Library and Archives Canada

Bibliothèque et Archives Canada

Canada

Library and Archives Canada
www.collectionscanada.gc.ca

Library and Archives Canada (LAC)
Local Digital Format Registry (LDFR)
File Format Guidelines for Preservation and Long-term Access
Version 1.0

Table of Contents

1 Introduction

1.1 Purpose

This document identifies the file formats that Library and Archives Canada (LAC) will be supporting within the Trusted Digital Repository (TDR). The formats are identified as:

- Recommended; or
- Acceptable for transfer.

"Recommended" formats are those that LAC believes will be sustainable over a long period of time, whereas the formats considered "acceptable for transfer" are those formats that LAC considers to be most representative of commonly used formats (formats in widespread use) in the collections that LAC will be preserving in the TDR (e.g., most commonly used formats in digital publications and Government of Canada (GoC) electronic records).

The list of file formats to be supported will evolve over time, particularly as new formats are introduced or older formats become obsolete. It should be noted that for any given collection submitted for preservation within LAC's TDR, file formats that do not fall within the category of "recommended" or "acceptable for transfer" will be evaluated on the basis of their content: where the content is deemed of preservation value, the content will be normalized/migrated to a "recommended" preservation format[1].

1.2 Background

1.2.1 Preserving digital information

Canadians have been generating digital information for decades. Our books, music, movies and the records of our private and public organizations are increasingly being created in digital formats. The preservation of this digital information is a problem that touches all sectors – academic, government, private and non-profit – and ultimately all Canadians.

By its very nature, digital information is fragile. Digital bits can be preserved, but our ability to use the information is at risk if the computer hardware and software needed to interpret/render the information are no longer available, or the format specifications are not accessible (e.g., the format is proprietary, is subject to intellectual property rights, or the specifications are no longer available). Preserving digital information is complicated. It involves the active commitment of organizations, the development of appropriate policies and plans, and the implementation of sound practices. It requires all organizations with an interest in preserving digital information to share expertise, advice and best practices.

Among these best practices, the identification and use of appropriate file formats is critical for preserving digital information. Due to a mix of technical and practical issues, certain file formats are more suitable for digital preservation. This document identifies and describes digital formats which LAC is recommending for long-term preservation and access to digital information.

[1] Note: Within the TDR, automatic normalization will be performed on the "acceptable for transfer" formats identified in the guidelines (conversion or migration to a "recommended" format): all other formats will be addressed on an individual case basis. Should the format prove to be a commonly used format, automated normalization/migration will be considered for future submissions.

These recommendations are contextualized within LAC's Digital Preservation Policy[2] and the development of LAC's TDR. The TDR is LAC's digital preservation infrastructure supporting secure acquisition, storage, management and continuing access to Canada's digital memory.

1.2.2 Digital content preservation strategy

LAC has adopted the following strategy for preserving digital content:

- When digital content is first accepted/approved for preservation in the TDR (that is, the content has been evaluated by LAC and deemed to be of preservation value), a preservation master is created (termed a "preservation master (0)" or PM(0));
- As part of the acceptance/approval process, the digital content is normalized as required (that is, migrated from the submitted/transferred format to one of the appropriate recommended preservation formats), thereby creating a new preservation master (termed a "preservation master (+1)" or PM(+1));
- From the current preservation master (i.e., PM(0) or PM(+1)), a copy of the digital content is created to service access requests by internal and external users (termed a "service copy")[3];
- The service copies can be presented using LAC-supported play-out services as well as client-based play-out services where needed or desired (an example of a play-out service would be an Apache server for HTML pages combined with a browser on the client, or a video streaming server; on the client, the Adobe Reader is an example of a client-based play-out service).

1.3 Target audience and use

LAC has developed these guidelines for a broad audience including the public, academic and private sectors. Whether it is a government department producing a budget or a citizen self-publishing, this document is intended to provide guidance on which digital file formats are most suitable for preservation and long-term access.

These guidelines also serve as the policy foundation for LAC's Local Digital Format Registry (LDFR), the underpinning set of guidelines for file format normalization/migration services within LAC's TDR.

1.4 Scope

These guidelines and recommendations are concerned with media-independent content; that is digital content that is managed as file types and is not inextricably linked to a physical storage medium (in contrast to videotape which is dependent both on the physical carrier and the playback equipment). These guidelines do not address recommendations for physical preservation media[4].

The file formats covered in this document have been clustered into the following content types:

- Text
- Audio
- Digital video
- Still images
- Web archiving

[2] http://www.collectionscanada.gc.ca/digital-initiatives/012018-2000.01-e.html

[3] A service copy may be created as part of the acceptance/approval process or may be produced dynamically.

[4] A policy addressing storage media for use in preservation is currently under development.

- Geospatial
- Structured data, including:
    - Databases
    - Statistical and Qualitative Analysis Data
    - Scientific Data
- Computer Aided Design (CAD):
    - Technical drawings
    - Computer-aided Software Engineering (CASE)

This document consists of file format recommendations based on LAC's experience in collecting and preserving digital content as well as international best practices.

1.5 Summary of recommendations

1.5.1 Definition of file formats

Generally speaking, file formats are specific patterns or structures which organize and define data. Some formats contain only one 'stream' of uncompressed data, others may contain codecs to encode and compress the data[5], and others still may support several 'streams' of media.

In addition to file formats, there are also 'container' or 'encapsulating' formats. These formats can contain and support various types or layers of audio, video, still imagery, and their associated metadata. Each of these formats may be handled by different programs, processes, or hardware; but for the multimedia data stream to be interpreted properly, the information must be encapsulated together. Library of Congress define three types of container formats:

- "wrapper" format: *wrapper* is often used by digital content specialists to name a file format that encapsulates its constituent bitstreams and includes metadata that describes the content within. Archetypal examples include WAVE and TIFF. Files that are instances of these wrappers are distinguished in terms of their underlying bitstreams, e.g., WAVE files may contain (a) linear pulse code modulated (LPCM) audio, (b) highly compressed audio as used for digital telephony, or (c) other representations of sound. Meanwhile, the self-describing, content-declaring feature of a wrapper is typified by the familiar TIFF header. Relatively more complex and facile wrappers like QuickTime may contain multiple objects, e.g., one or more video streams and separate audio streams;
- "simple bundling" formats: these formats encapsulate their constituent files and, save for a directory that provides the filenames, do not describe the content and the relationships that may exist between files. Archetypes include ZIP, StuffIt, and TAR, the latter associated with the UNIX operating system. Simple bundling formats tend to be generic, i.e., they may be used for a wide range of content types;
- "self-describing bundling" formats: these formats are employed to represent the bundle of files that comprise a complex digital work, e.g., a book text with supporting illustrations or a movie with multiple segments and sound tracks in different languages. Self-describing bundling formats list the component parts and their relationships (information about the relationships is often called *structural metadata*) and may indicate how the work as a whole can be rendered or used. Bundling formats often incorporate technical details about each component, since a single object may include a mix of texts, sound, images, etc. They may or may not encapsulate their constituent

[5] Please see Appendix C: Concepts and Definitions - Codecs.

files. They include metadata that describes their content and the relationships between files. Archetypes for this subcategory include METS (Metadata Encoding and Transmission Standard) and MPEG-21 (Multimedia Framework).

For further information on formats, see the working definition[6] on the Library of Congress Web site on Sustainability of Digital Formats.

There are thousands of file types now in existence: LAC's guidelines specify only the file formats that will be supported in the TDR. For a more complete registry please refer to PRONOM[7], the Unified Digital Format Registry[8] or the Library of Congress Web site on Sustainability of Digital Formats[9].

1.5.2 Evaluating the sustainability of file formats

In developing these guidelines, LAC has attempted to balance the requirements for quality, stability, potential longevity and industry acceptance. Where possible, a preference has been placed on the selection of non-proprietary national and international standards, or failing the availability of non-proprietary standards on, de facto standard industry formats. De facto standard formats are widely used and recognized formats that have become industry standards because of their ubiquitous use and support, and not because they have been formally approved by a standards organization. LAC has also reserved the right to select formats that it believes will become more widely adopted by the preservation community in the near future (e.g., SIARD).

Based on a review of criteria published by Library of Congress, the National Archives (UK), and the National Library of the Netherlands[10], Library and Archives Canada has established the following criteria for evaluating file formats for long-term preservation and access

1. ***Openness/Transparency***
   The relative ease with which knowledge of the file format and its technical information can be accumulated.
2. ***Adoption as a preservation standard***
   The extent to which the format has been formally adopted by national libraries, archives, and other memory institutions internationally.
3. ***Stability/Compatibility***
   a) The degree to which the format is backward and forward compatible.
   b) The degree to which the format is protected against file corruption.
   c) The relative frequency of release of newer or replacement versions of the format over time.
4. ***Dependencies/Interoperability*** The degree to which the format relies on a particular hardware or software, reader, etc.

---

6 http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml#what
7 http://www.nationalarchives.gov.uk/pronom/
8 http://www.gdfr.info/udfr.html
9 http://www.digitalpreservation.gov/formats/content/content_categories.shtml
10 *See* Gillesse *et al* 2008; Rauch, Carl *et al*. 'File-Formats for Preservation: Evaluating the Long-Term Stability of File-Formats." Proceedings ELPUB2007 Conference on Electronic Publishing : Vienna, Austria , 2007. http://elpub.scix.net/data/works/att/122_elpub2007.content.pdf; National Archives (UK). "Selecting File Formats for Long-Term Preservation." (2003). http://www.nationalarchives.gov.uk/documents/selecting_file_formats.rtf; Library of Congress. "Sustainability of Digital Formats: Planning for Library of Congress Collections." (2007). http://www.digitalpreservation.gov/formats/sustain/sustain.shtml.

5. ***Standardization*** The degree to which the format has gone through a rigorous formal standardization process.

Table 1, below, summarizes the evaluation scheme used, whereas Table 2, following, provides a definition for each evaluation criterion along with the rating to be assigned based on the degree to which the criterion has been met.

Table 1: Rating Scheme

Rating	
Symbol	Description
✓ ✓$	Evaluation criterion fully met Evaluation criterion fully met, however a cost is associated with meeting the criterion (e.g., to acquire the specification)
✶	Evaluation criterion partially met
×	Evaluation criterion not met
✓/× ✓/✶	Evaluation criterion met in one sector (e.g., for Government of Canada content) but not met / partially met in another sector (e.g., for non-government / commercial content)

1.5.3 File format recommendations

Table 3, following, summarizes the files formats that LAC recommends for the preservation of and long term access to digital content, and also identifies the file formats that are acceptable for the transfer of digital content to LAC.

Please note that there is no implied migration path from the "acceptable for transfer" formats and the "recommended" for preservation formats. The selection of a preservation format will be based on the degree to which the significant properties of the source format (and of individual instances of the format) are retained in the target preservation format (and the relative importance (or weigthing) of specific properties).

Table 4 summarizes the ratings of LAC's recommended file formats against the criteria identified in Section 1.5.2, whereas Appendix A – Recommended Preservation Format Evaluation provides detailed rating information. Please note that there is no implied order of preference / precedence in the list of formats.

Appendix B – Applying the Guidelines to LAC Preservation Policies, graphically demonstrates the mapping of the recommended preservation formats to LAC's preservation strategy (outlined in Section 1.2.2).

Table 2: Evaluation Criteria Definition and Rating

Criterion	Evaluation Basis	Rating
Openness/Transparency	Specifications available from one or more of the following: a) Open membership organization (such as the W3C (World Wide Web Consortium), the OMG (Object Management Group)) b) International standards organization (such as the ISO) c) Industry-based open membership organization	✓
	Specifications available only at cost	✓$
	Specifications potentially available from multiple sources (could not be confirmed)	✶
	Specifications only available from / under the control of a single vendor or small group of vendors	×
Adoption as a preservation standard	The majority of the organizations investigated use/are planning to use the format as a preservation standard (50% or more of the organizations)	✓
	Some of the organizations investigated use/are planning to use the format as a preservation standard (less than 50% of the organizations)	✶
	None of the organizations investigated use/are planning to use the format as a preservation standard	×
Stability/Compatibility		
a) degree of forward/backward compatibility	A format is backward compatible if it provides all of the functionality of a previous release or version of the format A format is forward compatible if it has the ability to gracefully accept content intended for later versions of the format (that is, software designed to interpret / render a prior version of a format can also interpret / render the current version of the format) Forward/backward compatibility: a) High compatibility: A format is both forward and backward compatible b) Medium compatibility: A format is backward compatible only c) Low compatibility: A format is neither forward nor backward compatible	 ✓ ✶ ×
b) degree of protection against file corruption	Corruption protection: Resilience to random bit-level/byte-level changes in content a) High resilience: Changes have little or no impact to renderability/interpretability / uses methods for detecting/recovering from changes b) Medium resilience: Changes affect renderability but not interpretability / some ability to recover from changes c) Low resilience: Any change affects the ability to interpret and render the format	 ✓ ✶ ×
b) frequency of version releases	Format stability demonstrated by the number of version releases and/or extensions; format's use in derivatives and/or industry-specific applications High format stability	 ✓

<table>
<tr><th>Criterion</th><th>Evaluation Basis</th><th>Rating</th></tr>
<tr><td></td><td>Medium format stability
Low format stability</td><td>✶
×</td></tr>
<tr><td>Dependencies/Interoperability</td><td>
<table>
<tr><th></th><th>Dependency</th><th>Interoperability</th></tr>
<tr><td>Low</td><td>High availability of low-cost/free software to render/interpret the format; “humanly readable” format; little or no dependency on other formats / dependency only on non-proprietary formats</td><td>Format renderable on a very small set of platforms (such as, electronic book formats limited to one or two hardware platforms, or supported by a single software vendor (e.g., Microsoft LIT readable only with proprietary reader))</td></tr>
<tr><td>Medium</td><td>Availability of software from many vendors to interpret / render the format</td><td>Format renderable on a small set of mainstream hardware / software platforms</td></tr>
<tr><td>High</td><td>Some/high dependency on proprietary formats; low availability of software to interpret/render the format; format not “humanly readable” (e.g., binary format)</td><td>Format renderable on a large number of platforms (e.g., multiple OS, hardware (such as, EPUB format support on PDAs))</td></tr>
</table>
</td><td></td></tr>
<tr><td></td><td>Low dependency / High interoperability
Low dependency / Medium interoperability
Medium dependency / High interoperability</td><td>✓</td></tr>
<tr><td></td><td>Low dependency / Low interoperability
Medium dependency / Medium interoperability
Medium dependency / Low interoperability</td><td>✶</td></tr>
<tr><td></td><td>High dependency / Low interoperability
High dependency / Medium interoperability
High dependency / High interoperability</td><td>×</td></tr>
<tr><td>Standardization</td><td>Format follows a formal process enacted by any of the following:
a) Open membership organization (such as the W3C (World Wide Web Consortium), the OMG (Object Management Group))
b) International standards organization (such as the ISO)
c) Industry-based open membership organization</td><td>✓</td></tr>
<tr><td></td><td>Format is subject to documented processes implemented by a single vendor or small group of vendors or no documented process</td><td>×</td></tr>
</table>

Table 3: Recommended and Acceptable for Transfer File Formats

Content Type	Recommended	Acceptable for transfer
Text	• EPUB for electronic books • Extensible Hypertext Markup Language (XHTML) • Extensible Markup Language (XML) • Hypertext Markup Language (HTML) • Multipurpose Internet Mail Extensions (MIME) • Open Document Format (ODF) • PDF for long-term preservation (PDF/A) • Rich Text Format (RTF) • Standard General Markup Language (SGML) • Text (plain text)	• Office Suites: o Microsoft Office including: Word Document Format, Excel Spreadsheet Format, Powerpoint Presentation Format o WordPerfect Suite including: WordPerfect Document Format, Quattro Pro Spreadsheet Format, Corel Presentations Format o Lotus Smartsuite including: WordPro Document Format, 1-2-3 Spreadsheet Format, Freelance Graphics Format • Portable Document Format (PDF)
Audio	• Broadcast Wave Format (BWF) (for newly digitized content (i.e., creating)) • Waveform Audio Format (WAV) (for migrating born digital audio content)	• Audio Interchange File Format (AIFF) • Mpeg-1 layer-3, Mepg-2 layer-3 (MP3) • Mpeg-4 aac – advanced audio coding (AAC) • Musical instrument digital interface (MIDI) • Window media audio (WMA)
Digital Video	• Motion JPEG 2000	• Audio video interleave (AVI) • Moving pictures expert group (MPEG-2) • Moving pictures expert group (MPEG-4) • Quicktime (MOV) • Windows media video (WMV)
Still Images	• Joint photographic experts group (JPEG) • Joint photographic experts group jpeg 2000 (JP2) • Tagged image file format (TIFF) • TIFF - GeoTIFF	• Digital imaging and communications in medicine (DICOM v. 3.0) • Encapsulated postscript (EPS) • Graphics interchange format (GIF) • Portable network graphics (PNG)
Web Archiving	• Internet archive format (ARC) • Web archive format (WARC)	
Structured Data - Databases	• Software Independent Archiving of Relational Databases (SIARD)	• dBase Format (DBF)

Content Type	Recommended	Acceptable for transfer
	• Delimited Flat file with DDL	
Structured Data – Statistical and Qualitative Analysis	• Data Documentation Initiative (DDI) Version 3.0 • Data Exchange and Conversion Utilities and Tools (DExT) • Statistical Data and Metadata Exchange (SDMX) • Delimited Flat File with Variable Descriptions	• SAS • SPSS
Structured Data – Scientific	• XML Container	
Geospatial[11]	• ISO 19115 Geographic Information – Metadata (NAP – Metadata) (North American Profile)	• Canadian Council on Geomatics Interchange Format (CCOGIF) • Digital Elevation Model (DEM) • Digital line graphics – level 3 (DIG-3)* • Environmental systems research institute (ESRI) export format – (E00)* • Environmental systems research institute (ESRI) shape file format (SHP)* • International Hydrographic Organization (IHO) S-57, Edition 3.1*
Computer Aided Design – Technical Drawing	• Drawing Interchange File Format/Data eXchange Format (DXF)	• Computer Graphics Metafile (CGM)
Computer Aided Design – CASE	• XML Metadata Interchange (XMI)	
Source Code and Scripts	• XML Container	• Text

[11] For geospatial information, the "acceptable for transfer" formats with asterisks will be preserved as is (not migrated) until such time as the adoption rate of the Treasury Board Secretariat (TBS) standard (identifying ISO 19115), and the avalaibility of tools supporting the standard is more fully understood (exception to preservation strategy for the near future).

Table 4: Summary Evaluation of Recommended File Formats

Content Type	Format	Openness / Transparency	Adoption	Stability / Compatibility			Dependencies / Interoperability	Standardization
				Forward/Backward Compatibility	Corruption Protection	Release Stability		
Text	EPUB (underlying standard for eBooks)	✓	✶	✶		✓	✓	✓
	Extensible Markup Language (XML)	✓	✓	✓		✓	✓	✓
	Extensible HyperText Markup Language (XHTML)	✓	✶	✶		✓	✓	✓
	HyperText Markup Language (HTML)	✓	✓	✶		✓	✓	✓
	Multipurpose Internet Mail Extensions (MIME)	✓	✶	✶		✓	✓	✓
	Open Document Format (ODF)	✓	✓	✶		✶	✓	✓
	PDF for long-term preservation: PDF-Archive (PDF/A)	✓$	✓	✶		✓	✓	✓
	Rich Text Format (RTF)	×	✓	✶		✶	×	×
	Standard Generalized Markup Language (SGML)	✓$	✓	✓		✓	✓	✓
	Text (TXT)	✓$	✓	✶		✓	✓	✓
Audio	Broadcast Wave Format (BWF)	✓	✓	✓		✓	✓	✓
Digital Video	JPEG 2000 MXF (MOTION JPEG 2000)	✓$	✓	✓	✓	✓	✓	✓
Still Images	Joint Photographic Experts Group (JPEG)	✓$	✓			✓	✓	✓
	Joint Photographic Experts Group JPEG2000 (JP2)	✓$	✓			✓	✓	✓
	Portable Network Graphics (PNG)	✓$	✓			✓	✓	✓
	Tagged Image File Format (TIFF)	✓	✓			✓	✓	✓
	TIFF - GeoTIFF	✓	×	✓		✶	✓	✓
Structured Data - Database	Software Independent Archiving of Relational Databases (SIARD)	✶	✶			✶	✓	✶
	Delimited Flat File with Data Description	✓$	✓	✶		✓	✓	✓
Structured Data - Statistical and Qualitative Analysis Data	Data Documentation Initiative (DDI) Version 3.0	✓	✶			✓	✓	✓
	Data Exchange and Conversion Utilities and Tools (DExT)	✓	✶			✶	✓	✓
	Statistical Data and Metadata Exchange (SDMX)	✓	✶	✶		✶	✓	✓
	Delimited Flat File with Variable Description	✓$	✶	✶		✓	✓	✓
Structured Data - Scientific Data	Not applicable at this time							
Geospatial Data	ISO 19115 Geographic Information – Metadata (NAP – Metadata) (North American Profile)	✓$	✓GoC /n.a.					✓
Computer-Aided Design (CAD) – Technical Drawings	Drawing Interchange File Format (DXF)	×	✓				✓	✶
Computer-Aided Design (CAD) – CASE	XML Metadata Interchange (XMI)	✓	×			✶	✓	✓
Source Code and Scripts	Not applicable at this time							

2 Guidelines and recommendations

Please note: The formats in each Content Category section are organized alphabetically and do not imply an order of preference for any given format.

2.1 Content Category: Text[12]

There are two general types of text in this category; plain and formatted. Plain text files contain encoded ASCII or Unicode data[13] that has no formatting or layout code to influence the presentation of the data. These files can be opened and read by any software package that is capable of interpreting and displaying the numeric and alpha-numeric data values that are contained in a file.

Formatted text files such as rich text format (rtf) contain encoded ASCII data and format definitions that display the information in a defined pattern. Formatted text files are primarily created using word processing software such as WordPerfect and Microsoft Word. Markup languages such as HTML and XML are also considered to be formatted text.

Generally speaking, textual file formats can be subdivided into three categories: Word Processing, Structural Markup, and Page Layout.

The current de facto formatted document file format is Microsoft Word, which uses the .doc extension (and more recently the .docx extension). However, the OpenDocument format (ODF) is slowly gaining recognition, and may emerge as the preferred file format for the creation of office documents in the future (see Appendix C, Formats Under Investigation for further details regarding the ODF format).

LAC does not recommend the acquisition of proprietary binary formats that are created by various word processing software applications such as Microsoft Word and WordPerfect, or desktop-publishing applications such as Quark. However, the institution recognizes that this may be impossible to apply in all situations. Furthermore, the Library of Congress recommends that text documents that are created by the most popular word processing applications be converted to the PDF format (preferably PDF/A), or to a non-proprietary format such as OpenOffice, which is XML-based.

2.1.1 Recommended

2.1.1.1 EPUB (Electronic Publications)

Electronic books are available in a variety of formats aimed for distribution on computer and personal digital assistant (PDA) devices. Each device has specialized reading systems (such as Adobe Digital Editions, Sanza/iPhone, MobiPocket, Kindle, eBookwise, Sony Reader, BEBOOK, Bookworm, iRex, ETI's eBook Technology Suite).

The International Digital Publishing Forum (IDPF)[14]:

- is the Trade and Standards organization for the digital publishing industry;
- is a Business Special Interest Group (BSIG); and
- develops and maintains industry standards for text-based digital reflowable books and publications[15].

[12] This section does not address the digitization of textual material (books, newspapers, magazines) which often uses image formats for conversion. For recommendations on imaging text for the purpose of digitization, please refer to Section 2.4.1.
[13] Please refer to Appendix C: Concepts and Definitions: Character Sets for a detailed definition of character sets.
[14] Source: EPUB 101.pdf available from International Digital Publishing Forum (http://www.idpf.org/).

EPUB is the current standard for the production of electronic publications. EPUB is in essence a packaging of Open Publication Structure (OPS) publications into an Open Container Format (OCF) container, whereby:

- OCF is the container in which Publications are packaged for transport and potential delivery;
- OPS is the Publication standard (the "markup" or "source" of a Publication) – XHTML or DTBook based[16];
- OPS includes Open Packaging Format (OPF): OPF contains the Publication manifest (list of files, images, style sheets, et al), the "spine" (linear reading order), the metadata (title, author, language, etc.), navigation information.

EPUB addresses the content and presentation without digital rights management (DRM) allowing:

- publishers to produce a single format of the publication; and
- rendering of content to be performed from a single version of the document to any number of devices (using device specific reading systems (or play-out services).

EPUB does not currently address DRM, although DRM can be added for/within individual reading systems (using "just in time" approach).

2.1.1.2 eXtensible Hypertext Markup Language (XHTML)[17]

XHTML is a reformulation of HTML 4 as a XML application. XHTML 1.0 became a W3C recommendation in January 2000. XHTML 1.1 reformatted XHTML 1.0 into XHTML modules.

This modularization provided the ability to extend and create subsets of XHTML, which made it easier to combine markup tags for vector graphics, multimedia, math, e-commerce and other applications. Version 1.1 became a W3C recommendation in May 2001. XHTML version 2.0 is currently being developed and will not be backwards compatible with previous versions. At the time of writing, version 2.0 cannot be considered stable.

As a result, LAC only recommends the use of XHTML versions 1.0 and 1.1. LAC will continue to monitor the development of version 2.0.

2.1.1.3 eXtensible Markup Language (XML)

XML is a simple, flexible, and platform independent markup language that is considered to be a subset of SGML. It describes how you should format your tags, how you should document your definitions and how you will define your schema.

XML tags are fully extensible and user defined. They are used to describe the content of the text rather than its appearance. This allows for more efficient searching, but documentation of the tags is critical for one to be able to interpret a XML document.

---

[15] Reflowable refers to the ability to automatically adjust the display of text in a document based on user changes to the display area.

[16] The OPS standard also incorporates the DAISY DTBook (Digital Talking Book) specification alongside XHTML as a Preferred Vocabulary (DAISY/NISO Z39.86, DTBook).
Note: Audio books are also produced using existing audio formats (and as a result will be subject to the recommended/acceptable file formats identified in Section 2.2 Content Category: Audio); however, with the move towards a single publication production format (EPUB), it is likely that this format will also be used for audio books in the future as an accessibility feature of the electronic publication itself (e.g., through the next evolution of standards such as the National Instructional Materials Accessibility Standard (NIMAS)).

[17] XHTML, XML, and HTML documents are subject to being "well formed" and "valid" – please refer to Appendix C: Concepts and Definitions - Well formed/Well formedness and Document Validity.

XML became a World Wide Web Consortium (W3C) recommendation in 1998 and it is now fully supported by all the leading software providers.

Since the use of XML is practiced at differing levels of technical maturity among federal government agencies and departments, LAC is monitoring developments in the creation and use of domain specific XML schema definitions. LAC will continue to monitor, evaluate and adopt specific XML formats as the schema definitions are developed, reviewed and approved by specific user communities (such as for online journals publishing (e.g., Érudit))[18].

The Library of Congress has indicated that they would use XML as a preferred format if it was conformant to an appropriate standard or community agreed upon DTD or schema that can be used for technical validation.

Preservation: XML documents to be preserved in the TDR will conform to a published schema or DTD, or will have a schema or DTD provided with the documents (and the schema/DTD preserved in the TDR).

2.1.1.4 HyperText Markup Language (HTML)[17]

HTML is a simple markup language derived from SGML (see Section 2.1.1.9 for a detailed definition of SGML). It is used to create hypertext documents that are portable from one computer platform to another and it has become the standard format for producing documents for the World-Wide Web.

Each HTML version contains a specific, non-extensible set of tags that are used to specify the appearance of the document being created. LAC recommends that GoC departments and agencies produce HTML 4.01 documents rather than HTML 4.0 documents[19].

2.1.1.5 Multipurpose Internet Mail Extensions (MIME)[20]

The MIME format is an Internet standard that specifies how messages must be formatted so they can be exchanged between different email systems. MIME allows email messages to contain:

- Multiple objects in a single message.
- Text having unlimited line length or overall length.
- Character sets other than ASCII, allowing non-English language messages.
- Multi-font messages.
- Binary or application specific files.
- Images, Audio, Video and multi-media messages.

2.1.1.6 Open Document Format (ODF)

The OpenDocument format (ODF) is a file format for representing electronic documents such as spreadsheets, charts, presentations and word processing documents.

While the specifications were originally developed by Sun Microsystems, the standard was developed by the OASIS Open Document Format for Office Applications (OpenDocument) TC - OASIS ODF TC, committee of the Organization for the Advancement of Structured Information Standards (OASIS)

[18] XML is also being used by LAC as an information interchange format (e.g., METS for TDR content, or ERTA (GoC electronic records (e-Records) Transfer Application)).
[19] HTML 4.01 is the basis for the international standard ISO/IEC 15445:2000.
[20] MIME will be used to retain mail messages in a non-proprietary format (e.g., used to preserve mail from systems such as Microsoft Exchange), whereas the attachments will be subject to the preservation formats for the content type of the attachment. LAC will also be monitoring developments in XML-based formats for preserving electronic mail.

consortium and based on the XML format originally created and implemented by the OpenOffice.org office suite (see OpenOffice.org XML).

In addition to being a free and open OASIS standard, it is published (in one of its version 1.0 manifestations) as an ISO/IEC international standard, ISO/IEC 26300:2006 Open Document Format for Office Applications (OpenDocument) v1.0. Published ODF standards meet the common definitions of an open standard, meaning they are freely available and implementable.

The most common filename extensions used for OpenDocument documents are:

- .odt for word processing (text) documents
- .ods for spreadsheets
- .odb for object-oriented database
- .odp for presentations
- .odg for graphics
- .odf for formulae, mathematical equations

A basic OpenDocument file consists of an XML document that has <document> as its root element. OpenDocument files can also take the form of a ZIP compressed archive containing a number of files and directories; these can contain binary content and benefit from ZIP's lossless compression to reduce file size. The OpenDocument Format benefits from the separation of concerns: it separates the content, styles, metadata and application settings into four separate XML files.

2.1.1.7 PDF for long-term preservation: PDF-Archive (PDF/A)[21]

The Association for Suppliers of Printing, Publishing and Converting Technologies (NPES), and the Association for Information and Image Management International (AIIM International) have developed an international standard that defines the use of PDF for archiving and preserving documents. The format is known as PDF-Archive (PDF/A) and has been adopted by the ISO (ISO standard 19005-1:2005).

PDF/A will be desirable where the original content rendering characteristics are of greater or of equal importance to the ability to repurpose the content (as supported in XML).

2.1.1.8 Rich Text Format (RTF)[22]

The RTF Specification provides a format for text and graphics interchange that can be used with different output devices, operating environments, and operating systems. RTF uses the ANSI, PC-8, Macintosh, or IBM PC character set to control the representation and formatting of a document, both on the screen and in print. With the RTF Specification, documents created under different operating systems and with different software applications can be transferred between those operating systems and applications. RTF files created in Word 6.0 (and later) for the Macintosh and Power Macintosh have a file type of "RTF."

2.1.1.9 Standard Generalized Markup Language (SGML)

SGML is defined in international standard ISO 8879:1986. It is a markup language used for formally describing the structure and contents of documents. Tags in SGML are used to identify, name and

[21] PDF-Archive or PDF/A is "a file format based on PDF, known as PDF/A, which provides a mechanism for representing electronic documents in a manner that preserves their visual appearance over time, independent of the tools and systems used for creating, storing or rending the files." (from ISO 19005-1). It should be noted PDF/A does not support all features that are available in current releases of the PDF file format (such as, bookmarks, internal document links), although some of these features may be added to revisions of the PDF/A standard.

[22] RTF is an example of a de facto standard that many organizations have elected to preserve.

describe relationships between data, so they can be managed and manipulated. SGML-based applications are platform independent and are used for a variety of functions[23].

An SGML document has three elements:

- The Declaration which describes the processing environment that is required;
- The Document Type Definition (DTD) which is a defined tag set that forms a template for describing the structure and content of a specific type of document; and,
- The Document stream itself.

Some relevant SGML DTDs are:

- EAD (Encoded Archival Description) which is a DTD for archival material;
- US MARC DTD**;**
- HTML (Hyper Text Markup Language);
- XHTML (eXtensible Hyper Text Markup Language);
- XML (eXtensible Markup Language)

2.1.1.10 Plain Text (Text)

LAC will accept plain text files that use the ISO/IEC 8859-1:1998 ASCII character set for encoding[24]. Plain text files typically use the extension .txt and contain ASCII-encoded text with no formatting or layout information[25]. They can be opened and read by any program that reads text.

2.1.2 Acceptable for transfer

2.1.2.1 Office Suites[26]

In computing, an **office suite**, sometimes called an **office software suite** or **productivity suite** is a collection of programs intended to be used by knowledge workers. The components are generally distributed together, have a consistent user interface and usually can interact with each other, sometimes in ways that the operating system would not normally allow.

The currently dominant office suites are Microsoft Office, which is available for Microsoft Windows and Apple Inc.'s Mac OS X, and OpenOffice.org, a free software / open source alternative available for many operating systems. Microsoft Office's binary file formats have been the *de facto* default, but are being supplanted by the open standard ODF, OpenOffice.org's (among other programs) native format. ODF is becoming the standard choice for governments, educational environments and businesses seeking neutral document formats for information exchange or seeking to save money. There are numerous office suites attempting to challenge Microsoft and OpenOffice.org.

The formats in this section are presented in descending order of use in industry and the GoC.

---

[23] Functions include: documentation (such as, CALS (Continuous Acquisition and Life-cycle Support) is a US Department of Defense (DoD) initiative for electronically capturing military documents and for linking related data and information), and the Text Encoding Initiative (TEI) (an academic consortium that designs, maintains, and develops technical standards for digital-format textual representation applications), and information exchange (such as EDGAR (Electronic Data-Gathering, Analysis, and Retrieval) a system that effects the automated collection, validation, indexing, acceptance, and forwarding of submissions, by companies and others, who are legally required to file data and information forms with the US Securities and Exchange Commission (SEC)).

[24] Please refer to Appendix C: Concepts and Definitions: Character Sets.

[25] Examples of software used to create text files include Notepad on Windows operating systems, or Vi text editor on Unix systems, or TextEdit on Mac OSX systems.

[26] http://en.wikipedia.org/wiki/Office_suite

Note: The primary types of file that are of preservation concern in this category are those created with the word processing, spreadsheet, and presentation components of these suites. Databases (such as, Lotus Approach) will be addressed under Section 2.6.

2.1.2.1.1. Microsoft Office Suite[27]

Microsoft Office is an office suite of interrelated desktop applications, servers and services for the Microsoft Windows and Mac OS X operating systems. Microsoft Office was introduced by Microsoft in 1989 for Mac OS, with a version for Windows in 1990. Initially a marketing term for a bundled set of applications, the first version of Office contained Microsoft Word, Microsoft Excel, and Microsoft PowerPoint. Additionally, a "Pro" (Professional) version of Office included Microsoft Access and Schedule Plus. Over the years, Office applications have grown substantially closer with shared features such as a common spell checker, OLE data integration and Microsoft Visual Basic for Applications scripting language. Microsoft also positions Office as a development platform for line-of-business software under the Office Business Applications (OBA) brand.

The current versions are Office 2007 for Windows which was released on January 30, 2007, and Office 2008 for Mac OS X, released January 15, 2008. Office 2007/Office 2008 introduced a new user interface and new Office Open XML document formats (docx, xlsx, pptx). Consequently, Microsoft has made available, free of charge, an add-on known as the *Microsoft Office Compatibility Pack* to allow Office 2000-2003 for Windows and Office 2004 for Mac editions to open, edit, and save documents created under the new formats for Office 2007.

2.1.2.1.2. Corel WordPerfect Office Suite[28]

WordPerfect Office is an office suite developed by Corel Corporation. As of April 2008, the latest version is **WordPerfect Office X4** (representing 14), which is available in various editions, including Standard, Professional, Home & Student and Family Pack Edition. WordPerfect Office 12 Small Business Edition, including a wider range of applications, including tools for photo editing and Internet security, is also available.

Its predecessor was **WordPerfect Suite**, assembled by Novell in 1994 and sold to Corel in 1996.

The major components of WordPerfect Office X4 - Standard Edition are:

- WordPerfect X4, word processor (***.wpd**-files*)
- Quattro Pro X4, spreadsheet (***.qpw**-files*)
- Presentations X3, slide show creation (***.shw**-files*)
- WordPerfect MAIL, Email client and personal information manager

2.1.2.1.3. Lotus SmartSuite[29]

SmartSuite is an office suite from Lotus Software. Lotus made versions for IBM's OS/2, as well as Microsoft Windows versions.

SmartSuite is in maintenance mode, and supported with fixes and fixpacks on Windows 2000 and Windows XP. SmartSuite is not officially supported by IBM on the Windows Vista operating system, but it does work on the 32-bit version of Vista if the installer and applications are run in XP compatibility mode (this isn't needed to install or run Organizer 6). IBM has no plans to release specific Vista-compatible versions of SmartSuite or Organizer.

[27] http://en.wikipedia.org/wiki/Microsoft_Office
[28] http://en.wikipedia.org/wiki/WordPerfect_Office
[29] http://en.wikipedia.org/wiki/Lotus_SmartSuite

In 2007, IBM introduced a new office suite called IBM Lotus Symphony.

The following applications are included in SmartSuite for Microsoft Windows:

- Lotus Word Pro — word processor (previously called Ami Pro) - *.lwp* files
- Lotus 1-2-3 — spreadsheet - *.123, .wk1, .wk3, .wk4* files
- Lotus Freelance Graphics — presentation software - *.prz* files
- Lotus Approach — relational database - *.apr* (data entry & reports), *.dbf* (database) files
- Lotus Organizer — personal information manager - *.org .or2 .or3* files
- Lotus SmartCenter — a toolbar that lets users quickly access programs, their calendar, Internet bookmarks, and other resources
- Lotus FastSite — web design software - *.htm* files
- Lotus ScreenCam — recording of screen activity for demos and tutorials - *.scm, .exe, .wav* files

2.1.2.2 Portable Document Format (PDF)

PDF is an open, de facto standard that was developed by Adobe for the electronic distribution of textually based documents in raster format. It is a widely used format that preserves all the fonts, formatting, graphics and colours contained in the original source document after its conversion to the PDF format. PDF is fully backwards compatible and platform independent.

2.2 Content Category: Audio

Sound is digitized by "measuring" the voltage (produced by a microphone) representation of the sound wave at regular intervals. How often the sound wave is measured is called the sampling rate and is generally expressed in kHz (i.e. thousands of times per second). The standard sampling rate for compact disc recordings is 44.1 kHz or 44,100 times per second. Master quality recordings have a minimum sampling rate of 96 kHz.

The "measurements" are expressed in bits. Audio CD's have a bit depth (number of bits used to measure the voltage) of 16 which yields a total of 65,536 values. Master quality recordings commonly use 24 bits, which yields 16,777,216 values.

Bit rate is the total number of bits used per second (sampling rate X bit depth). Therefore a CD quality representation of sound of 44.1 kHz/16 bits would have a bit rate of: 44,100 X 16 = 705,600 bits per second per channel (seeing that CDs are stereo the total would be 1,411,200 bits per second).

Audio analog-to-digital converters allow for 192 kHz sampling rate and 24 bit amplitude resolution. The International Association of Sound Archives (IASA) recommends a minimum digital resolution of 48 kHz sampling rate at 24-bit resolution for analog originals. But, IASA acknowledges that the higher resolution of 96 kHz/24 bit has become the standard for heritage organizations. IASA also recommends that spoken word recordings be captured at the same rate as music recordings.

2.2.1 Recommended

Preferred digitization quality: **96 kHz/ 24**

Minimum digitization quality: **48 kHz/ 24**

The recommended audio file formats are classified as being uncompressed file formats.

2.2.1.1 Broadcast Wave Format (BWF)

The European Broadcast Union (EBU) introduced BWF in 1996 to allow files to be exchanged between digital audio workstations during radio and television productions. It is now used in every aspect of professional audio. Based on Microsoft's and IBM's WAV format, BWF can carry PCM (Pulse Code Modulation) or MPEG encoded audio which can be enhanced with metadata describing information about the originator, date and coding history of the recording. A BWF file is fully compatible with any playback software that supports regular WAV files. The International Association of Sound and Audiovisual Archives (IASA) recommends the use of BWF as an archival audio file format, and LAC has recently switched to this from standard WAV.

2.2.2 Acceptable for transfer

2.2.2.1 Audio Interchange File Format (AIFF)

The AIFF format was developed by Apple Computer in 1988 as the standard audio format for Macintosh computers. The audio data in a standard AIFF file is uncompressed pulse-code modulation (PCM). There is also a compressed variant of AIFF known as AIFF-C or AIFC, with various defined compression codecs.[30] The AIFF file format, much like its WAV counterpart, is an uncompressed digital audio format of choice for professional audio and video applications. AIFF files do have one enhanced characteristic over standard WAV files: in addition to containing audio data, they can include loop point data and also be used as a wrapper for MIDI data.

[30] Basic definition of the AIFF audio file format from Wikipedia http://en.wikipedia.org/wiki/AIFF

2.2.2.2 Moving Pictures Expert Group (MPEG) MPEG-1 layer-3, MPEG-2 layer-3 (MP3)

Preferred fidelity: **192 kbit/s @ 44.1 KHz**

Minimum fidelity: **128 kbit/s @ 44.1 KHz**

MP3 is a lossy codec, and is the most widely adopted digital audio dissemination format in use today. It works by eliminating inaudible frequencies in combination with a high compression rate (rates of 10:1 are possible).

The primary advantage of MP3 is its universality; unlike most other file formats, just about every digital music player and player program can handle the MP3 format for music.

2.2.2.3 MPEG-4 AAC – Advanced Audio Coding (AAC)

Preferred fidelity: **192 kbit/s @ 44.1 KHz**

Minimum fidelity: **128 kbit/s @ 44.1 KHz**

AAC is a lossy codec that was created by Fraunhofer-Gesellschaft as the next generation MP3 codec. Due to advances in technology, AAC files generally achieve better sound quality at smaller bit rates than a similarly encoded MP3 file.

The format comes in 2 varieties: MPEG-2 AAC and MPEG-4 AAC and both are referred to as AAC. MPEG-2 AAC is used for several purposes. It is part of the specification for the DVD-Audio Recordable (DVD-AR) format and it can also be used for streaming and downloading online audio.

MPEG-4 AAC is a newer specification that includes more capabilities. It serves as a multimedia container format and delivers higher-quality sound than MPEG-2. It is currently the default codec for Apple's iTunes, ensuring its continued use; the media player which powers the iPod (the most popular line of portable digital audio players on the market). This format was selected by Apple because it supports the implementation of Digital Rights Management (DRM).

NOTE: The m4p extension indicates that a file is protected, whereas the .m4a extension indicates that a file is unprotected.

2.2.2.4 Musical Instrument Digital Interface (MIDI)

MIDI is a standard adopted by the electronic music industry for controlling devices such as synthesizers and sound cards that emit music. At a minimum, a MIDI representation of a sound includes the note's pitch, length and volume, but it also can include other characteristics like attack and delay time. MIDI files are different from regular digital audio files as they do not contain any waveform data, but merely a set of playback instructions. The correct playback of a MIDI file requires an external hardware or software synthesizer to recreate the recorded composition. LAC will accept a MIDI composition for works natively created in the format that cannot be obtained in a standard waveform digital audio file.

2.2.2.5 WAVeform Audio Format (WAV)

The WAV audio file format was developed jointly by Microsoft and IBM, with native support for WAV files built into the Windows operating system. The format supports many different bit resolutions, sample rates, audio channels and a number of lossless compression methods; however WAV files are very rarely compressed.

The most common WAV format contains uncompressed audio in the Pulse Code Modulation (PCM) format. PCM audio is the standard audio file format for recording compact discs at 44,100 samples per second, 16 bits per sample. Since PCM uses an uncompressed, lossless storage method, the WAV format maintains maximum audio quality.

As a long-standing digital audio format, WAV remains the *de facto* standard for lossless PC audio files in use today.

2.2.2.6 Window Media Audio (WMA)

Preferred fidelity: **192 kbit/s @ 44.1 KHz**

Minimum fidelity: **128 kbit/s @ 44.1 KHz**

This is a proprietary compressed audio file format developed by Microsoft and is similar to the popular MP3 format. With the introduction of WMA Pro and Apple's iTunes Music Store, WMA has positioned itself as a competitor to the AAC format used by Apple and is part of Microsoft's Windows Media framework.

2.3 Content Category: Digital Video

Digital video is comprised of a sequence of bitmap digital images displayed in rapid succession at a constant rate. In the context of video these images are called frames. Every bitmap frame comprises a raster of pixels.[31]

The higher the frame rate the better the motion, and the higher the bits per pixel, the better the colour quality. Unlike analog video which is typically stored on magnetic tape, subject to physical deterioration and signal degradation with each subsequent copy, it is possible to copy multiple generations of digital video files with no loss in quality.

There are a myriad of different digital video file formats and codecs available in the marketplace today; each one designed for a different purpose from professional video and film production to the smallest consumer camcorders and cell phone devices. The technical trends within the video industry are constantly in motion. The task of narrowing the list or stating precise recommended specifications and formats is difficult. The metrics applied to measure "video quality" are often subjective.

In the perfect scenario, it would be desirable to ensure that all submissions to LAC be uncompressed. Obviously, this demand may not be feasible as many professional video acquisition and post-production digital formats still use some form of compression and the potential storage requirements involved may not be available. Historically significant video documents can come in all forms, including consumer digital video formats that are highly compressed.

2.3.1 Recommended

2.3.1.1 JPEG 2000 MXF (or Motion JPEG 2000)

LAC is currently using MXF wrapped JPEG 2000 for the lossless preservation of video in digital form and as the format of choice for the migration of obsolete analog video recordings into digital files. In this format, a video is stored as a sequence of JPEG 2000 encoded frames contained within a .MXF file container. JPEG 2000 offers a lossless, completely reversible, intra-frame encoding scheme that can contain multiple resolutions and quality layers within the same file. It can support standard and high definition resolutions. The .MXF file container is a standard interchange format developed by the SMPTE (Society of Motion Picture and Television Engineers) that supports frame accurate time code, the inclusion of technical and descriptive metadata and was designed to allow for greater compatibility between different software and hardware implementations.

2.3.2 Acceptable for transfer

Digital video files destined for transfer to LAC for long-term preservation should adhere to established NTSC standard broadcast resolutions for SD (Standard Definition) and HD (High Definition) video:

Standard Definition NTSC:
720 x 480 29.97fps (480i, 480p)
Aspect ratios: 4:3 or 16:9

High Definition NTSC:
1280 x 720 (720p60, 720p30, 720p24)
1920 x 1080 (1080i60, 1080p30, 1080p24)
Aspect ratio: 16:9

[31] Basic definition of digital video from Wikipedia http://en.wikipedia.org/wiki/Digital_video

Compliance with a broadcast standard resolution is not the only variable that determines digital video quality, but it is the most important characteristic as it ensures that a video file can be played back on standard display equipment without visual alteration and can be transferred for preservation without requiring changes to the original frame size or aspect ratio.

The encoding bit rate of a digital video file, normally calculated as the amount of data allocated per second (bits per second), determines the level of detail or image clarity of the video. Typically, the higher the bit rate, the more visual information is retained in the video. A video that has been encoded with a very low bit rate may contain compression artefacts and an insufficient level of detail. As there are many different file formats and compression codecs, each performing better at different bit rates and file sizes, it is difficult to dictate specific requirements. LAC endeavours to collect the highest quality version of a video document that is available. If the only available source material originated in a non-standard resolution or a low resolution, it may still be a candidate for preservation.

The file formats that follow in this section are examples of some of the most common file containers and standard codecs for the distribution of digital video – but is by no means an attempt at an exhaustive list of formats and options.

2.3.2.1 Audio Video Interleave (AVI)

Microsoft developed AVI as a multimedia container format that conforms to RIFF (Resource Interchange File Format) specifications. An AVI file may carry audio or visual data in almost any compression scheme, including: Full Frames (Uncompressed), Motion JPEG, DV-NTSC and MPEG-4.

AVI is considered by some to be an outdated format, yet despite its limitations it remains popular among file-sharing communities as a container for “Divx” or “Xvid” encoded video.

2.3.2.2 MPEG-2

The Moving Pictures Expert Group (MPEG) is an ISO working group that is responsible for defining standards for the coded representation of digital audio and video. MPEG uses a lossy compression schema that sequentially stores changes from one picture and audio frame to the next. Currently, there are three major MPEG video standards, MPEG-1, MPEG-2 and MPEG-4.

The most widely applied MPEG standard is MPEG-2. While MPEG-2 is based on MPEG-1 and is fully backward compatible, it produces much higher quality video and sound files. It has become a dominant format for the transmission of broadcast digital video and is the standard video compression for the DVD disc format.

2.3.2.3 Moving Pictures Expert Group (MPEG-4)

MPEG-4 was introduced in 1998 and is built on the MPEG-1, MPEG-2 and Apple Quicktime standards, that are all in turn based upon the established ISO/IEC 14496 specification. Initially, MPEG-4 was developed for low bit rate video communications, providing a higher level of video quality at smaller data rates. However, its enhanced encoding efficiency has led to adoption at higher resolutions and even used in camera acquisition file formats.

2.3.2.4 Quicktime MOVie (MOV)

The Quicktime MOV multimedia container file format was developed by Apple Computer to create, play and stream high-quality audio and video files on both Macintosh and Windows based computers. The International Organization for Standardization chose the Quicktime framework as the basis for the development of the MPEG-4 standard.

Quicktime containers can carry a video encoded in a wide range of codecs, including: Full Frames (Uncompressed), H.264/AVC, Apple Prores 422, DVCPRO, HDV and DV.

2.3.2.5 Windows Media Video (WMV)

Windows Media Video (WMV) is a compressed video file format for several proprietary codecs developed by Microsoft, based on Microsoft's continued customization of MPEG-4. As of March 2006, an implementation of WMV 9 was officially declared an independent *Society of Motion Picture and Television Engineers* (**SMPTE**) standard, and it is now considered to be a unique independent codec better known as VC-1. VC-1 is one of only two codecs approved for use on Blu-Ray discs.

2.4 Content Category: Still Images

Digital image files are made up of either raster ("bit-mapped") images or vector ("object-oriented") images.

A raster image is comprised of bits of information representing uniquely valued pixels in the form of a grid. Image resolution is measured by pixels per inch (PPI); however the printing abbreviation DPI (dots per inch) is also commonly used to describe image resolution. All digital photographs, regardless of file type, are raster images.

150 PPI

50 PPI

The more pixels there are in relation to the area, the higher the resolution. The higher the resolution, the sharper the image is and the larger the file. The picture to the left has a ppi of 150; the picture below has a resolution of 50 ppi.

Digital image resolution is greatly misunderstood. Digital images themselves have no size other than the number of pixels they contain. The image only has real dimensions (inches or cm) when it is in an analogue form before digitization, or after it has been printed.

There are two basic measures for digital imagery characteristics:

- Spatial resolution – capturing detail (PPI) and,
- Tonal resolution – colour, bit depth and dynamic range.

Generally, the higher the PPI and the larger the bit depth, the more accurate the image will be to its original colour. Black and white images are not characterized by colour resolution. They are comprised of brightness values that represent 256 different shades of gray.

Colour Depth

The number of colours available in a digital image is determined by the number of bits assigned to each pixel. The more bits per pixel, the more colours can be displayed.

Colour Depth	Number of Colours Visible
1 bit (monochrome)	2
4 bit	16
8 bit (indexed colour)	256
24 bit (true colour)	16,777,216

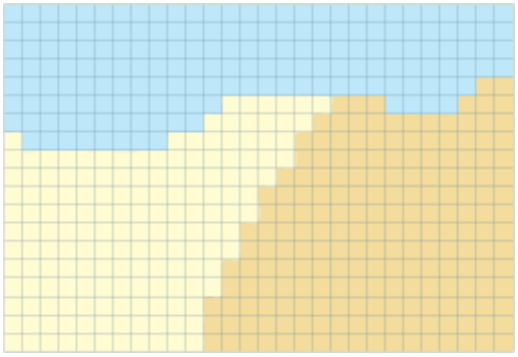
Common "colour resolutions" are 1 bit per pixel, for solid black-and-white nonrealistic images; 8 bits per pixel for grayscale images, nonrealistic colour images, and coarse realistic images; and 24 bits per pixel, for "photographic quality" realistic images. 48 bits per pixel is in increasing use for ultrahigh quality images.

Grayscale images have a maximum colour depth of 8 bits. This is because when defining shades of gray in terms of RGB, each of the 3 red, green and blue components must be equal (i.e. R=192 G=192 B=192, or R=128 G=128 B=128). Since these three components must be equal, there are only 256 possible combinations, which equals 8 bits of colour.

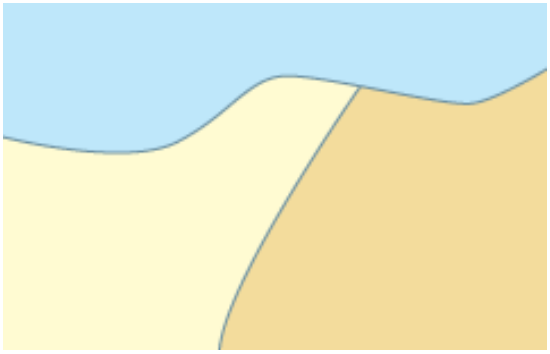
Indexed colour images are limited to a maximum of 256 colours (8-bit), which can be any 256 colours from the set of 16.7 million 24 bit colours. Each image file contains it's own palette which provides a reference index number used by the computer to identify each colour.

Vector graphics (or images), on the other hand, use the geometry of points, lines, curves, and polygons to represent images. This format is commonly used in graphic design, architectural drawings, engineering drawings, and Geographic Information Systems (GIS).

Because their parameters can be modified, vector images can be moved, scaled, rotated, filled, etc., all without degradation of the drawing. Vector[32] file size is directly proportional to the complexity of the image, but is typically smaller than an equivalent raster image file.



Raster representation



Vector representation

2.4.1 Recommended

2.4.1.1 Joint Photographic Experts Group (JPEG)

JPEG refers to both a compression and a file format. The JPEG compression is a standardized lossy image compression designed for compressing full-colour and grayscale images. The International Organization for Standardizaton (ISO) standardized the JPEG compression format in 1990.

It uses a 24-bit colour depth, and the compression is designed to exploit the fact that humans perceive small colour changes less accurately than small changes in brightness.

JPEG works well for photographs and artwork, but does not accurately represent lettering, cartoons or line drawings. This algorithm can be used to compress data within several different file formats, including JFIF, TIFF, or PDF to name a few.

The JPEG file fomat is **JFIF** (**JPEG Image File Format**), but uses the JPG extension. When most people refer to JPEG, JFIF is the file format to which they are referring. JFIF is fully compliant with the JPEG standard.

[32] Because all commonly-used computer monitors display pixels, true vector representation is impossible. When viewing vector images, the monitor translates vector representations of the image to a high resolution raster to view.

2.4.1.2 Joint Photographic Experts Group JPEG 2000 (JP2)

JPEG 2000, Part 1 (the core system, .jp2), is an open, published international ISO standard (ISO/IEC 15444-1). It offers both lossless and lossy compression and provides superior image quality with more efficient compression than JPEG.

JPEG 2000, Part 6 (for mixed raster content, .jpm), is also a published international ISO standard (ISO/IEC 15444-6). It is aimed at compressing scanned colour documents containing both bi-tonal elements as well as images.

Compared to JPEG (which is limited strictly to the RGB colour space), JPEG 2000 compression supports grayscale, RGB and CMYK colour models in up to16-bit colour; in addition to alpha and spot color channels. It can also contain XML compliant metadata.

It should be noted that Microsoft is promoting its own file format which is reported to be comparable to JPEG 2000; HD Photo (.hdp, .wdp) (ISO/IEC 29199-2 (known as JPEG XR)). HD Photo is the new name for Windows Media Photo, and is the native image file format in the Windows Vista operating system.

2.4.1.3 Portable Network Graphics (PNG)

PNG is an open lossless extensible file format that was designed to provide a patent-free, high quality replacement for the GIF and TIFF file formats. It supports the capability of storing up to 16-bits (gray-scale) or 48-bits (truecolour) per pixel, and up to 16-bits of alpha data. It also handles the progressive display of image data and the storage of gamma, transparency and metadata.

2.4.1.4 Tagged Image File Format (TIFF)

TIFF is a bitmapped image format developed by Aldus (now part of Adobe) in 1986. TIFF files may or may not be compressed, and are extensible and portable. They do not favour a particular computer operating system, compiler or processor.

Baseline TIFF images support monochrome, gray-scale, palette (i.e., indexed), and RGB (i.e., true colour) colour spaces. A common extension of TIFF also allows for CMYK images.

TIFF Uncompressed baseline[33] RGB TIFF v.6.0 is LAC's preferred standard for describing and storing raster image data from scanners, faxes and digital photography applications, however because this format can come in a range of types, it is important to specifically note that our preference is for **v.6.0 uncompressed baseline RGB TIFF** files.

The TIFF format uses a 32-bit colour depth, and is therefore limited to a maximum file size of 4 gigabytes.

2.4.1.5 TIFF – GeoTIFF

GeoTIFF files are TIFF images that have geographic coordinate data embedded. The GeoTIFF specification defines a set of TIFF tags which describe cartographic information associated with TIFF imagery including projections[34], coordinate systems[35], ellipsoids[36], datums[37], and anything else required to establish the spatial reference of an image.

[33] For baseline specs see http://www.awaresystems.be/imaging/tiff/tifftags/baseline.html
[34] http://en.wikipedia.org/wiki/Projection
[35] http://en.wikipedia.org/wiki/Coordinate_system
[36] http://en.wikipedia.org/wiki/Ellipsoid
[37] http://en.wikipedia.org/wiki/Datum

GeoTIFF files make use of a public tag structure that is platform independent, and is fully compliant with TIFF 6.0. Image content includes satellite imaging systems, aerial photography, scanned maps, digital elevation models, or the results of geographic analyses.

GeoTIFF files are LAC's preferred format for the transfer of geographically referenced maps in raster format.

2.4.2 Acceptable for transfer

2.4.2.1 Digital Imaging and COmmunications in Medicine (DICOM v. 3.0)

DICOM is a universal image format standard used by virtually all modern ultrasound devices, X-ray photography systems and computer tomographs (CT scans) for storage and transmission of medical images.

DICOM standards are designed to achieve compatibility and improve workflow efficiency between imaging and other information systems in healthcare environments, and are maintained by the DICOM Standards Committee.

The DICOM file format is supported by special software shipped together with corresponding medical equipment, however the DICOM standard is also supported by most devices which allow data exchange regardless of equipment or examination type.

2.4.2.2 Encapsulated PostScript (EPS)

An EPS file is a Post Script language program describing the appearance of a single page, and may contain any combination of text, graphics and images.

LAC receives EPS files from Canada Post, as it is their most common file format used in relation to stamp design.

2.4.2.3 Graphics Interchange Format (GIF)

GIF is primarily an exchange and storage format, although it is supported by many applications. CompuServe released GIF in 1987 as a free and open specification for the storage of raster imagery and to facilitate the exchange of digital imagery between different computer platforms and operating systems. GIF images are lossless and compressed using the LZW scheme.

GIF files are limited to 256-bit images, so the format is more suited to monochrome logos and graphics than for colour photographs.

2.5 Content Category: Web Archiving

Web archiving is the process of collecting portions of the World Wide Web and ensuring the collection is preserved in an archive, such as an archive site, for future researchers, historians, and the public. Due to the massive size of the Web, web archivists typically employ web crawlers for automated collection. The largest web archiving organization based on a crawling approach is the Internet Archive which strives to maintain an archive of the entire Web.

2.5.1 Recommended

2.5.1.1 Internet ARChive Format (ARC)

This format was created by the Internet Archive as a method for combining multiple digital resources into a single archival file. It is used to store 'web crawls' as sequences of content blocks harvested from the World Wide Web.

2.5.1.2 Web ARChive Format (WARC)

WARC is the next generation of the ARC format. WARC generalizes the older format to better support the harvesting, access, and exchange needs of archiving organizations.

Besides the primary content currently recorded, the revision accommodates related secondary content, such as assigned metadata, abbreviated duplicate detection events, and later-date transformations.

2.5.2 Acceptable for transfer

Not applicable.

2.6 Content Category: Structured Data - Databases

Structured data is anything that has an enforced composition to atomic data types. Structured data is managed by technology that allows for querying and reporting against predetermined data types and understood relationships. Structured data typically refers to data that resides in fixed fields within a record or file, typically stored in relational databases and statistical analysis tools.

2.6.1 Recommended

2.6.1.1 Software Independent Archiving of Relational Databases (SIARD)

The Software Independent Archiving of Relational Databases (SIARD) format was developed by the Swiss Federal Archives (SFA)[38] in response to the lack of a standardized archiving format for database content: a situation which has often rendered the preservation (i.e. archiving) task highly complex.

The SIARD format is based on internationally accepted standards. SIARD treats the most common type of relational databases. It enables structure (schemas, tables, etc.) and content of any give relational database to be stored in a simple XML coding. It can handle databases from varied provenances (e.g., MS-Access, Oracle and MS-SQL).

A SIARD archive consists of a content file (content.xml) and a metadata file (metadata.xml) which includes metadata from all levels of the database. Both files are stored in a single uncompressed ZIP-container.

SIARD is based on ISO standards (SQL:199 and XML 1.0). The use of these standardized code assures long-term preservation of databases archived in the SIARD format.

2.6.1.2 Text: Delimited Flat File with Data Description

Tabular data from legacy databases applications may be transferred to LAC in an acceptable ASCII, EBCDIC or Unicode delimited flat file format (e.g., comma separated values (.csv) file). A flat file contains a sequentially arranged set of computer records that must be delimited with an end-of-record marker.

Computer records are composed of a common logical grouping of data fields, which must contain an end-of-field delimiter for variable length records. Flat files are commonly used to transfer and import data files between users who do not use compatible software applications.

At a minimum, the data definition language (DDL) for the database will need to be transferred along with the flat file representation of the database. Additional information related to the interpretation of the data (such a, physical/logical/conceptual models (i.e., entity relationship diagrams), data dictionaries) should also be transferred in a suitable format (e.g., see textual file formats).

2.6.2 Acceptable for transfer

2.6.2.1 dBase Format (DBF)

The dBase file format is widely used for the transfer of files between databases. The format was originally created for dBase database applications. The file header contains information about the record and is encoded in binary, while the record itself is encoded in ASCII.

[38] See ICA2008_Comment_SIARD.pdf (available from PLANETS (http://www.planets-project.eu/)) and SIARD+Format_en.pdf (available from SFA (http://www.bar.admin.ch/)).

2.7 Content Category: Structured Data - Statistical and Qualitative Analysis Data[39]

Much data used in research (particularly in the social and health sciences domain) is stored in statistical and qualitative data analysis tools, such as, SAS, SPSS, Stata, Atlas-ti, QDAMiner, Nvivo, and MAxQDA. These proprietary formats present challenges from a curation, preservation and data exchangeability perspective.

Data in these domains are typically classified as "microdata", representing the lowest level of data collection (which can include personal information), and "aggregate data", representing "microdata" summarized along one or more specific dimensions (such as time, geography, social structure).

In the area of social and health sciences, two major standards have evolved:

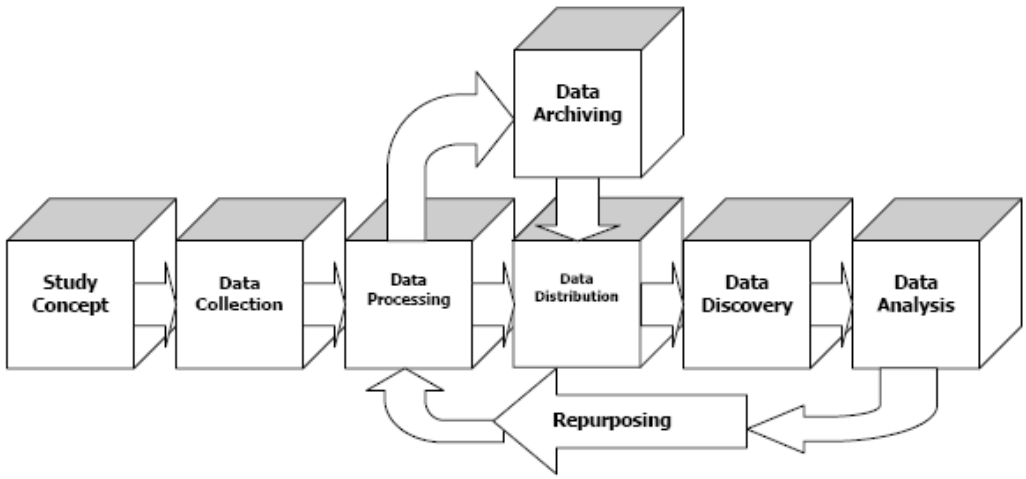
- The Data Documentation Initiative (DDI) for "microdata"; and
- The Statistical Data and Metadata Exchange (SDMX) initiative for "aggregate data" (typically used by major time series data publishers).

DDI is also an underlying standard of the Data Exchange and Conversion Utilities and Tools (DExT) initiative at the UK Data Archive at the University of Essex, for survey and structure data.

2.7.1 Recommended

2.7.1.1 Data Documentation Initiative (DDI)

The Data Documentation Initiative (DDI) is an XML specification for social science metadata that is being developed by an international group called the DDI Alliance[40]. DDI Version 3.0 introduces the capability to document the rich complexity of social science data across its life course as reflected in the Combined Life Cycle Model (as per diagram below)[41].



[39] Information for sections 2.7 and 2.8 is based on a teleconference discussion with Chuck Humphrey, Data Library Coordinator, Libraries, University of Alberta.
[40] http://www.ddialliance.org/org/index.html
[41] Source: DDI_3.0_Part_I_Overview.pdf available from http://www.icpsr.umich.edu/DDI/ddi3/index.html#ddi2.

DDI provides the conceptual model, and the XML Schemas and DTDs which are derived from it. This is a common approach to the standardization of XML vocabularies, and one which provides many benefits to users: the vocabulary itself becomes more consistent and comprehensible, and the conceptual model can prove a valuable asset to developers of applications which need to support the standard, as many tools now allow for XML binding directly from a model expressed in the Universal Modeling Language (UML) or its derivatives[42].

2.7.1.2 Data Exchange and Conversion Utilities and Tools (DExT)

The UK Data Archive at the University of Essex has developed an XML schema and supporting tools to address these issues: the Data Exchange and Conversion Utilities and Tools (DExT). The DExT project builds on existing standards and has constructed a new dedicated schema called QuDEx. This schema enables the transformation from CAQDAS (Computer Assisted Qualitative Data AnalysiS) packages to an open format and can represent annotated textual and multimedia data[43].

LAC will assess the DExT initiative to determine its applicability to content to be submitted to LAC for preservation.

2.7.1.3 Statistical Data and Metadata Exchange (SDMX)

The Statistical Data and Metadata Exchange (SDMX) initiative (http://www.sdmx.org) sets standards that can facilitate the exchange of statistical data and metadata using modern information technology, with an emphasis on aggregated data.

The Version 1.0 specification of the technical standards has been approved by the International Organization for Standardization (ISO) as a Technical Specification (ISO/TS 17369: 2005 SDMX).

The Version 2.0 specification (November 2005) broadens the framework to support wider coverage of metadata exchange as well as a more fully articulated architecture for data and metadata exchange. Steps are also being taken to bring this work forward within the context of ISO, assuring that SDMX technical standards build on other recognized standards and providing the basis for interoperability with them.

LAC will assess the SDMX initiative to determine its applicability to content to be submitted to LAC for preservation.

2.7.1.4 Text: Delimited Flat File with Variable Description

Tabular data from CAQDAS applications may be transferred to LAC in an acceptable ASCII, EBCDIC or Unicode delimited flat file format (e.g., comma separated values (.csv) file)[44].

At a minimum, variable descriptive information must accompany the delimited flat file (i.e., the variables, their characteristics and their meaning need to be provided). Ideally, additional information related to the interpretation of the data should also be transferred in a suitable format (e.g., see textual file formats).

[42] See http://www.icpsr.umich.edu/DDI/ddi3/index.html#ddi2
[43] http://www.data-archive.ac.uk/dext/about/introduction.asp
[44] See Section 2.7.1.4 (first two paragraphs) for additional information on ".csv" files.

2.7.2 Acceptable for transfer

2.7.2.1 SAS[45]

SAS is an integrated system of software products provided by the SAS Institute that enables the programmer to perform:

- data entry, retrieval, management, and mining
- report writing and graphics
- statistical analysis
- business planning, forecasting, and decision support
- operations research and project management
- quality improvement
- applications development
- data warehousing (extract, transform, load)
- platform independent and remote computing

In addition, SAS has many business solutions that enable large scale software solutions for areas such as IT management, human resource management, financial management, business intelligence, customer relationship management and more. SAS is widely used in GoC departments for statistical analysis.

2.7.2.2 SPSS[46]

SPSS is a computer program used for statistical analysis. In 2009 SPSS re-branded its software packages as PASW (Predictive Analytics SoftWare). Statistics included in the base software are:

- Descriptive statistics: Cross tabulation, Frequencies, Descriptives, Explore, Descriptive Ratio Statistics
- Bivariate statistics: Means, t-test, ANOVA, Correlation (bivariate, partial, distances), Nonparametric tests
- Prediction for numerical outcomes: Linear regression
- Prediction for identifying groups: Factor analysis, cluster analysis (two-step, K-means, hierarchical), Discriminant

SPSS is also widely used in GoC departments for statistical analysis.

[45] From Wikipedia: http://en.wikipedia.org/wiki/SAS_System.
[46] From wikepedia: http://en.wikipedia.org/wiki/SPSS.

2.8 Content Category: Structured Data - Scientific Data

There are no standards across all domains of scientific data (with the exception of the social and health sciences identified in Section 2.7). Data in the scientific community will be:

- Stored in relational databases;
- Stored in binary files with supporting scientist-developed code to interpret the data;
- Transformed prior to use in databases or by programs (e.g., algorithms applied to raw satellite data collections).

Some domains in the scientific community have begun to develop data formats that enable sharing of the data by scientists. Some examples include: scientific data based on numerical arrays that are stored in scientific data formats such as NetCDF[47] (network common data format), HDF5[48] (hierarchical data format) and FITS[49] (Flexible Image Transport System). The data formats support efficient mechanisms for accessing and manipulating arrays and allow machine-independent data exchange. They also provide mechanisms for storing metadata information such as simulation parameters[50].

The following provides a brief overview of each scientific data format:

- NetCDF is a set of data formats, programming interfaces, and software libraries that help read and write scientific data files for use in geoscience education and research. (NetCDF was developed and is maintained by Unidata, part of the University Corporation for Atmospheric Research (UCAR) Office of Programs (UOP) with funding primarily provided by the National Science Foundation.)
- HDF5 is a file format and library for storing scientific data. HDF5 is a unique technology suite that makes possible the management of extremely large and complex data collections. The HDF5 technology suite includes:
    - A versatile data model that can represent very complex data objects and a wide variety of metadata.
    - A completely portable file format with no limit on the number or size of data objects in the collection.
    - A software library that runs on a range of computational platforms, from laptops to massively parallel systems, and implements a high-level API with C, C++, Fortran 90, and Java interfaces.
    - A rich set of integrated performance features that allow for access time and storage space optimizations.
    - Tools and applications for managing, manipulating, viewing, and analyzing the data in the collection.
    - The HDF5 data model, file format, API, library, and tools are open and distributed without charge.
- Flexible Image Transport System (FITS) is the standard computer data format widely used by astronomers to transport, analyze, and archive scientific data files.

[47] http://www.unidata.ucar.edu/software/netcdf/
[48] http://www.hdfgroup.org/HDF5/
[49] http://fits.gsfc.nasa.gov/
[50] Format identification/definitions from: http://www.nersc.gov/nusers/analytics/sdm/; http://www.unidata.ucar.edu/software/netcdf/docs/; http://www.hdfgroup.org/HDF5/doc/index.html; and http://fits.gsfc.nasa.gov/fits_documentation.html

2.8.1 Recommended

The format for the preservation of scientific data will be at the discretion of the preservation specialist (or conservator); however, it will use an XML-based container format.

Depending on the source of the scientific data, standards adopted for databases, source code, and statistical and qualitative analysis data may be applied.

2.8.2 Acceptable for transfer

Not applicable.

2.9 Content Category: Geospatial

Geospatial data consists of spatially defined geographic data that are analyzed through the use of Geographic Information System (GIS) software, image processing systems, or similar types of modeling software and technology. Geographic and spatial data comes in a variety of formats, some of which can be used in GIS software applications. The most popular formats that are used in the Canadian federal government appear below.

2.9.1 Recommended

2.9.1.1 TC 211 ISO 19115 Geographic Information - Metadata (NAP – Metadata) (North American Profile)[51]

Geospatial data is defined as data with implicit or explicit reference to a location relative to the Earth. This standard establishes the information infrastructure to support the discovery and use of geospatial information and to enable information sharing among departments, with other jurisdictions, and with the private sector.

Geospatial data important to social, economic and cultural well-being is produced or used by federal departments, the provinces, territories, and others. This includes mapping products to support activities such as search and rescue, geospatial intelligence, and fire fighting. Standardization is essential in this context. It allows data from one source to be easily used with those from another source to create a richer and more useful product. The Standard on Geospatial Data adopts measures that have been endorsed by federal departments, provincial and territorial governments, as well as by academic and private sector participants in the Canadian Geospatial Data Infrastructure.

This standard will allow departments to share data and maximize utility of existing mapping and related products. Departments will also be able to exploit commercially available tools and software in common use around the world to discover, access and use geospatial data. This will result in significant efficiencies in the sharing and use of public sector mapping products. More broadly, it will support departments' mandated programs and services, allowing them to address and respond to economic, environmental and societal challenges more effectively.

2.9.2 Acceptable for transfer

2.9.2.1 Canadian Council on Geomatics Interchange Format (CCOGIF)

This standard specifies the format for the exchange of digital spatial data among Canadian survey and mapping agencies. CCOGIF provides a national standard that preserves the accuracy and content of the exchanged information, and is machine and language independent.

2.9.2.2 Digital Elevation Model (DEM)

A DEM data file consists of an array of terrain elevation samples for ground positions at regular intervals. It is used to create 3D graphics that display the slope, aspect and terrain profiles of a given area. The USGS DEM standard was recently altered to conform to the SDTS format.

2.9.2.3 Digital Line Graphics – level 3 (DIG-3)

The DLG standard was originally developed by the U.S. Geological Survey (USGS) as a National Mapping Program (NMP) standard for the digital representation of many of the country's traditional 7.5-

[51] Please refer to http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=16553 for additional information.

minute quadrangle cartographic paper maps. The format was created to define topological (i.e., spatial relationships between the data elements) vector-based line data such as roads, rivers and boundaries.

The DLG format is one of the more efficient and widely recognized data formats used for the distribution of vector data. DLG-3 is gradually being replaced by the Spatial Data Transfer Standard (SDTS) interchange format (see below) in the United States Government.

Preservation: The DLG-3 format will be used as a preservation format only if the geospatial content is provided in this format and is not available in a recommended format.

2.9.2.4 Environmental Systems Research Institute (ESRI) Export Format – (E00)

E00 is an interchange data format that was developed by Environmental Systems Research Institute (ESRI) to enable users to move data into and out of its geographic information system (GIS) software package known as ARC/INFO.

A single E00 file describes a complete ARC/INFO coverage. An E00 file is actually an archive of smaller sub-files. Standard sub-files, which have fixed names and are comprised of a fixed data format that does not change from coverage to coverage. The second includes Info sub-files that contain user-defined attribute information.

Preservation: The E00 format will be used as a preservation format only if the geospatial content is provided in this format and is not available in a recommended format.

2.9.2.5 Environmental Systems Research Institute (ESRI) SHape File Format (SHP)

ESRI introduced the Shapefile to provide GIS users with a simple and effective means to disseminate geospatial information, as an alternative to the E00 export file format.

"While the term "shapefile" is quite common, a "shapefile" is actually a set of several files"[52]. Three individual files are normally mandatory to store the core data that comprises a shapefile. There are a further eight optional files which store primarily index data to improve performance. Each individual file should conform to the MS DOS 8.3 filenaming[53] convention (8 character filename prefix, fullstop, 3 character filename suffix such as shapefil.shp) in order to be compatible with past applications that handle shapefiles, though many recent software applications work fine with longer file names. For this same reason, all files should be located in the same folder.

Mandatory files :

- .shp — shape format; the feature geometry itself
- .shx — shape index format; a positional index of the feature geometry to allow seeking forwards and backwards quickly
- .dbf — attribute format; columnar attributes for each shape, in dBase III format

Optional files :

- .prj — projection format; the coordinate system and projection information, a plain text file describing the projection using well-known text[54] format
- .sbn and .sbx — a spatial index[55] of the features
- .fbn and .fbx — a spatial index of the features for shapefiles that are read-only
- .ain and .aih — an attribute index of the active fields in a table or a theme's attribute table

[52] http://en.wikipedia.org/wiki/Shapefile
[53] http://en.wikipedia.org/wiki/8.3_filename
[54] http://en.wikipedia.org/wiki/Well-known_text
[55] http://en.wikipedia.org/wiki/Spatial_index

- .ixs — a geocoding index for read-write shapefiles
- .mxs — a geocoding index for read-write shapefiles (ODB format)
- .atx — an attribute index for the .dbf file in the form of *shapefile.columnname*.atx (ArcGIS 8 and later)
- .shp.xml — metadata in XML format
- .cpg — used to specify the code page[56] (only for .dbf) for identifying the character encoding to be used.

The Shapefile is the de facto standard for geospatial data exchange and desktop GIS applications. The openly published Shapefile format is based upon a non-proprietary geospatial data structure.

Preservation: The SHP format will be used as a preservation format only if the geospatial content is provided in this format and is not available in a recommended format.

2.9.2.6 International Hydrographic Organization (IHO) S-57, Edition 3.1

The S-57: IHO Transfer Standard for Digital Hydrographic Data, Edition 3.1 was officially made available in November 2000. IHO S-57 is a standard that describes a data format for the transfer of digital hydrographic data. The standard is based on the ISO/IEC 8211:1994 specification for a data descriptive file for information exchange.

The interchange standard is a media and content independent standard which allows users to name and describe data fields containing both character and binary data. Data structures in the S-57 format can be encoded in either binary or ASCII. The data structure is a tree with a finite number of levels: each file comprises records, each record fields, each field sub-fields.

Preservation: The S-57 format will be used as a preservation format only if the geospatial content is provided in this format and is not available in a recommended format.

[56] http://en.wikipedia.org/wiki/Code_page

2.10 Content Category: Computer Aided Design (CAD) - Technical drawing

2.10.1 Recommended

2.10.1.1 Drawing Interchange File Format/Data eXchange Format (DXF)[57]

The DXF format is a tagged data representation of all the information contained in an AutoCAD® drawing file. DXF files enable the interchange of drawings between different CAD software applications. DXF was originally released in 1982 as part of AutoCAD 1.0, and the specification was intended to provide programmers the data model for the AutoCad native file format known as DWG. DXF files can be in either ASCII or binary formats. LAC supports the ASCII format.

2.10.2 Acceptable for transfer

2.10.2.1 Computer Graphics Metafile (CGM)

Although CGM is not widely supported and has been supplanted by other formats, it is still prevalent in engineering, aviation, and other technical applications. CGM is a file format for 2D vector graphics, raster graphics, and text, and is defined by ISO/IEC 8632.

[57] DXF is an example of an industry *de facto* standard.

2.11 Content Category: Computer Aided Design (CAD) – Computer-Aided Software Engineering (CASE) [58]

Computer-Aided Software Engineering (CASE), in the field of Software Engineering is the scientific application of a set of tools and methods to a software development process which is meant to result in high-quality, defect-free, and maintainable software products. It also refers to methods for the development of information systems together with automated tools that can be used in the software development process.

CASE information may be transferred to LAC for preservation as part of database and/or business logic information transfers to further assist in their understanding and reuse, along with graphical representations of the CASE models (see 2.4 Content Category: Still Images) and/or descriptive information provided in the appropriate formats (see 2.1 Content Category: Text).

Where CASE information is to be included as part of a transfer, the following format(s) will apply.

2.11.1 Recommended

2.11.1.1 XML Metadata Interchange (XMI) [59]

The **XML Metadata Interchange** (XMI) is an Object Management Group (OMG)[60] standard for exchanging metadata information via Extensible Markup Language (XML). It can be used for any metadata whose metamodel can be expressed in Meta-Object Facility (MOF)[61]. The most common use of XMI is as an interchange format for UML[62] models, although it can also be used for serialization of models of other languages (metamodels).

XMI integrates four industry standards:

- XML - eXtensible Markup Language, a W3C standard.
- UML - Unified Modeling Language, an OMG modeling standard.
- MOF - Meta Object Facility, an OMG language for specifying metamodels.
- MOF Mapping to XMI.

The integration of these four standards into XMI allows tool developers of distributed systems to share object models and other metadata.

XMI is an international standard:

- ISO/IEC 19503:2005 Information technology -- XML Metadata Interchange (XMI).

2.11.2 Acceptable for transfer

Not applicable.

[58] From Wikipedia: http://en.wikipedia.org/wiki/Computer-aided_software_engineering.
[59] From Wikipedia: http://en.wikipedia.org/wiki/XMI.
[60] http://en.wikipedia.org/wiki/Object_Management_Group
[61] http://en.wikipedia.org/wiki/Meta-Object_Facility
[62] http://en.wikipedia.org/wiki/Unified_Modeling_Language

2.12 Content Category: Source Code and Scripts

Source code (or scripts) is contained in an unstructured plain text file that has collections of statements or declarations written in some human- readable computer programming language which special programs (compilers or interpreters) understand to derive structures and semantics. Depending on the programming language used, usefulness of code-level comments and other factors, some source code may be more difficult to read and/or understand.

Source code is sometimes protected using a mechanism called "obfuscation" which removes semantic information and further modifies the source code to make it difficult to recompile and removes all practical human readability. Obfuscated code[63] is of very little use and reduces the value of the preserved assets.

Source code may have one of hundreds of file extensions based on the programming language, platform and era it was developed under. It is also possible that certain extensions are already obsolete and it may even be possible that certain files do not carry any file extensions at all.

Source code rarely is contained within a single file but rather in multiple files with embedded dependencies (sometimes packaged in a hierarchical structure). Source code, along with as much contextualization information (i.e. metadata) as possible in regards to the program (comments, technical environment and platform description, functional information, etc.) should be submitted as a unit/package and all material should be preserved alongside. Containers such as METS or ZIP should be used to ensure relational integrity and facilitate their overall management.

From a digital preservation standpoint, source code is only as useful as the overall context information that accompanies it. The only reasonable intent for preserving source code should be to understand parts of important business logic, functionality and overall architecture of a given program in relation to one or more intellectual asset. Due to the complexity of modern development environments (multiple tools, compiler versioning, external and internal dependencies, libraries, etc.), it is not realistic to preserve source code with the intention of re-building it in the near or distant future.

Therefore it is recommended to ingest and preserve source code on an exception basis and follow proper process to assess desirability and usability of the material: as a result, the decision to accept and preserve source code and scripts will be at the discretion of LAC, or in accordance with pre-established transfer agreements

2.12.1 Recommended

2.12.1.1 Container-based Format (XML)

The format for the preservation of source code and scripts will be at the discretion of the preservation specialist (or conservator), however, it will use an XML-based container format.

2.12.2 Acceptable for transfer

2.12.2.1 Text

Transfers of source code and scripts will be accepted in standard text-based files.

[63] Obfuscated code is source or machine code that has been made difficult to understand. Programmers may deliberately obfuscate code to conceal its purpose (a form of security through obscurity), to deter reverse engineering, or as a puzzle or recreational challenge for readers. Programs known as obfuscators transform human-readable code into obfuscated code using various techniques (see http://en.wikipedia.org/wiki/Obfuscated_code).

2.13 Other Content Format Considerations and Guidelines

This section identifies guidelines for organizations to consider when planning the transfer of digital content to LAC. Specifically, this section provides:

- Guidelines for the use of backup functionality (for operational considerations) versus the use of "archiving" software/formats;
- Guidelines for including executable code in support of "archiving" digital content.

2.13.1 Backup and Archiving Guidelines

Business systems for managing information resources of business value generally do not provide for the ongoing, long-term preservation of the object. Within some specialized communities of practice there is a clear distinction between backup or archiving systems.

In the Information Technology community the use of the term backup refers to a complex series of scheduled processes that make copies of information resources (including the operating system and file systems) to a portable media so that they (i.e., the media) can be used to restore the original operating environment and file system information after a system crash where there is any type of loss. The primary purpose of backups is to reinstate a set of information resources after a disaster or as a result of data loss through deletion or corruption. Backups are not usually portable between operating systems and different versions of the same system.

Archiving applications, such as LAC's Trusted Digital Repository, are designed and implemented with authenticity, portability / interoperability and information reliability being the key considerations. This includes the design of the technological solution and its practical implementation that ensures the ability to demonstrate that authenticity, reliability and usability of electronically stored information resources (records / data / information) are addressed. These types of systems are usually aligned with the concepts of digital preservation based on open and international standards and frameworks such as OAIS, Open Archival Information System Reference Model and TRAC, Trusted Digital Repository Audit Checklist. These digital preservation systems include technologies supporting integrity services such as format validation and verification, virus checking, migration and emulation services, persistent identification, replication services, preservation planning and preservation metadata.

Depending on the operating system an archive file can be composed of one or more files along with associated metadata that includes source volume, logical format information, file directory information, error detection (integrity checks), logic to detect and correct errors (recovery record), and may employ some type of lossless compression.

2.13.2 Object Code (Executables)

Object code, or executable files, causes a computer "to perform indicated tasks according to encoded instructions", as opposed to a file that only contains data (or data and format codes, such as word processing documents). Files that contain instructions for an interpreter or virtual machine may be considered executables, but are more specifically called scripts or "bytecode". Once they are compiled, executables are also called "binaries"[64] in contrast to the program's source code.

An executable file is created to run on very specific environment settings. Operating System dependent by nature, they may also rely on the presence of a core system library, third-party libraries, drivers, CPU architecture, hardware and several configuration elements of the original platform. Therefore, the preservation of executable files represents a significant challenge and may not be possible without the

[64] "Binary" code is "machine readable" whereas source code is humanly readable.

implementation of a digital preservation strategy based on *virtualization/emulation*[65]. Most memory institutions, including Library and Archives Canada have instead opted for the *migration/refreshment*[66] approach. This strategy offers several advantages but makes the preservation of executables (with the intention of re-executing them in the future) impossible.

Depending on the original target operating system, executables may have one of several file extensions (.exe, .com, etc.) or may not have any file extensions but rather metadata elements that characterize the file as having "execute permission", such as the case in UNIX environments.

Exceptionally, for instances where the specific executables or programs "render" or "give access to" important information that is considered of preservation value, a dedicated effort/project will proceed to the extraction of those assets and, if necessary, convert them to contemporary recommended preservation formats[67]. If the executables are transferred to LAC via one of the ingest channels, this strategy will require that all dependencies are also made available to LAC staff to allow proper execution[68].

LAC will not preserve transferred executables in the Trusted Digital Repository (TDR) [69], however, the executables that are deemed to be required to access the content being transferred will be retained as part of the preservation environment in a preservation software toolkit library.

[65] http://en.wikipedia.org/wiki/Digital_preservation#Emulation
[66] http://en.wikipedia.org/wiki/Digital_preservation#Refreshing
[67] The content to be converted may fall into any one or more content categories, such as, text, audio, video, structure data, etc.
[68] Dependencies include relationships to other executable code (such as, specialized libraries).
[69] A notable execption to the preservation of executables is in the area of "harvested" websites that are conatined in WARC files (see Section 2.5.1.2 for a definition). "Harvested" websites (website content collected using specialized tools such as web crawlers) may contain executable code: these sites may be subject to a virtualizaton/emulation strategy in lieu of a migration strategy. Developments in the broader preservation community are being monitored by LAC, to determine an appropriate long term strategy for preserving "harvested" web sites (e.g., retaining the executables and the software necessary to run the executables within the WARC itself).

3 Bibliography

1. About.com, *Graphics Software*, Nov 2007
   http://graphicssoft.about.com/library/glossary/bldefmetafile.htm
2. Adobe Developers Association. *TIFF Revision 6.0.* Mountain View, CA. (1992).
   http://partners.adobe.com/asn/developer/pdfs/tn/TIFF6.pdf
3. Adobe Systems Inc., *Whitepaper-- PDF as a Standard for Archiving*, 2002.
   http://www.adobe.com/enterprise/pdfs/pdfarchiving.pdf
4. Aitken, Peter. *I Never Metafile I Didn't Like*, 3 November 2006.
   http://www.devsource.com/article2/0,1895,2050820,00.asp
5. Audio Engineering Society,
   http://www.aes.org/publications/downloadDocument.cfm?accessID=14703162000122117
6. Aware Systems, *The BigTIFF File Format Proposal*, 2007
   http://www.awaresystems.be/imaging/tiff/bigtiff.html
7. Bachmann, Erik. *Xbase Data File (*dbf),* 18 Aug 2007
   www.clicketyclick.dk/databases/xbase/format/dbf.html#DBF_STRUCT
8. Biblioscape, *Rich Text Format (RTF) Version 1.5 Specification*, 2007.
   http://www.biblioscape.com/rtf15_spec.htm#Heading1
9. Broadcastpapers.com,
   http://www.broadcastpapers.com/sigdis/Snell&WilcoxMXF01.htm
10. Broadcastpapers.com, *The online library for technical & business whitepapers*, November 2007.
    http://www.broadcastpapers.com/whitepapers.cfm?objid=2
11. Brooke, Simon. *XML Representation of Nautical Chart Data*. Scaffie Ltd. Auchencairn, Scotland. Retrieved June 003 from:
    http://www.weft.co.uk/library/xmlchart/documentation/overview-summary.html
12. Brooks, Alfred A. *Overview – ISO/IEC 8211:1994.* 1996
13. Brown, Adrian. *Digital Preservation Guidance Note: 4, Graphics File Formats,* The National Archives, United Kingdom, 9 July 2003.
    http://www.nationalarchives.gov.uk/documents/graphic_file_formats.pdf
14. Brown, David, et. al. *Management and Preservation of Geospatial Data.* Report written for the Ad-Hoc Committee on Archiving and Preserving Geospatial Data, GeoConnections, Policy Advisory Network Node, July. 2003.
15. Brown, David. *Guidelines for Computer File Types, Interchange Formats and Information Standards*, Library and Archives Canada, Version 1, 28 June 2004.
16. California Digital Library. *Digital Image Format Standards.* (2001).
    http://www.cdlib.org/about/publications/CDLImageStd-2001.pdf

17. Canadian Council on Geomatics, *Standard File Exchange Format for Digital Spatial Data, Version 2.3*, October 1994.

www.cits.rncan.gc.ca/fich_ext/1/text/products/ntdb/ccogif.pdf

18. Canadian Standards Association, *Beyond Canada: health informatics around the globe*, 2007.

http://www.csa.ca/standards/health_care/newsletter/default.asp?load=news5&language=english

19. CARIS, *Geomatics Software Solutions*, 2007

www.caris.com

20. Carson, Steve. *Basic Image Interchange Format (BIIF)*, GSC Associates Inc., 2007.

http://www.acm.org/tsc/biif.html

21. Committee on Earth Observing Satellites (CEOS). *The CEOS Working Group on Information Systems and Services (WGISS)*, 2007

http://wgiss.ceos.org/ceos.htm

22. CoOL, a project of the Preservation Department of Stanford University Libraries and Academic Information Resources, Aldus/Microsoft Technical Memorandum, *TIFF Revision 5.0*, November 2007.

http://palimpsest.stanford.edu/bytopic/imaging/std/tiff5.html

23. Cudlip, W. *Guidelines on Standard Formats and Data Description Languages Version 1.0.* Committee on Earth Observation Satellites. 1998.

24. Curtin, Dennis P. *Sensors, Pixels and Image Sizes*, 2007.

http://www.shortcourses.com/pixels/colourdepth.htm

25. Data Compression Dogma, *What is the state of the art in lossless image compression?*, 26 October 2006

http://datacompression.dogma.net/index.php?title=FAQ:What_is_the_state_of_the_art_in_lossless_image_compression%3F#Benchmarks

26. Data Documentation Initiative (DDI) Alliance, July 2009.

DDI_3.0_Part_I_Overview.pdf available from
http://www.icpsr.umich.edu/DDI/ddi3/index.html#ddi2

27. Developer Shed, *Bringing Yourself Up to Speed with AAC, MP3, and Digital Audio,* November 2007.

http://www.devhardware.com/c/a/Software/Bringing-Yourself-Up-to-Speed-with-AAC-MP3-and-Digital-Audio/3/

28. Doughty, Mike. Mike's Sketchpad, *A Little Bit More About Color*, 2007.

http://www.sketchpad.net/basics6.htm

29. Durham University, *Accessibility Glossary*, 2007.

http://www.dur.ac.uk/its/services/web/accessibility/glossary/

30. ER Mapper, *Geospatial Imagery Solutions Forum*, 2007.

http://forum.ermapper.com/viewforum.php?f=11

31. European Broadcast Union (EBU), *Broadcast Wave Format (BWF) User Guide*, 11 May 2007.

http://www.ebu.ch/en/technical/publications/userguides/bwf_user_guide.php

32. Federal Ministry of the Interior, *SAGA: Standards and Architectures for eGovernment Applications, KBSt Publication Series, Volume 56,* February. Berlin, AG. (2003).

http://www.kbst.bund.de/saga

33. FileFormatInfo, Encyclopedia of Graphics File Formats, *Microsoft Windows Metafile File Format Summary.*

http://www.fileformat.info/format/wmf/

34. File format evaluation, July 2009.

Gillesse *et al* 2008; Rauch, Carl *et al*. 'File-Formats for Preservation: Evaluating the Long-Term Stability of File-Formats." Proceedings ELPUB2007 Conference on Electronic Publishing : Vienna, Austria , 2007. http://elpub.scix.net/data/works/att/122_elpub2007.content.pdf

National Archives (UK). "Selecting File Formats for Long-Term Preservation." (2003). http://www.nationalarchives.gov.uk/documents/selecting_file_formats.rtf Library of Congress. "Sustainability of Digital Formats: Planning for Library of Congress Collections." (2007). http://www.digitalpreservation.gov/formats/sustain/sustain.shtml

35. Future Publishing Limited, *Video Codecs*, November 2007.

http://www.pcanswers.co.uk/tutorials/default.asp?pagetypeid=2&articleid=30900&subsectionid=781&subsubsectionid

36. *GIF Graphics Interchange Format.* CompuServe, Inc. Columbus, Oh. (1987). http://www.w3.org/Graphics/GIF/spec-gif87.txt

37. Glagola, Michael J. mglagola@cox.net' *Digital Image Characteristics; Understanding Pixels, Dots, Samples & Viewing*, Washington Apple Pi iLife SIG, January 2005

http://www.wap.org/imovie/DigitalImagingPresentation.pdf

38. Grand, Mark. *MIME Overview*, 26 Oct 1993

http://mgrand.home.mindspring.com/mime.html

39. Hamilton, Eric. *JPEG File Interchange Format Version 1.02.* C-Cube Microsystems. Milpitas, Ca. (1992).

http://www.w3.org/Graphics/JPEG/jfif3.pdf

40. International Business Machines Corp., *IBM Character Data Representation Architecture, Reference and Registry, SC09-2190-00,* December 1996.

41. International Digital Publishing Forum (IDPF), July 2009.

EPUB 101.pdf available from http://www.idpf.org/

42. International Telecommunications Union, *Telecommunication Standardization Sector (ITU-T),* 2007

http://www.itu.int/publications/sector.aspx?lang=en&sector=2

43. Interoperability Framework Coordination Group. *The HKSARG Interoperability Framework: Version 1.0. G*overnment of the Hong Kong Special Administrative Region Information Technology Services Department. November 2002.

44. International Hydrographic Organization, *IHO Transfer Standard for Digital Hydrographic Data, Publication S-57 Edition 3.1*, Nov 2000

www.iho.shom.fr/publicat/free/files/31Main.pdf

45. International Organization for Standardization, *Coding of Moving Pictures and Audio, MPEG-4 Overview*, March 2002.

http://www.chiariglione.org/mpeg/standards/mpeg-4/mpeg-4.htm

46. International Organization for Standardization. ISO/TC211, *Geographic Information/Geomatics Scope.* (2002).

http://www.isotc211.org/scope.htm#scope

47. International Organization for Standardization, ISO/IEC 8859-1:1998 Information technology -- 8-bit single-byte coded graphic character sets -- Part 1: Latin alphabet No. 1, 2007.

www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=28245&IC1=35&ICS2=40&ICS3

48. ISO/TC171/SC2. *NWI Ballot for Document management – Long-term electronic preservation – Use of PDF (PDF/A).* International Organization for Standardization. Document N 226 E. April. (2003).

http://www.aiim.org/documents/standards/SC2N226.pdf

49. Joint Photographic Experts Group, *JPEG 2000 FAQ.*, 2007.

http://www.jpeg.org/.demo/FAQJpeg2k/index.htm

50. Korpela, Jukka. *A tutorial on character code issues*, 13 July 2007.

http://www.cs.tut.fi/~jkorpela/chars.html

51. Kunze, John. California Digital Library, *WARC: an Archiving Format for the Web*, 22 September 2005

http://www.iwaw.net/05/kunze.pdf

52. Lane, Tom. *JPEG Image Compression FAQ, part ½.* (1999)

http://www.faqs.org/faqs/jpeg-faq/part1/

53. Liam, Quin. *XML Core Working Group Public Page* – Revision 1.24. World Wide Web Consortium. (2003).

http://www.w3.org/XML/Core/#Publications

54. Library and Archives Canada Digital Preservation Policy http://www.collectionscanada.gc.ca/digital-initiatives/012018-2000.01-e.html

55. Library of Congress, *Digital Preservation, Digital Formats, Sound Quality and Functionality Factors*, 07 March 2007.

http://www.digitalpreservation.gov/formats/content/sound_quality.shtml

56. Lim, Mark. *National Archives of Canada: Digital Media Formats Study*. 1514486 Ontario Inc. Contract No. 02011-2-0257. 2003.

57. McGowan, John F. *AVI Overview*. (1999)

http://www.2dreamers.com/tutorials/John%20McGowan%27s%20AVI%20Overview.htm

58. Moving Pictures Experts Group. *The MPEG Home Page*, November 2007.

http://www.mpeg.org/

59. Mpeg***TV***, *The reference site for MPEG*, November 2007.

http://www.mpeg.org/MPEG/

60. NCH Swift Sound, *Audio File Formats*, November 2007.

http://www.nch.com.au/acm/formats.html

61. New Zealand E-government Unit. *New Zealand E-government Interoperability Framework (NZ e-GIF)*. State Services Commission. Version 1.1. July. (2003).

62. NOAA Satellite and Information Service, National Satellite, Data and Information Services (NESDIS), *NOAA Metadata Manager and Repository,* 2007

www.ngdc.noaa.gov/nmmr/public/viewRecord.do?xmlstyle=FGDC&edit=&recuid=1858&recordset=NCDCà

63. OASIS, *Open Document Format for Office Applications (OpenDocument) TC,* 2007.

http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=office

64. Open GIS Consortium Inc. *OpenGIS Geography Markup Language (GML) Implementation Specification*. Document OGC 02-023r4, Version 3.0. editors, Simon Cox, et. al., January 2007.

www.opengis.org/

65. Oracle Corporation, *Image File and Compression Formats,* 2003.

http://www.stanford.edu/dept/itss/docs/oracle/10g/appdev.101/b10829/mm_formats.htm

66. PDF Tools AG, *White Paper – PDF Primer*, 6 October 2005.

http://www.pdf-tools.com/public/downloads/whitepapers/whitepaper-pdfprimer.pdf

67. Pearson, Glenn and Michael Gill, *"An Evaluation of Motion JPEG 2000 for Video Archiving"*, Proc. Archiving 2005 (April 26-29, Washington, D.C.), IS & T (www.imaging.org), pp. 237-243.

http://archive.nlm.nih.gov/pubs/pearson/MJ2_video_archiving.pdf

68. Peterson, Kit A. Digital Conversion Specialist, Prints & Photographs Division, Library of Congress, Washington, D.C. 20540-473, *Introduction to Basic Measures of a Digital Image for Pictorial Collections,* June 2005.

http://www.loc.gov/rr/print/tp/IntroDgtlImage.pdf

69. Radiological Society of North America, *DICOM (Digital Imaging and Communications in Medicine), The Value and Importance of an Imaging Standard,* 2007.

http://www.rsna.org/Technology/DICOM/

70. Roelofs, Greg., *An Open, Extensible Image Format with Lossless Compression*, 6 September 2007.

http://www.libpng.org/pub/png/

71. Ruth, Mike. *GeoTIFF FAQ Version 2.1*. 1999

http://remotesensing.org/geotiff/faq.html

72. Ruth,Mike. *GeoTIFF FAQ Version 2.3*, February, 2005

http://remotesensing.org/geotiff/faq.html#What%20is%20GeoTIFF%20and%20how%20is%20this%20different%20from%20TIFF?

73. Scientific Data Formats, July 2009.

http://www.nersc.gov/nusers/analytics/sdm/; http://www.unidata.ucar.edu/software/netcdf/docs/; http://www.hdfgroup.org/HDF5/doc/index.html; and http://fits.gsfc.nasa.gov/fits_documentation.html

74. Swiss National Archives, Software Independent Archiving of Relational Databases (SIARD), July 2009.

ICA2008_Comment_SIARD.pdf (available from PLANETS (http://www.planets-project.eu/)) and SIARD+Format_en.pdf (available from SFA (http://www.bar.admin.ch/)

75. TeamCom Books, *The MP3 and Internet Audio Handbook, Your Guide to the Digital Music Revolution!,* March 2000.

http://www.teamcombooks.com/mp3handbook/MP3_Handbook.htm

76. Techsmith,

http://fr.techsmith.com/products/studio/tutorials/1104.asp

77. Treasury Board Secretariat (TBS), Government of Canada, Standard on Geospatial Data, July 2009.

http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=16553

78. UK Digital Archives, DeXT, July 2009.

http://www.data-archive.ac.uk/dext/about/introduction.asp

79. Unicode Inc., *The Unicode Standard, Version 3.0*, 2007

www.unicode.org/book/u2.html

80. United Kingdom Office for Library and Information Networking (IKOLN). *NOF-Digitization Technical Standards and Guidelines*. New Opportunities Fund, UKOLN, University of Bath in association with Resource: The Council for Museums, Archives & Libraries. Bath. Version Five: revised March. (2003).

81. United States Geological Survey, Digital Line Graph Standards, 29 Aug 2007

http://rockyweb.cr.usgs.gov/nmpstds/dlgstds.html

82. United States Geological Survey, Digital Elevation Model Standards, 29 Aug 2007

http://rockyweb.cr.usgs.gov/nmpstds/demstds.html

83. United States Geological Survey, *Spatial Data Transfer Standard,* 2007

http://mcmcweb.er.usgs.gov/sdts

84. United States General Service Administration. *Government Without Boundaries: A Management Approach to Intergovernmental Programs.* Office of Intergovernmental Solutions. May. (2002).

85. United States National Snow and Ice Data Center, *Global Digital Sea Ice Data Bank (GDSIDB),* 2003)

http://nsidc.org/noaa/gdsidb/s3development.html

86. University of Connecticut, *The file types*, 12 November 2007

http://www.gifted.uconn.edu/siegle/HonorsSeminar/filetype.html

87. Usdin, B. Tommie et. al. What is SGML? Mulberry Technologies, Inc. Rockville, MD. 1997.

88. Webopedia.com, *CCITT, Comité Consultatif International Téléphonique et Télégraphique*, November 2007.

http://www.webopedia.com/TERM/C/CCITT.html

89. Webopedia.com, *Data Compression*, 2007.

http://www.webopedia.com/TERM/d/data_compression.html

90. Wikipedia, The Free Encyclopedia, *Linear Pulse Code Modulation used in communications (or LPCM)*, Nov 2007

http://en.wikipedia.org/w/index.php?title=LPCM&oldid=88811980>.

91. Wikipedia, The Free Encyclopedia, *Sound quality*, 20 Sept 2007

http://en.wikipedia.org/wiki/Sound_quality

92. Wikipedia, The Free Encyclopedia, *Windows Media Audio (WMA)*, 10 November 2007.

http://en.wikipedia.org/wiki/Windows_Media_Audio

93. Wikipedia, The Free Encyclopedia, *Audio File Format*, 8 Nov 2007.

http://en.wikipedia.org/wiki/Audio_file_format

94. Wikipedia, The Free Encyclopedia, *Document File Format*, 2007

http://en.wikipedia.org/wiki/Document_file_format

95. Wikipedia, The Free Encyclopedia, *GeoTIFF*, 28 August 2007.

http://en.wikipedia.org/wiki/Geotiff

96. Wikipedia, *NetCDF (Network Common Data Form)*, November 9, 2007.

en.wikipedia.org/wiki/Netcdf

97. Wikipedia, The Free Encyclopedia, *Vector graphics*, 6 November 2007.

http://en.wikipedia.org/wiki/Vector_graphics

98. Wikipedia, The Free Encyclopedia, *AIFF*, July 2009.

http://en.wikipedia.org/wiki/AIFF

99. Wikipedia, The Free Encyclopedia, *Digital Video*, July 2009.

http://en.wikipedia.org/wiki/Digital_video

100. Wikipedia, The Free Encyclopedia, *SAS*, July 2009.

http://en.wikipedia.org/wiki/SAS_System

101. Wikipedia, The Free Encyclopedia, *SPSS*, July 2009.

http://en.wikipedia.org/wiki/SPSS

102. Wikipedia, The Free Encyclopedia, *CASE*, July 2009.

http://en.wikipedia.org/wiki/Computer-aided_software_engineering

103. Wikipedia, The Free Encyclopedia, *XMI*, July 2009.

http://en.wikipedia.org/wiki/XMI

4 Appendices

4.1 Appendix A – Recommended Preservation Format Evaluation

The following sources of information were used in evaluating the adoption rate:

- Library of Congress: Sustainability of Digital Formats - Planning for Library of Congress Collections (review performed in February 2008) http://www.digitalpreservation.gov/formats/content/content_categories.shtml
- Florida Digital Archives: Florida Center for Library Automation (FCLA), Recommended Data Formats for Preservation Purposes in the Florida Digital Archive, March 2007[70]
- UK Digital Archives: Assessment of UKDA and TNA Compliance with OAIS and METS Standards, Appendix 2, created in 2005/updated in 2007 (http://www.jisc.ac.uk/media/documents/programmes/preservation/oaismets.pdf)
- Canadian Museum of Civilization Corporation (CMCC): Request for Proposal, DAM System, January 2008 (DAMIT7.PDF)
- National Archives of Australia: XENA Digital Preservation Software target format for MS Office products
- Swiss Federal Archives: PDF_Archivtaugliche_Dateiformate_07_en.pdf , July 2007 (available from SFA (http://www.bar.admin.ch/))

The following definitions apply:

- LOC:
    - "Preserve" refers to formats for which full preservation activities will be undertaken
    - "Preserve if only format available" refers to formats for which full preservation activities will be undertaken if the source of the content is only available in this format
- FCLA:
    - High Confidence Level[71]: Formats with a High confidence level are best suited for long-term preservation, including potential format migration. Note that these formats do not necessarily have Action Plans implemented at this time. High confidence level formats:
        - have specifications that are publicly available
        - are uncompressed or use lossless compression
        - are non-proprietary
        - adhere to published standards
    - Medium Confidence Level: Formats with a Medium confidence level are potential candidates for format migration, though they do not have all of the characteristics of High confidence formats. Migration may be imperfect.
    - Low Confidence Level: Formats that are unsuited for long-term preservation via format migration:
        - do not have published specifications
        - use lossy compression
        - are proprietary

[70] http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf
[71] http://fclaweb.fcla.edu/node/893

        - are not governed by any published standards

      FCLA do not intend to provide full preservation for formats listed under Low Confidence Level, as the characteristics of these formats make format migration extremely difficult.

- UKDA:
    - "Preferred file formats are those that are not platform or software dependent or that can be easily transferred to a suitable preservation format"
- CMCC:
    - "Preferred" represents "sole archival formats kept at the Archive for long term preservation purposes"
    - "Accepted" represents "formats for access and operational purposes" and are "kept at the Active, Semi-active and Dormant stages"
    - "Monitored" represents "format developments" that are being monitored by CMCC for applicability to preservation or access/operational purposes
- SFA: All formats endorsed by SFA are considered "standards" within SFA

4.1.1 Content Category: Text

Content Type: Text		Justification	Rating
Format: EPUB (replacement for Open eBook standard)			
	Openness	Freely available from www.idpf.org	✓
	Adoption as a preservation standard	Library of Congress (LOC): Preserve (OEBPS/DTB standards) Florida Digital Archives (FDA): UK Digital Archives (UKDA): Canadian Museum of Civilization Corporation (CMCC):	★
	Stability/Compatibility	a) Forward/backward compatibility Backward compatibility: Compatibility with OEPBS is primary focus of format development Forward compatibility: Existing reader software must be upgraded to accommodate the new standard b) Protected against corruption c) Frequency of version releases OEBPS: Version 1.0: 1999 Version 1.0.1: 2001 Version 1.2: 2002 DTB: Standardized as ANSI/NISO Z39.86-2005 EPUB adopted as a standard in 2007: Version 2.0: OPS and OPF Version 1.0: OCF	★ ✓
	Dependencies/Interoperability	Low/High: Supports XHTML (XHTML 1.1) and DTBook (an XML standard provided by the DAISY (Digital Accessible Information System) Consortium) to represent the text, XML (XML 1.1) for packaging and zip as the container format; free software available for many different hardware (computer and personal digital assistant (PDA)) platforms	✓
	Standardization	**EPUB** (electronic publication) is an e-book standard, by the International Digital Publishing Forum (IDPF) (and supersedes the Open eBook standard); EPub consists of three specifications: 1) Open Publication Structure (OPS) 2.0, contains the formatting of its content. 2) Open Packaging Format (OPF) 2.0, describes the structure of the .epub file in XML. 3) OEBPS Container Format (OCF) 1.0, collects all files as ZIP archive. Adopted by the publishing industry as a means to "publish once" and distribute for	✓

Content Type: Text		Justification	Rating
		many platforms (such as PDAs) applying any required Digital Rights Management (DRM) as the reader (software) level on distribution	
Format: eXtensible HyperText Markup Language (XHTML)			
	Openness	Specification freely available on the W3C web site (http://www.w3.org/)	✓
	Adoption as a preservation standard	Library of Congress (LOC): Florida Digital Archives (FDA): Full preservation (rated at high confidence level) UK Digital Archives (UKDA): Canadian Museum of Civilization Corporation (CMCC):	✶
	Stability/Compatibility	a) Forward/backward compatibility Backward compatibility: Newer software can interpret/render HTML documents based on an earlier version of the format Forward compatibility: Older software can interpret/render documents based on the newer version (to this point – unknown whether this will occur with the XHTML5) b) Protected against corruption c) Frequency of version releases Release 1.0 – January 2000 Release 1.1 – May 2001 Release 2.0 (not backward compatible) – to be superseded by XHTML5 (which is being defined alongside HTML5) (backward compatible)	✶ ✓
	Dependencies/Interoperability	Low/High: Cross-platform content type; renderable with browser software with/without formatting; browser software freely available; humanly readable	✓
	Standardization	W3C recommendation; reformulation of HTML 4 as XML	✓
Format: eXtensible Markup Language (XML)			
	Openness	Specification freely available on the W3C web site (http://www.w3.org/)	✓
	Adoption as a preservation standard	Library of Congress (LOC): Preserve Florida Digital Archives (FDA): Full preservation (rated at high confidence level) UK Digital Archives (UKDA): Preserve Canadian Museum of Civilization Corporation (CMCC): Preferred	✓
	Stability/Compatibility	a) Forward/backward compatibility Backward compatibility: Newer software can interpret/render XML documents based on an earlier version of the format Forward compatibility: Features introduced by new derivatives or applications that conform to the rules of XML can be interpreted by existing XML software b) Protected against corruption c) Frequency of version releases	✓ ✓

Content Type: Text		Justification	Rating
		Release 1.0: 1998 (5 editions) (strict adherence on allowed control characters) Release 1.1: 2004 (2 editions) (expanded use of control characters)	
	Dependencies/Interoperability	Low/High: Cross-platform content type; renderable with browser software with/without formatting; browser software freely available; humanly readable	✓
	Standardization	W3C recommendation; Based on SGML (subset of SGML)	✓
Format: HyperText Markup Language (HTML)			
	Openness	Specification freely available on the W3C web site (http://www.w3.org/) Standard available (at cost) from the International Organization for Standardization (ISO) web site (http://www.iso.org)	✓
	Adoption as a preservation standard	Library of Congress (LOC): Preserve if only format available Florida Digital Archives (FDA): Full preservation (rated at medium confidence level) UK Digital Archives (UKDA): Accepted Canadian Museum of Civilization Corporation (CMCC): Accepted	✓
	Stability/Compatibility	a) Forward/backward compatibility Backward compatibility: Newer software (i.e., browsers) can render older HTML documents Forward compatibility: Features introduced in later versions (e.g., HTML 5) cannot be rendered in earlier software b) Protected against corruption c) Frequency of version releases Draft specification: mid-1993 to 1994 Release 2.0: 1995 – 1997 Release 3.2: January 1997 Release 4.0: December 1997 Release 4.01: December 1999	★ ✓
	Dependencies/Interoperability	Low/High: Cross-platform content type; renderable with browser software that conforms to W3C standard / browser software freely available; humanly readable	✓
	Standardization	W3C standard based on SGML International standard: ISO/IEC 15445:2000	✓
Format: Multipurpose Internet Mail Extensions (MIME)			
	Openness	Specifications available from http://www.ietf.org/	✓
	Adoption as a preservation standard	Library of Congress (LOC): Florida Digital Archives (FDA): UK Digital Archives (UKDA): Canadian Museum of Civilization Corporation (CMCC): Preferred (for e-mail data)	★

Content Type: Text		Justification	Rating
	Stability/Compatibility	a) Forward/backward compatibility Backward compatibility: Newer software can interpret/render content in an earlier version of the format Forward compatibility: Not forward compatible – e.g., 8 Bit MIME cannot be supported by e-mail clients/servers that support only 7 bit (ASCII) MIME b) Protected against corruption c) Frequency of version releases Introduced in 1992 Subsequently extended to address new functionality (e.g., security, attachments in application-specific formats)	★ ✓
	Dependencies/Interoperability	Medium/High: Cross-platform content type; rendered/interpreted by e-mail software; humanly readable for textual portions (e.g., not including attachments in an application-dependent format)	✓
	Standardization	Series of Internet Engineering Task Force (IETF) RFCs	✓
Format: Open Document Format (ODF)			
	Openness	Available from http://www.oasis-open.org/home/index.php, with version 1.0 also available (at cost) from the International Organization for Standardization (ISO) web site (http://www.iso.org)	✓
	Adoption as a preservation standard	Library of Congress (LOC): Florida Digital Archives (FDA): Full preservation (rated at medium confidence level) UK Digital Archives (UKDA): Canadian Museum of Civilization Corporation (CMCC): Monitored National Archives of Australia: Preserve (target for all office automation products)	✓
	Stability/Compatibility	a) Forward/backward compatibility Backward compatibility: Newer software can interpret/render content in an earlier version of the format Forward compatibility: Too early to determine whether forward compatibility will be supported b) Protected against corruption c) Frequency of version releases Version 1.0: May 2005 Version 1.1: February 2007 Version 1.2: Under development (may be released as 2.0) Format still under development	★ ★
	Dependencies/Interoperability	Low/Medium: Based on XML and a ZIP format, hence cross-platform content type; freely available software for interpreting/rendering (www.openoffice.org)	✓

Content Type: Text		Justification	Rating
		available on a number of platforms, as well as being supported by some commercial word processing/office automation products	
	Standardization	Developed by the Organization for the Advancement of Structured Information Standards (OASIS) consortium, also version 1.0 is an ISO standard (ISO/IEC 26300:2006 Open Document Format for Office Applications (OpenDocument) v1.0) Adopted by a number of countries (and NATO) for the production of documents/office automation products	✓
Format: PDF for long-term preservation: PDF-Archive (PDF/A)			
	Openness	Standard available (at cost) from the International Organization for Standardization (ISO) web site (http://www.iso.org)	✓$
	Adoption as a preservation standard	Library of Congress (LOC): Preserve (also use as preservation format for word processing formats, still images) Florida Digital Archives (FDA): PDF/A-1 Full preservation (rated at high confidence level) UK Digital Archives (UKDA): Canadian Museum of Civilization Corporation (CMCC): Monitored – Preferred standard published and tools available Swiss Federal Archives: Standard for "Office" documents	✓
	Stability/Compatibility	a) Forward/backward compatibility Backward compatibility: Newer software can interpret / render the format (e.g., PDF/A-1 will be compliant with PDF/A-2) Forward compatibility: Future versions of PDF-A (e.g., PDF/A-2) may not be compliant with PDF/A-1[72] b) Protected against corruption c) Frequency of version releases Based on Adobe's PDF version 1.4 (Adobe has published 7 versions in total) Adopted as a standard by the ISO in 2005 (PDF/A-1a (more strict with tagging) and PDF/A-1b) PDF/A-2 now being considered to address functionality introduced by Adobe in versions 1.5-1.7	★ ✓
	Dependencies/Interoperability	Medium/High: Designed to be a stable cross-platform format; interpretable/renderable with free software (not humanly readable)	✓
	Standardization	International standard: ISO 19005:2005	✓

[72] A new part to the standard, ISO 19005-1, Part-2 (PDF/A-2), is currently being worked on by the Technical Committee. PDF/A-2 will address some of the new feature added with versions 1.5, 1.6 and 1.7 of the PDF Reference. PDF/A-2 should be backwards compatible, i.e. all valid PDF/A-1 documents should also be compliant with PDF/A-2. However PDF/A-2 compliant files will not necessarily be PDF/A-1 compliant (http://www.pdfa.org/doku.php?id=pdfa:en:pdfa_whitepaper).

Content Type: Text		Justification	Rating
Format: Rich Text Format (RTF)			
	Openness	Specifications for early versions (up to 1.5) are available from various sources on the web (e.g. Sourceforge), later versions available from the Microsoft support web site or from Microsoft resource kits	✖
	Adoption as a preservation standard	Library of Congress (LOC): Florida Digital Archives (FDA): Full preservation (rated at medium confidence level) UK Digital Archives (UKDA): Preserve Canadian Museum of Civilization Corporation (CMCC): Preferred	✓
	Stability/Compatibility	a) Forward/backward compatibility Backward compatibility: Newer software can interpret/render content in an earlier version of the format Forward compatibility: Older software may be able to open later versions of the format with loss of functionality b) Protected against corruption c) Frequency of version releases Based on an early form of LaTeX 1987: RTF 1.0 January 1994: RTF 1.3 April 1997: RTF 1.5 May 1999: RTF 1.6 August 2001: RTF 1.7 April 2004: RTF 1.8 March 2008: RTF 1.9.1 (not anticipated to be further enhanced)	✶ ✶
	Dependencies/Interoperability	High/High: Proprietary format used for transferring textual documents between products; can be interpreted/rendered by a large number of platform-specific word processing/text processing software products, as well as free software; sample reader program provided with developer resources	✖
	Standardization	Under the control of a single company (Microsoft)	✖
Format: Standard Generalized Markup Language (SGML)			
	Openness	Standard available (at cost) from the International Organization for Standardization (ISO) web site (http://www.iso.org)	✓$
	Adoption as a preservation standard	Library of Congress (LOC): Preserve Florida Digital Archives (FDA): Full preservation (rated at medium confidence level) UK Digital Archives (UKDA): Preserve Canadian Museum of Civilization Corporation (CMCC): Accepted	✓
	Stability/Compatibility	a) Forward/backward compatibility Backward compatibility: Newer software can interpret/render older SGML	✓

Content Type: Text		Justification	Rating
		documents/applications Forward compatibility: Features introduced by derivatives or applications that conform to the rules of SGML can be interpreted by existing SGML software b) Protected against corruption c) Frequency of version releases Original GML developed in the 1960s Accepted as an ISO standard 1986 Other: SGML is used as the basis for other formats, including HTML (application), XML (derivative) and for industry-specific applications such as DocBook	✓
	Dependencies/Interoperability	Low/High: Cross-platform content type; software freely available; humanly readable	✓
	Standardization	International standard: ISO 8879:1986	✓
Format: Text (Plain Text)			
	Openness	Unicode specification available from published books or for its corresponding ISO/IEC standard (at cost) from the International Organization for Standardization (ISO) web site (http://www.iso.org)	✓$
	Adoption as a preservation standard	Library of Congress (LOC): Florida Digital Archives (FDA): Full preservation (rated at high confidence level) (ASCII/Unicode) (also used as preservation format for spreadsheets, and delimited text for databases) UK Digital Archives (UKDA): Preserve (ASCII/Unicode) (delimited text for statistical data, databases, spreadsheets) Canadian Museum of Civilization Corporation (CMCC): Preferred (ASCII/Unicode) Swiss Federal Archives: Standard for unstructured information (UTF-8, UTF-16, ISO 8859-1, ISO 8859-15, US ASCII)	✓
	Stability/Compatibility	a) Forward/backward compatibility Backward compatibility: Software designed to interpret/render extensions (i.e., ISO/IEC 8859, Unicode) can interpret/render ASCII Forward compatibility: Software designed for early forms of ASCII (such as, 7-bit ASCII) cannot interpret/render ASCII extensions b) Protected against corruption c) Frequency of version releases ISO/IEC 8859 (Latin Alphabet 1) developed as an extension to ASCII Unicode (double-byte character code) extends ISO/IEC 8859	✶ ✓
	Dependencies/Interoperability	Low/High: Cross-platform (ASCII/Unicode); interpretable/renderable by a large number of software products; human readable	✓

Content Type: Text		Justification	Rating
	Standardization	ASCII (American Standard Code for Information Interexchange): 128 character ASCII (8-bit ASCII) documented in ISO/IEC 8859, as well as in Unicode Unicode: Development coordinated by the Unicode Consortium (Unicode Inc.), a non-profit organization that cooperates with many standards development organizations, including ISO/IEC JTC1, W3C, IETF, and ECMA; documented in published books and in corresponding **ISO/IEC 10646** standard	✓

4.1.2 Content Category: Audio

Content Type: Audio		Justification	Rating
Format: Broadcast Wave Format (BWF)			
	Openness	Specifications available at: 1) EBU Tech 3285 - Specification of the Broadcast Wave Format (BWF) - Version 1 - first edition (2001) 2) EBU Tech 3285-s1 - Specification of the Broadcast Wave Format (BWF) - Supplement 1, MPEG Audio - first edition (1997) 3) EBU Tech 3285-s2 - Specification of the Broadcast Wave Format (BWF) - Supplement 2, Capturing Report - first edition (2001) 4) EBU Tech 3285-s3 - Specification of the Broadcast Wave Format (BWF) - Supplement 3, Peak Envelope Chunk - first edition (2001) 5) EBU Tech 3285-s4 - Specification of the Broadcast Wave Format (BWF) - Supplement 4, Link Chunk - first edition (2003) 6) EBU Tech 3285-s5 - Specification of the Broadcast Wave Format (BWF) - Supplement 5, <axml> Chunk - first edition (2003) WAV specifications available at http://www-mmsp.ece.mcgill.ca/Documents/AudioFormats/WAVE/WAVE.html	✓
	Adoption as a preservation standard[73]	Library of Congress (LOC): Preserve Florida Digital Archives (FDA): WAV is full preservation (rated at high confidence level) UK Digital Archives (UKDA): WAV is preserved Canadian Museum of Civilization Corporation (CMCC): WAV is preferred Swiss Federal Archives: WAVE Standard for audio	✓
	Stability/Compatibility	a) Forward/backward compatibility Backward compatibility: Extension of WAV format – any software that can interpret/render BWF content can also interpret/render WAV content	✓

[73] The rate of adoption of the "WAV" format is used, as BWF is an augmentation of / is forward compatible with the "WAV" format.

Content Type: Audio		Justification	Rating
		Forward compatibility: Software that can interpret/render WAV content can also interpret/render BWF content b) Protected against corruption c) Frequency of version releases WAV: Version 1.0: 1991 Version 3.0: 1994 Multichannel: 2001 BWF: Original: 1997 Updates: 2001, 2003	✓
	Dependencies/Interoperability	Low/High: Although based on WAV (originally proprietary), WAV specifications are openly available; broadly supported by digital audio software and devices	✓
	Standardization	Specified by the European Broadcasting Union; based on WAV format (originally developed by IBM/Microsoft)	✓

4.1.3 Content Category: Digital Video

Content Type: Digital Video		Justification	Rating
Format: JPEG 2000 MXF (MOTION JPEG 2000)			
	Openness	Standard available (at cost) from the International Organization for Standardization (ISO) web site (http://www.iso.org)	✓$
	Adoption as a preservation standard	Library of Congress (LOC): In use for preservation at the National Audiovisual Conservation Centre Florida Digital Archives (FDA): Full preservation (rated at high confidence level) UK Digital Archives (UKDA): Canadian Museum of Civilization Corporation (CMCC):	✓
	Stability/Compatibility	a) Forward/backward compatibility Backward compatibility: Forward compatibility: b) Protected against corruption High Resilience due to intra-frame encoding Data corruption causes gradual degradation due to wavelet based compression c) Frequency of version releases Released in 2001	✓ ✓ ✓
	Dependencies/Interoperability	Medium/High	✓

Content Type: Digital Video		Justification	Rating
	Standardization	International standard: ISO/IEC 15444-3	✓

4.1.4 Content Category: Still Images

Content Type: Still Images		Justification	Rating
Format: Joint Photographic Experts Group (JPEG)			
	Openness	Standard available (at cost) from the International Organization for Standardization (ISO) web site (http://www.iso.org)	✓$
	Adoption as a preservation standard	Library of Congress (LOC): Preserve Florida Digital Archives (FDA): Full preservation (rated at medium confidence level) UK Digital Archives (UKDA): Canadian Museum of Civilization Corporation (CMCC): Accepted – not preserved	✓
	Stability/Compatibility	a) Forward/backward compatibility Backward compatibility: Forward compatibility: b) Protected against corruption c) Frequency of version releases Originally released in 1992 Updated to JPEG2000 in 2000	 ✓
	Dependencies/Interoperability	Low/High:	✓
	Standardization	Originally developed by the Joint Photographic Experts Group in 1992, approved in 1994 as ISO 10918-1 Standards include: JPEG (lossy and lossless): ITU-T T.81, ISO/IEC IS 10918-1 JPEG (extensions): ITU-T T.84 JPEG-LS (lossless, improved): ITU-T T.87, ISO/IEC IS 14495-1 JBIG (black and white pictures): ITU-T T.82, ISO/IEC IS 11544-1	✓
Format: Joint Photographic Experts Group JPEG2000 (JP2)			
	Openness	Standard available (at cost) from the International Organization for Standardization (ISO) web site (http://www.iso.org)	✓$
	Adoption as a preservation standard	Library of Congress (LOC): Preserve Florida Digital Archives (FDA): Full preservation (rated at medium confidence level) UK Digital Archives (UKDA): Canadian Museum of Civilization Corporation (CMCC):	✓
	Stability/Compatibility	a) Forward/backward compatibility Backward compatibility: Programs designed to read JP2 files can read JPEG files	✶

Content Type: Still Images		Justification	Rating
		Forward compatibility: Not forward compatible (programs that can read JPEG files cannot read JP2 files b) Protected against corruption Provides error detection and concealment mechanisms c) Frequency of version releases Released in 2000	 ✓ ✓
	Dependencies/Interoperability	Low/Medium: Designed to be platform-independent; natively supported by some browsers, plug-ins are available	✓
	Standardization	International standards include: JPEG 2000 (successor of JPEG/JPEG-LS): ISO/IEC 15444-1 JPEG 2000 (extensions): ISO/IEC 15444-2 JPEG 2000 (JPM) (compound image file format: ISO/IEC 15444-6 JPEG XR (formerly called HD Photo): ISO/IEC 29199-2	✓
Format: Portable Network Graphics (PNG)			
	Openness	Standard available (at cost) from the International Organization for Standardization (ISO) web site (http://www.iso.org)	✓$
	Adoption as a preservation standard	Library of Congress (LOC): Preserve Florida Digital Archives (FDA): Full preservation (rated at high confidence level) UK Digital Archives (UKDA): Ingested Canadian Museum of Civilization Corporation (CMCC): Accepted – not preserved	✓
	Stability/Compatibility	a) Forward/backward compatibility Backward compatibility: Forward compatibility: b) Protected against corruption c) Frequency of version releases 1) October 1, 1996: Version 1.0 of the PNG specification was released, and later appeared as RFC 2083. It became a W3C Recommendation on October 1, 1996. 2) December 31, 1998: Version 1.1, with some small changes and the addition of three new chunks, was released. 3) August 11, 1999: Version 1.2, adding one extra chunk, was released. 4) November 10, 2003: PNG is now an International Standard (ISO/IEC 15948:2003). This version of PNG differs only slightly from version 1.2 and adds no new chunks. 5) March 3, 2004: ISO/IEC 15948:2004	 ✓
	Dependencies/Interoperability	Low/High	✓
	Standardization	International standard: ISO/IEC 15948:2003; ISO/IEC 15948:2004	✓
Format: Tagged Image File Format (TIFF, TIF)			
	Openness	Freely available from various web sites (including	✓

Content Type: Still Images		Justification	Rating
		http://partners.adobe.com/public/developer/tiff/index.html), the ISO standard is available (at cost) from the International Organization for Standardization (ISO) web site (http://www.iso.org)	
	Adoption as a preservation standard	Library of Congress (LOC): Preserve Florida Digital Archives (FDA): Full preservation (rated at high confidence level) UK Digital Archives (UKDA): Preserve Canadian Museum of Civilization Corporation (CMCC): Preferred Swiss Federal Archives: Standard for raster image	✓
	Stability/Compatibility	a) Forward/backward compatibility Backward compatibility: Forward compatibility: b) Protected against corruption c) Frequency of version releases Version 6: 1992	 ✓
	Dependencies/Interoperability	Low/High: Cross-platform; widely supported by image-manipulation applications, by publishing and page layout applications, by scanning, faxing, word processing, optical character recognition and other applications	✓
	Standardization	International standard: ISO 12639:1998 (TIFF/IT)	✓
Format: TIFF - GeoTIFF			
	Openness	Specification freely available from http://trac.osgeo.org/geotiff/	✓
	Adoption as a preservation standard	Library of Congress (LOC): Florida Digital Archives (FDA): UK Digital Archives (UKDA): Canadian Museum of Civilization Corporation (CMCC):	×
	Stability/Compatibility	a) Forward/backward compatibility Backward compatibility: Fully compatible with TIFF Forward compatibility: Software incapable of using metadata can interpret/render the content b) Protected against corruption c) Frequency of version releases Version 1.0	✓ ✶
	Dependencies/Interoperability	Low/High: Based on TIFF standard; cross-platform content type; as TIFF derivate is widely supported; freely available viewers for GeoTIFF	✓
	Standardization	A public domain metadata standard which allows georeferencing information to be embedded within a TIFF file	✓

4.1.5 Content Category: Structured Data - Databases

Content Type: Structured Data - Database		Justification	Rating
Format: Software Independent Archiving of Relational Databases (SIARD)			
	Openness	Specification available from Swiss Federal Archives SFA (http://www.bar.admin.ch/) and possibly from PLANETS http://www.planets-project.eu/	✶
	Adoption as a preservation standard	Library of Congress (LOC): Florida Digital Archives (FDA): UK Digital Archives (UKDA): Canadian Museum of Civilization Corporation (CMCC): Swiss Federal Archives: Standard for relational databases	✶
	Stability/Compatibility	a) Forward/backward compatibility Backward compatibility: Forward compatibility: b) Protected against corruption c) Frequency of version releases Newly established standard – unknown, however based on stable standards	 ✶
	Dependencies/Interoperability	Low/High: Uses XML and SQL 99 as a basis; designed for platform independence	✓
	Standardization	The published SIARD database format can be used independently of the *SIARD Suite* applications for long-term archiving of electronic databases. If structure and content of a database are migrated to the SIARD format, it will be possible to access the database data at any later time, even if the original database software is either unavailable or not executable. This has been achieved through the best use of suitable internationally supported standards in the SIARD format. This long term interpretability of database content is based essentially on the two ISO standards, XML and SQL:1999. SIARD was accepted in May 2008 as the official format for archiving relational databases of the European PLANETS project.	✶
Format: Text - Delimited Flat File with Data Description			
	Openness	See Text	
	Adoption as a preservation standard	Library of Congress (LOC): Florida Digital Archives (FDA): Full preservation (rated at high confidence level) (Database) UK Digital Archives (UKDA): Preserve (Database) Canadian Museum of Civilization Corporation (CMCC):	✓
	Stability/Compatibility	See Text	
	Dependencies/Interoperability	See Text	

Content Type: Structured Data - Database		Justification	Rating
	Standardization	See Text	

4.1.6 Content Category: Structured Data - Statistical and Qualitative Analysis Data

Content Type: Structured Data - Statistical and Qualitative Analysis Data		Justification	Rating
Format: Data Documentation Initiative (DDI) Version 3.0			
	Openness	Specification freely available from http://www.ddialliance.org/	✓
	Adoption as a preservation standard	Library of Congress (LOC): Florida Digital Archives (FDA): UK Digital Archives (UKDA): Canadian Museum of Civilization Corporation (CMCC): Adopted by individual organizations, such as UK Data Archives (underlying standard for the DExT initiative); currently has support from the National Science Foundation (NSF award SES0136447) and Health Canada (complete list of projects currently using the standard can be found at http://www.ddialliance.org/DDI/codebook/projects.html) Used by NESSTAR a Semantic Web application for statistical data and metadata that aims to streamline the process of finding, accessing and analysing statistical information	★
	Stability/Compatibility	a) Forward/backward compatibility Backward compatibility: Forward compatibility: b) Protected against corruption c) Frequency of version releases Version 1.0: 2000 Version 2.0: 2006 Version 3.0: 2008	✓
	Dependencies/Interoperability	Low/High: Uses XML, Dublin Core and the Council of European Social Science Data Archives (CESSDA) metadata standards; specifically designed for platform-neutral content	✓
	Standardization	The Data Documentation Initiative (DDI) is an international effort to establish a standard for technical documentation describing social science data, specifically it is an XML-based standard for the content, presentation, transport, and preservation	✓

Content Type: Structured Data - Statistical and Qualitative Analysis Data		Justification	Rating
		of documentation for datasets in the social and behavioral sciences	
Format: Data Exchange and Conversion Utilities and Tools (DExT)			
	Openness	Specification freely available from www.data-archive.ac.uk/dext; tools are open source	✓
	Adoption as a preservation standard	Library of Congress (LOC): Florida Digital Archives (FDA): UK Digital Archives (UKDA): Canadian Museum of Civilization Corporation (CMCC): Initiative of the UK Digital Archives (built on DDI 3.0)	✶
	Stability/Compatibility	a) Forward/backward compatibility Backward compatibility: Forward compatibility: b) Protected against corruption c) Frequency of version releases Relatively recent initiative, however built on open standards	✶ ✶
	Dependencies/Interoperability	Low/High: Built on existing standards (METS, XML, MODS, DDI); designed for preservation	✓
	Standardization	Funded under the JISC Repositories and Preservation Programme with the objective of providing a standard uniform format for richly encoding research and data (enhancing quantitative data with qualitative data), specifically to: • enabling the long-term preservation and re-use of metadata, data and annotation (software and platform-independent formats) • ensure consistency of presentation and description of data • facilitate the conversion of data to and from common statistical and qualitative data analysis (CAQDAS) packages using an open archival format specification • support the development of common web-based publishing and search tools • enable more precise searching/browsing of archived data beyond the collection-level descriptive record • and facilitate data interchange, sharing among dispersed collections and repositories (comparative analysis and e-science) Uses METS schema to encapsulate: • Metadata Object Description Schema (MODS) • Dublin Core (DC) • Text Encoding initiative (TEI)	✓

Content Type: Structured Data - Statistical and Qualitative Analysis Data		Justification	Rating
		• Data Documentation Initiative (DDI) • Synchronized Multimedia Integration Language (SMIL)	
Format: Statistical Data and Metadata Exchange (SDMX)			
	Openness	Specification freely available from www.data-archive.ac.uk/dext; tools are open source	✓
	Adoption as a preservation standard	Library of Congress (LOC): Florida Digital Archives (FDA): UK Digital Archives (UKDA): Canadian Museum of Civilization Corporation (CMCC): Initiative of the UK Digital Archives (built on DDI 3.0)	✶
	Stability/Compatibility	a) Forward/backward compatibility Backward compatibility: Forward compatibility: b) Protected against corruption c) Frequency of version releases Relatively recent initiative, however built on open standards	 ✶
	Dependencies/Interoperability	Low/High: Built on existing standards (METS, XML, MODS, DDI); designed for preservation	✓
	Standardization	Funded under the JISC Repositories and Preservation Programme with the objective of providing a standard uniform format for richly encoding research and data (enhancing quantitative data with qualitative data), specifically to: • enabling the long-term preservation and re-use of metadata, data and annotation (software and platform-independent formats) • ensure consistency of presentation and description of data • facilitate the conversion of data to and from common statistical and qualitative data analysis (CAQDAS) packages using an open archival format specification • support the development of common web-based publishing and search tools • enable more precise searching/browsing of archived data beyond the collection-level descriptive record • and facilitate data interchange, sharing among dispersed collections and repositories (comparative analysis and e-science) Uses METS schema to encapsulate: • Metadata Object Description Schema (MODS) • Dublin Core (DC)	✓

Content Type: Structured Data - Statistical and Qualitative Analysis Data		Justification	Rating
		• Text Encoding initiative (TEI) • Data Documentation Initiative (DDI) • Synchronized Multimedia Integration Language (SMIL)	
Format: Text - Delimited Flat File with Variable Description			
	Openness	See Text	
	Adoption as a preservation standard	Library of Congress (LOC): Florida Digital Archives (FDA): UK Digital Archives (UKDA): Preserve (Statistical) Canadian Museum of Civilization Corporation (CMCC):	✶
	Stability/Compatibility	See Text	
	Dependencies/Interoperability	See Text	
	Standardization	See Text	

4.1.7 Content Category: Geospatial

Content Type: Geospatial Data		Justification	Rating
Format: ISO 19115 Geographic Information – Metadata (NAP – Metadata) (North American Profile)			
	Openness	Standard available (at cost) from the International Organization for Standardization (ISO) web site (http://www.iso.org)	✓$
	Adoption as a preservation standard	Library of Congress (LOC): Florida Digital Archives (FDA): UK Digital Archives (UKDA): Canadian Museum of Civilization Corporation (CMCC): Government of Canada Materials: TBS Policy	✓(GoC) / n.a.
	Stability/Compatibility	a) Forward/backward compatibility Backward compatibility: Forward compatibility: b) Protected against corruption c) Frequency of version releases	
	Dependencies/Interoperability		
	Standardization	Treasury Board Secretariat (TBS) policy for geomatics produced by the Government of Canada International standard: ISO 19115	✓

4.1.8 Content Category: Computer-Aided Design (CAD) – Technical Drawings

Content Type: Computer-Aided Design (CAD) – Technical Drawings		Justification	Rating
Format: Drawing Interchange File Format (DXF)			
	Openness	Available from AutoDesk website http://usa.autodesk.com/adsk/servlet/item?siteID=123112&id=12272454&linkID=10809853	×
	Adoption as a preservation standard	Library of Congress (LOC): Preserve Florida Digital Archives (FDA): UK Digital Archives (UKDA): Preserve Canadian Museum of Civilization Corporation (CMCC): Preferred	✓
	Stability/Compatibility	a) Forward/backward compatibility Backward compatibility: Forward compatibility: b) Protected against corruption c) Frequency of version releases	
	Dependencies/Interoperability	Medium/High: Supported by a wide variety of software vendors; ASCII versions of DXF are human readable	✓
	Standardization	Current version of DXF based on ISO/IEC 29500-2:2008 Open Packaging Convention Under vendor control, however, the specification is broadly supported by CAD software vendors	✶

4.1.9 Content Category: Computer-Aided Design (CAD) – CASE

Content Type: Computer-Aided Design (CAD) – CASE		Justification	Rating
Format: XML Metadata Interchange (XMI)			
	Openness	Specification freely available from http://www.omg.org/; ISO standard available (at cost) from the International Organization for Standardization (ISO) web site (http://www.iso.org)	✓
	Adoption as a preservation standard	Library of Congress (LOC): Florida Digital Archives (FDA): UK Digital Archives (UKDA):	×

Content Type: Computer-Aided Design (CAD) – CASE		Justification	Rating
		Canadian Museum of Civilization Corporation (CMCC):	
	Stability/Compatibility	a) Forward/backward compatibility Backward compatibility: Forward compatibility: b) Protected against corruption c) Frequency of version releases Version 2.0.1: 2005 Version 2.1: December 2005 Version 2.1.1: 2007	✶
	Dependencies/Interoperability	Low/Medium: Based on open standards; designed for information exchange between CASE tool vendors	✓
	Standardization	XMI was developed by the Object Management Group (OMG) – it is a standard for exchanging metadata information via Extensible Markup Language (XML), and can be used for any metadata whose metamodel can be expressed in Meta-Object Facility (MOF). The most common use of XMI is as an interchange format for UML models, although it can also be used for serialization of models of other languages (metamodels). XMI integrates four industry standards: 1) XML - eXtensible Markup Language, a W3C standard. 2) UML - Unified Modeling Language, an OMG modeling standard. 3) MOF - Meta Object Facility, an OMG language for specifying metamodels. 4) MOF Mapping to XMI Version 2.0.1 adopted by ISO as ISO/IEC 19503:2005	✓

4.2 Appendix B – Applying the Guidelines to LAC Preservation Policies

The following sections present mindmaps of LAC's Format Guidelines, as follows[74]:

- Section 4.2.1 identifies by content category the formats that are "recommended" for preservation, as well as the formats that are "acceptable for transfer" (these formats are normalized to a recommended preservation format upon ingest as required)[75,76];
- Section 4.2.2 provides examples of normalization/migration paths that may be applied to transform the digital content from an "acceptable for transfer" format to a "recommended" preservation format: the actual normalization paths will vary by the type of digital content and the significant properties of the content that must be preserved (for example, for legal documents, both content and presentation are characteristics to be preserved – as a result a Word document would be migrated to PDF/A);
- Section 4.2.3 presents examples of mapping the "recommended" preservation formats to a service copy format: in many instances the preservation and service copy formats may be identical, however, where the size of the format is a factor, a format that allows for a lower resolution may be selected for servicing access requests;
- Section 4.2.4 presents examples of the mapping of service copy formats to the services that will be used to render the digital content – termed play-out services.

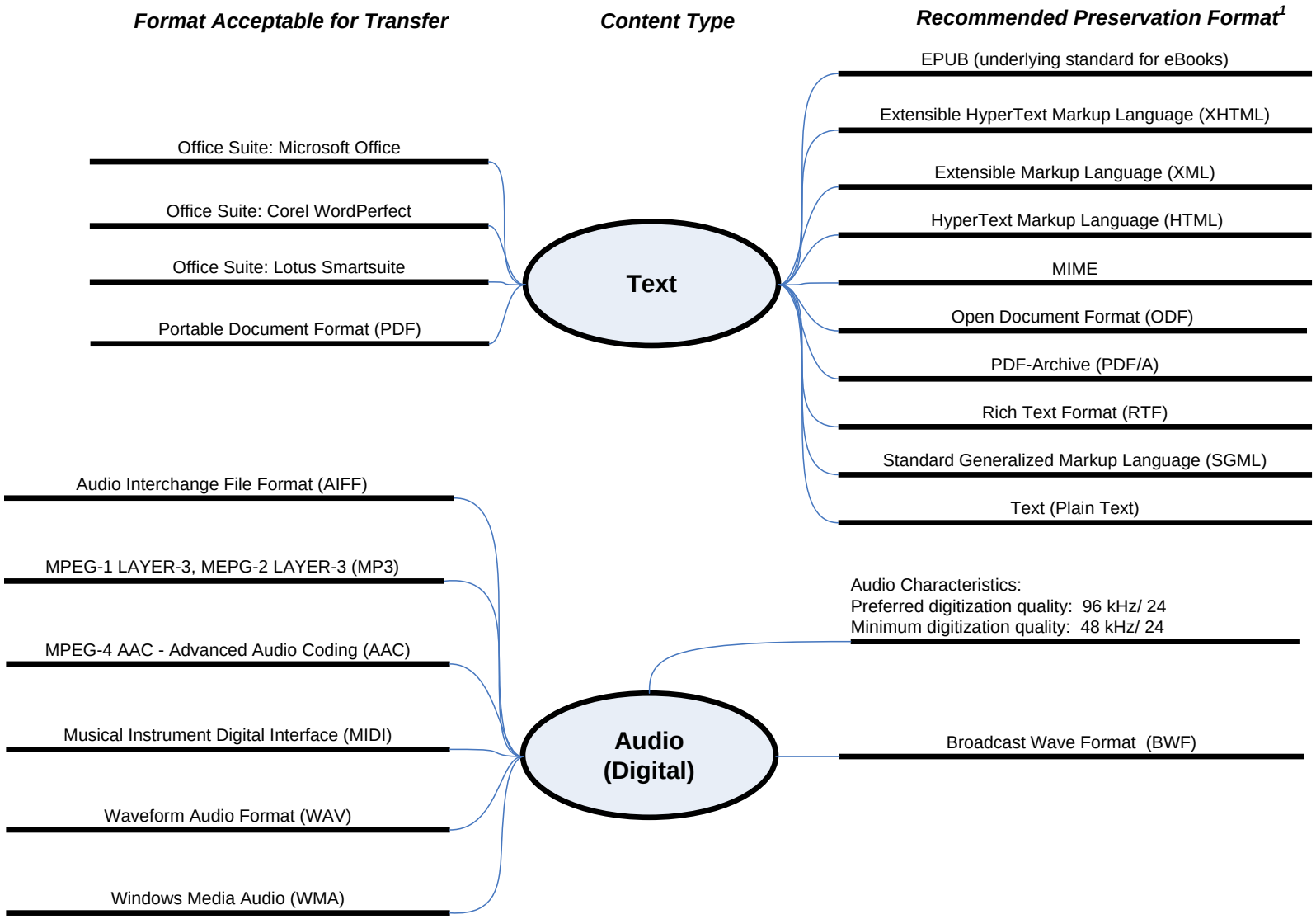
4.2.1 Summary of "Recommended" Preservation and "Acceptable for transfer" File Formats by Content Category

The diagrams following provide an overview of the file formats that LAC has identified as being "recommended" for preservation as well as those that are "acceptable for transfer" of digital content to LAC.

Please note that there is no implied migration path from the "acceptable for transfer" to the "recommended" formats in these diagrams: the selection of a migration target/path will be based on the significant properties of a submitted file format (which may/may not be an "acceptable for transfer" format) and the degree to which these properties need to be preserved (which may vary by the type of collection (e.g., GoC e-Records, or Legal Deposit e-Publications).

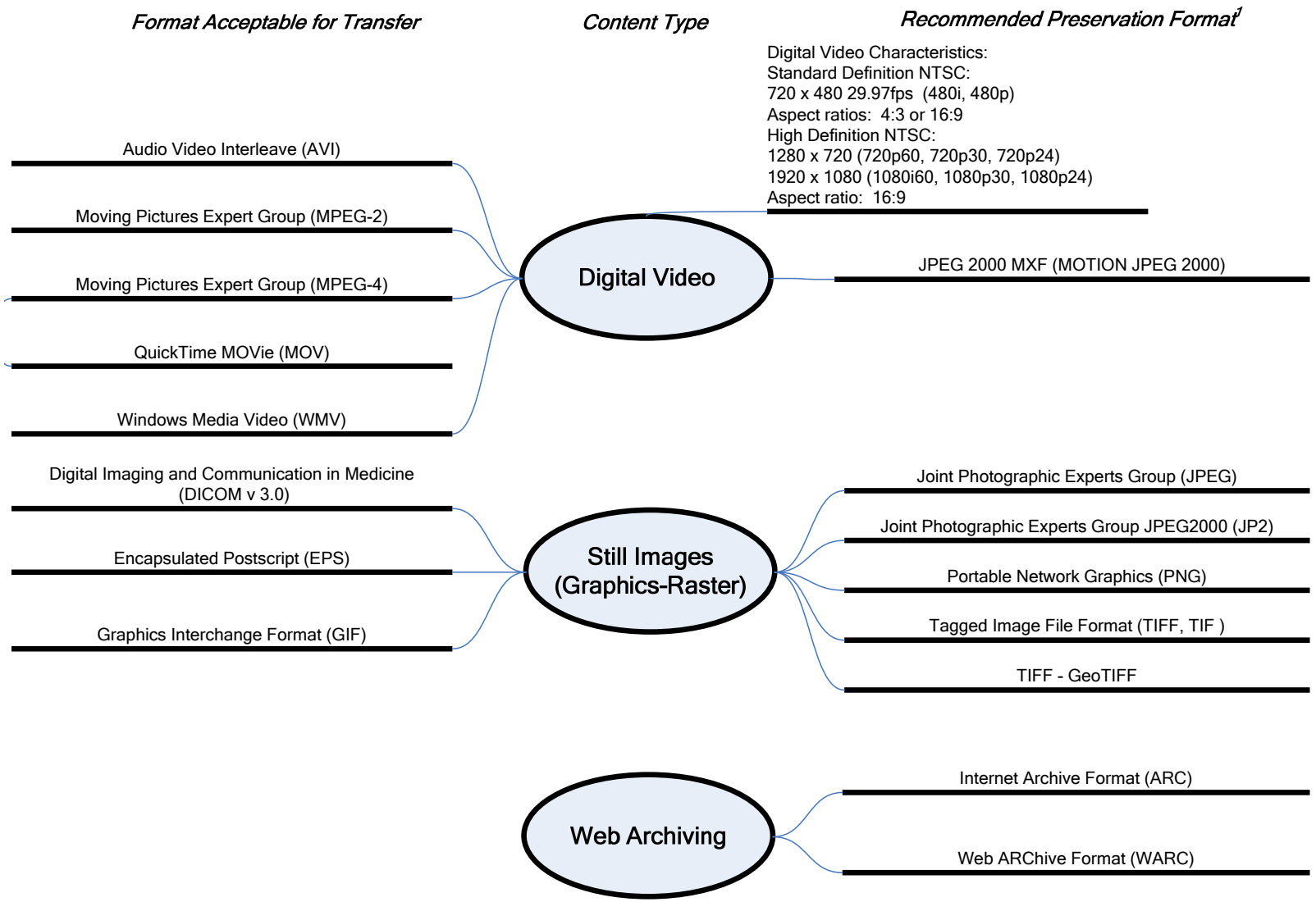
[74] Note: The last three sections of mindmaps represent "work in progress" and will be used as the starting point for implementing LAC's LDFR policies in the TDR.
[75] Note: All formats that are "recommended" for preservation are by default "acceptable for transfer".
[76] Some exceptions exist: for example, in the geospatial content category, some formats will be retained in their original format.

Figure 1: LAC Format Guidelines



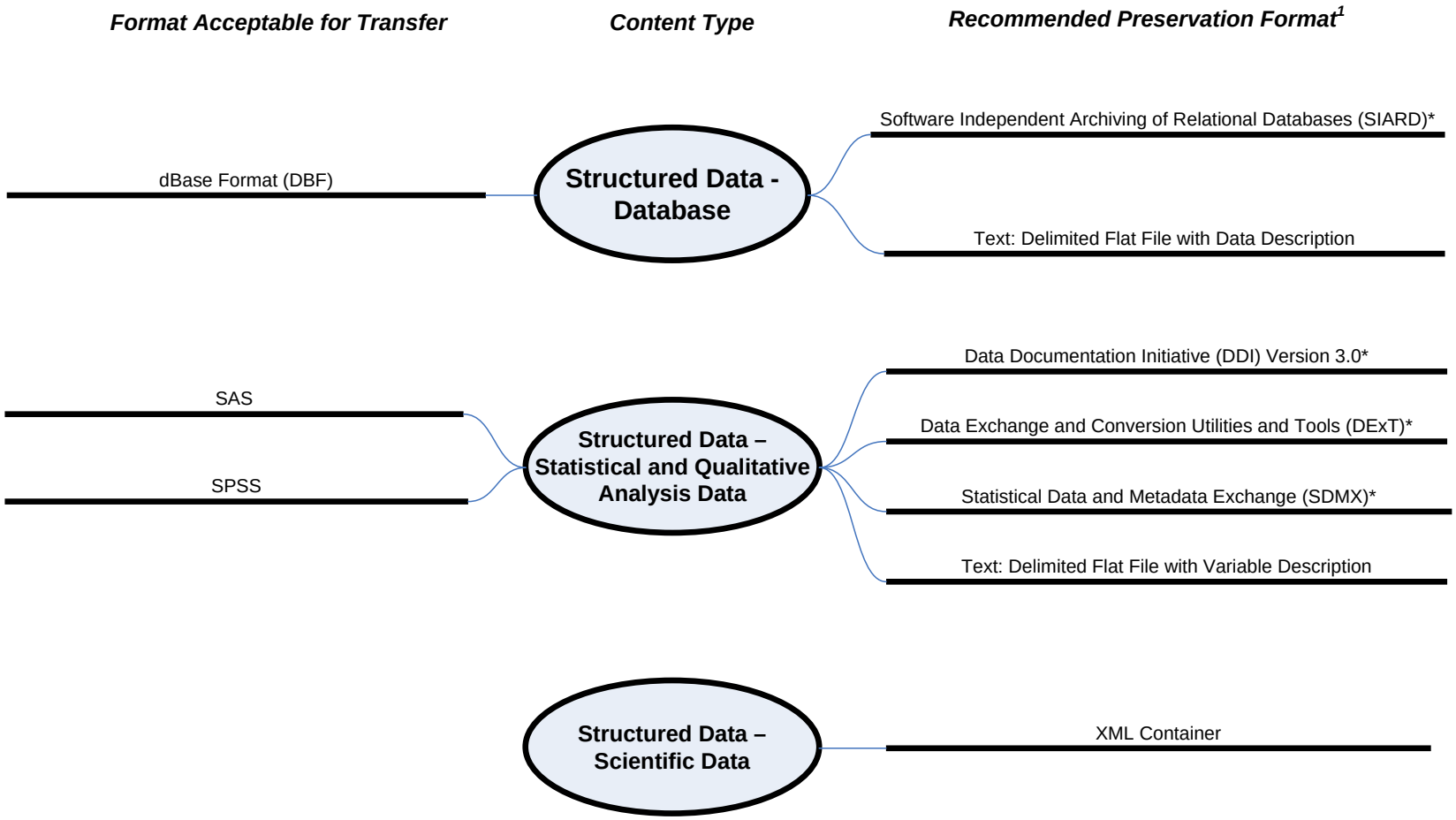
[1] All recommended preservation formats are acceptable for transfer

Figure 2: LAC Format Guidelines



[1] All recommended preservation formats are acceptable for transfer

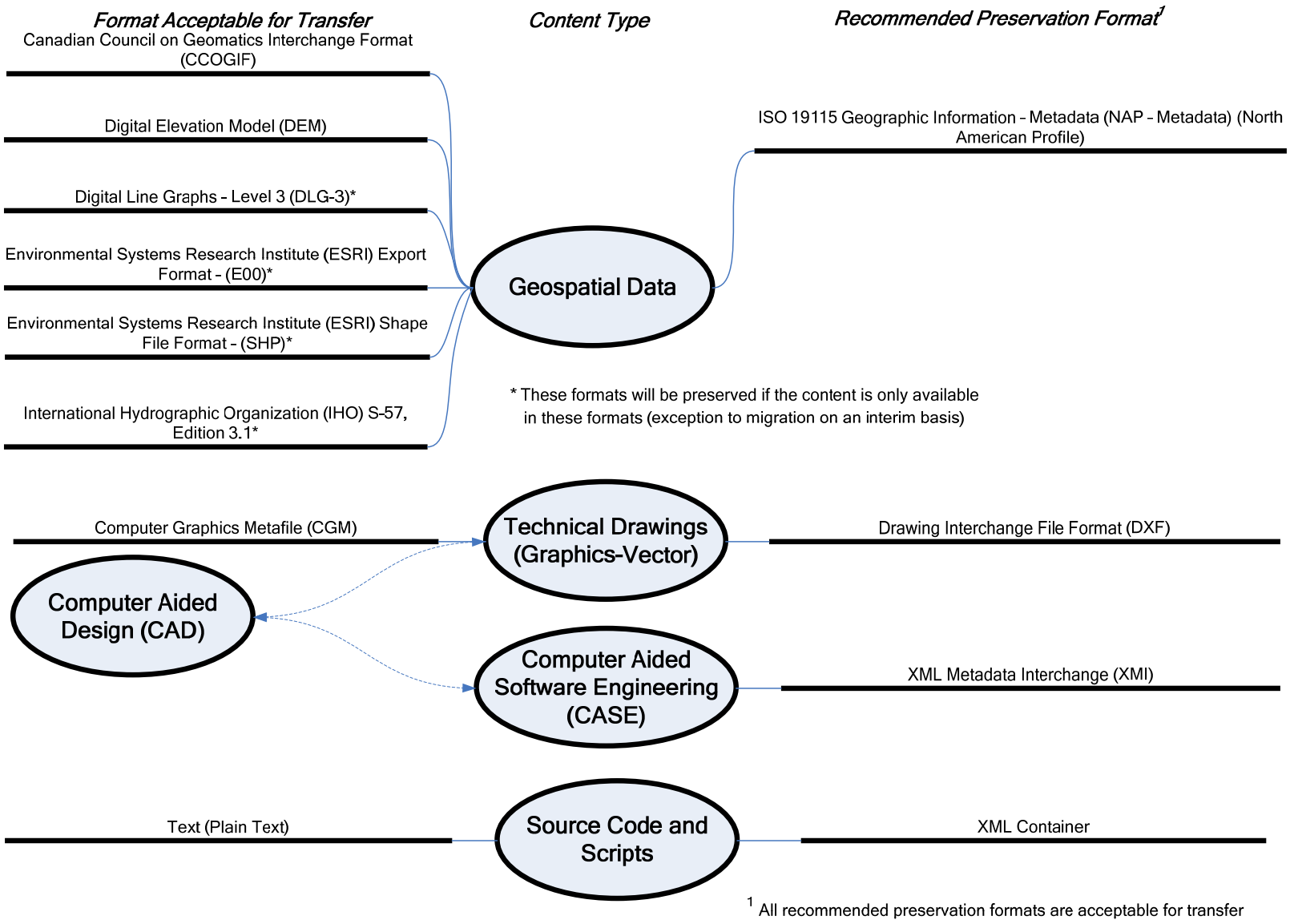
Figure 3: LAC Format Guidelines



* Indicates that the target format will be subject to detailed investigation and prototyping as part of the preservation activities in preparation for implementation Post-Release 3.0

[1] All recommended preservation formats are acceptable for transfer

Figure 4: LAC Format Guidelines

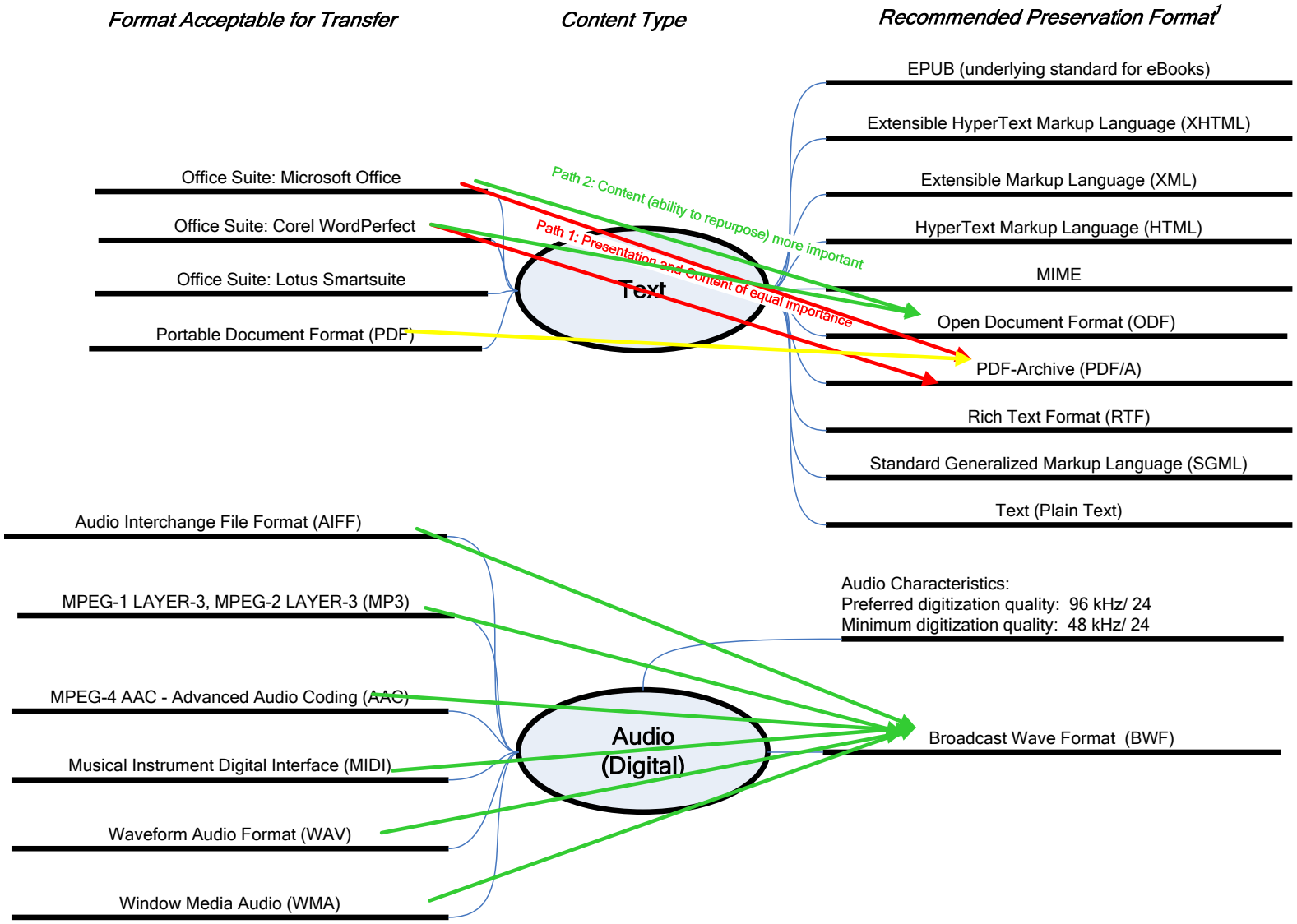


4.2.2 Examples of Migration Paths

The diagrams on the following pages provide examples of migration/normalization paths that may be adopted for some of the "acceptable for transfer" file formats. The actual target preservation file format and the migration/normalization path employed will be based on the significant properties to be retained from the source format[77].

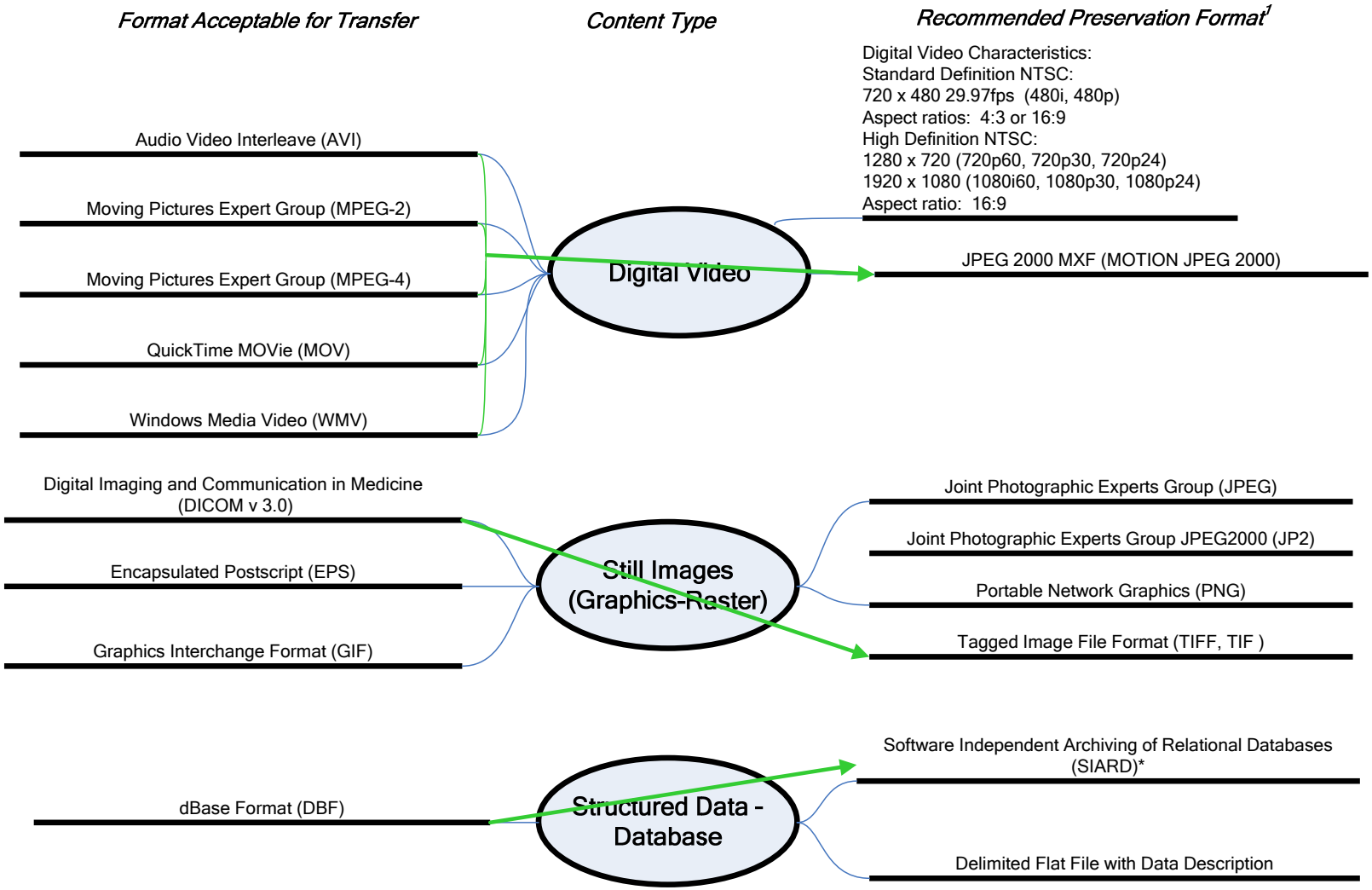
[77] Work is currently being undertaken by LAC's TDR Preservation Working Group to address migration/normalization targets/paths.

Figure 5: LAC Format Guidelines – Normalization/Migration Paths



[1] All recommended preservation formats are acceptable for transfer

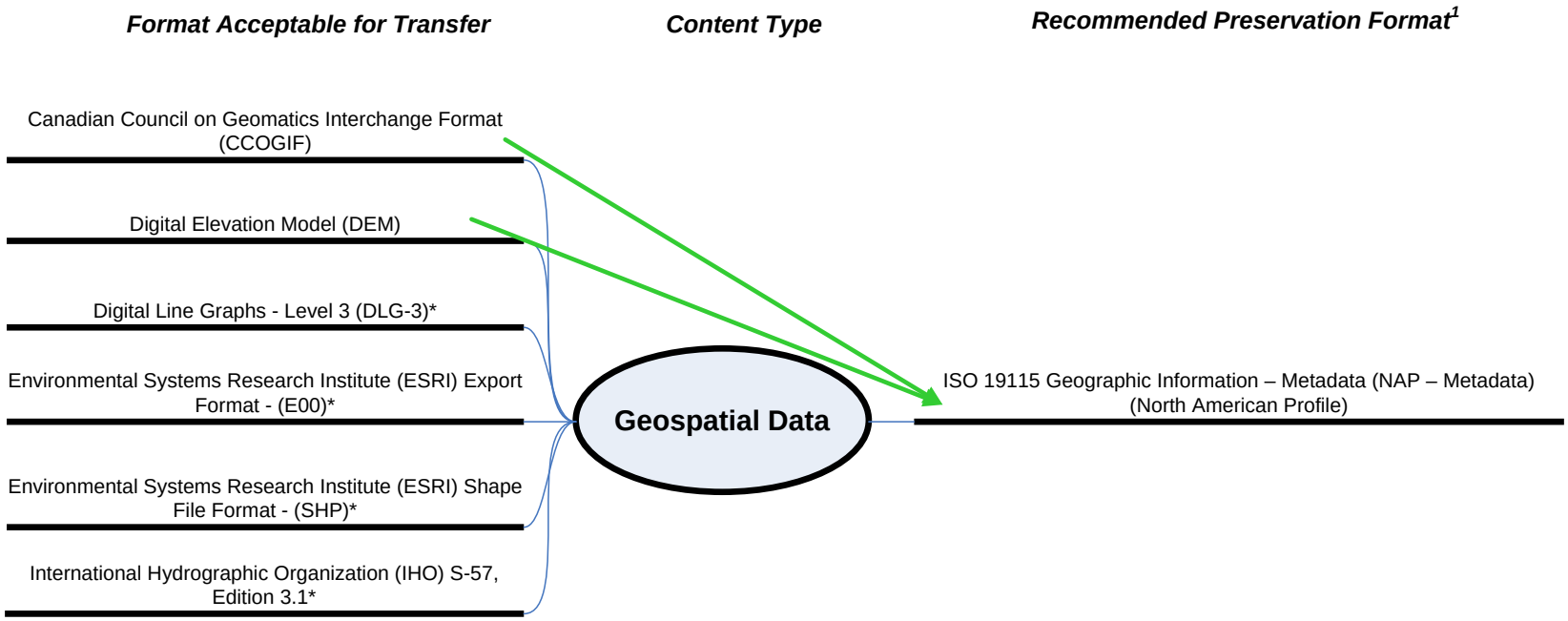
Figure 6: LAC Format Guidelines – Normalization/Migration Paths



* Indicates that the target format will be subject to detailed investigation and prototyping as part of the preservation activities in preparation for implementation Post-Release 3.0

[1] All recommended preservation formats are acceptable for transfer

Figure 7: LAC Format Guidelines – Normalization/Migration Paths



* These formats will be preserved if the content is only available in these formats (exception to migration on an interim basis)

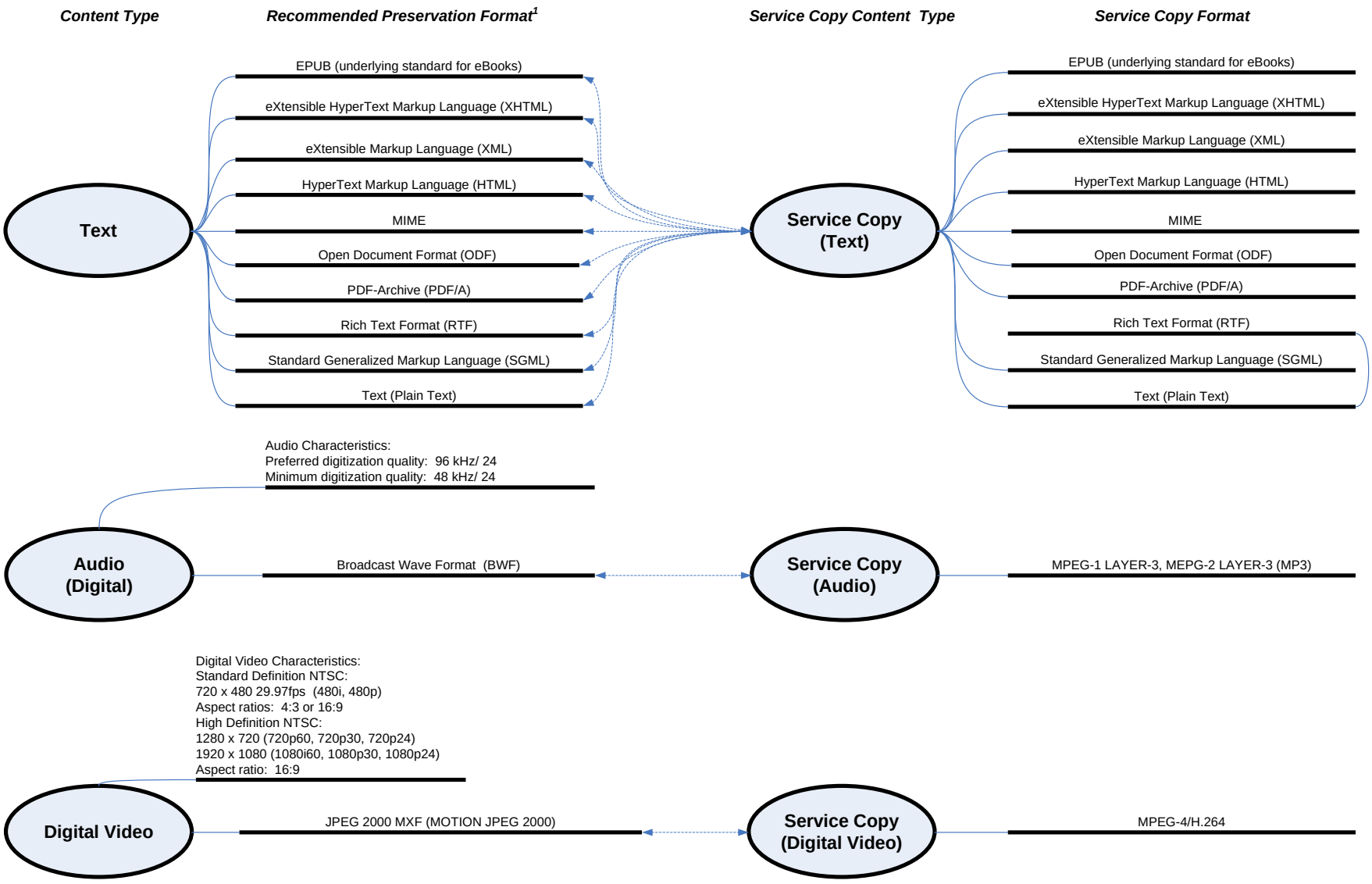
[1] All recommended preservation formats are acceptable for transfer

4.2.3 Mapping Preservation Formats to Service Copy Formats

The diagrams on the following pages provide examples of potential mappings from preservation formats to formats that will be used to service requests by internal/external users[78].

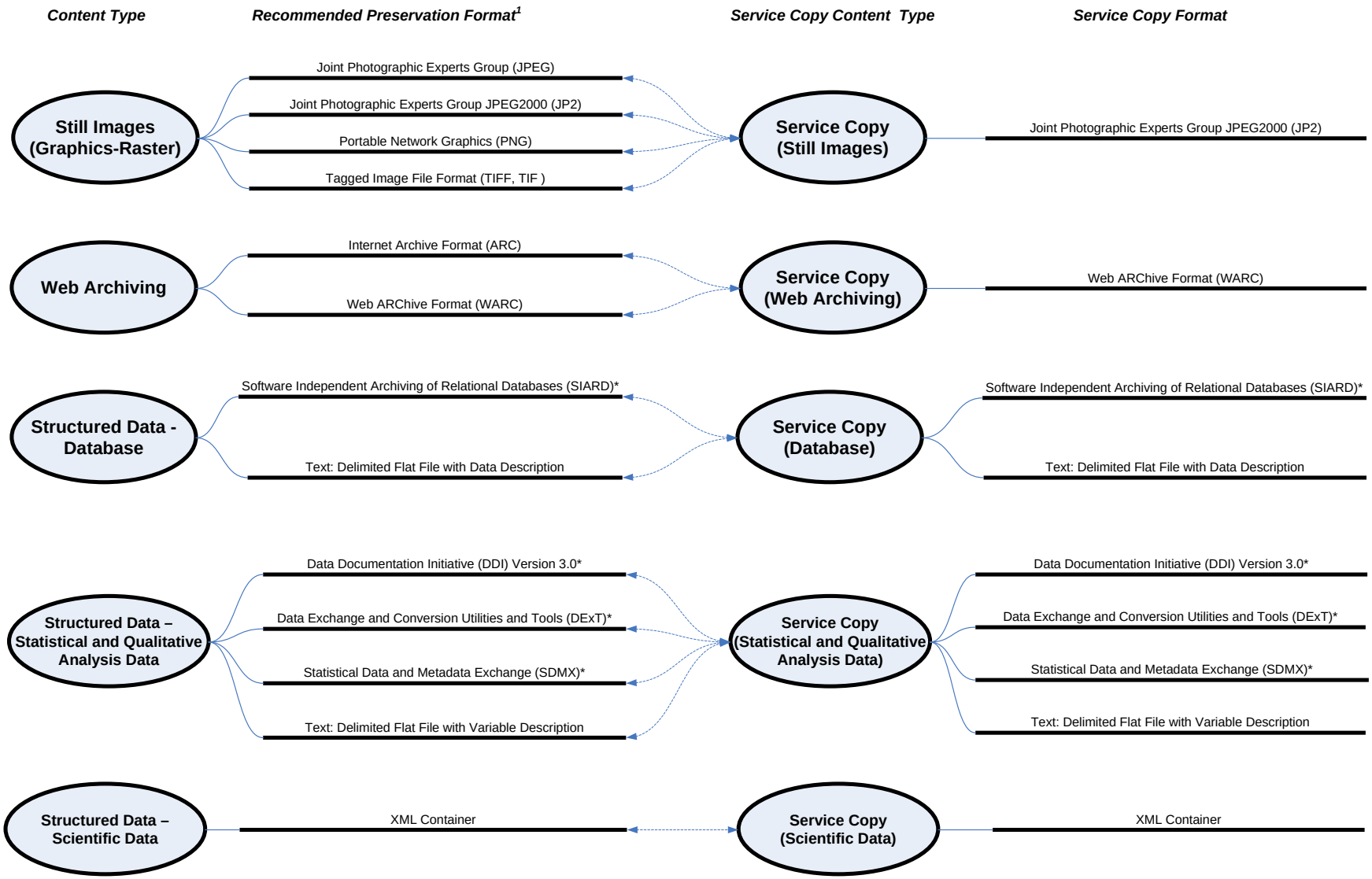
[78] Work is currently being undertaken by LAC's TDR Preservation Working Group to identify the specific rules for producing service copies and the actual formats to be used (e.g, displaying low resolution images for external users, generating service copies on demand, etc.).

Figure 8: LAC Format Guidelines – Service Copy Formats



[1] All recommended preservation formats are acceptable for transfer

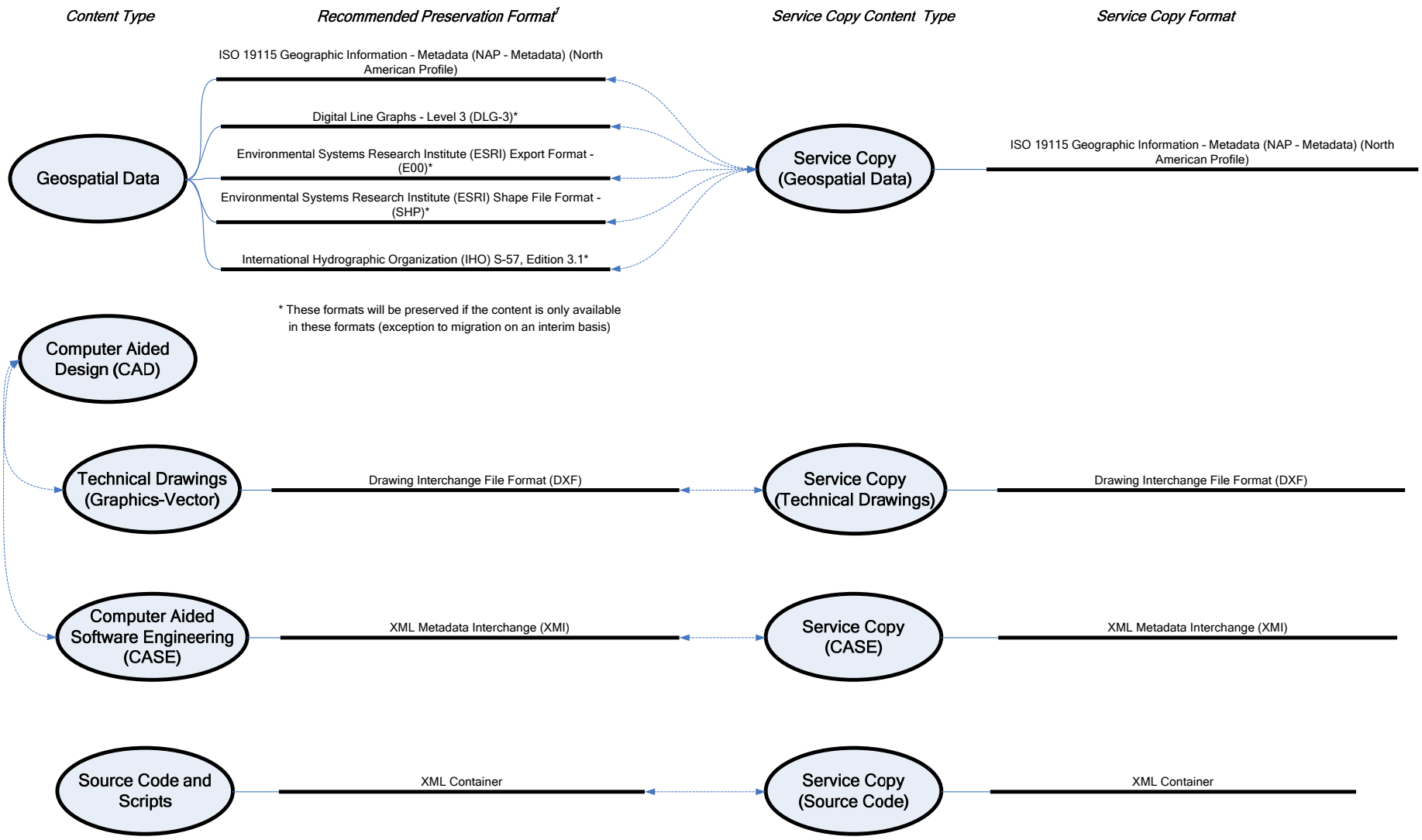
Figure 9: LAC Format Guidelines – Service Copy Formats



* Indicates that the target format will be subject to detailed investigation and prototyping as part of the preservation activities in preparation for implementation Post-Release 3.0

[1] All recommended preservation formats are acceptable for transfer

Figure 10: LAC Format Guidelines – Service Copy Formats



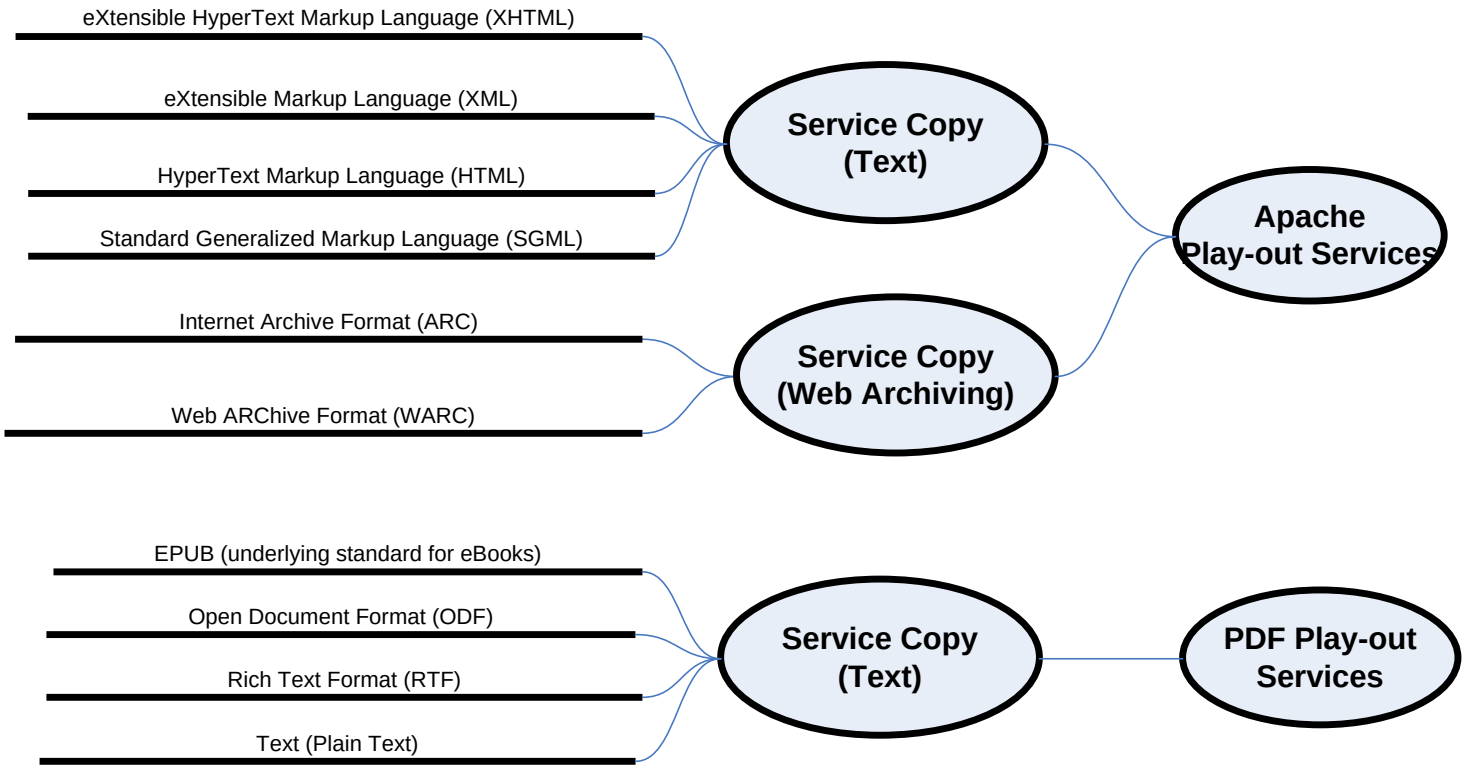
[1] All recommended preservation formats are acceptable for transfer

4.2.4 Mapping Service Copy Formats to Play-out Services

The final set of diagrams on the following pages provides examples of how different service copy formats may be rendered for viewing by an internal/external user (termed "play-out" services)[79].

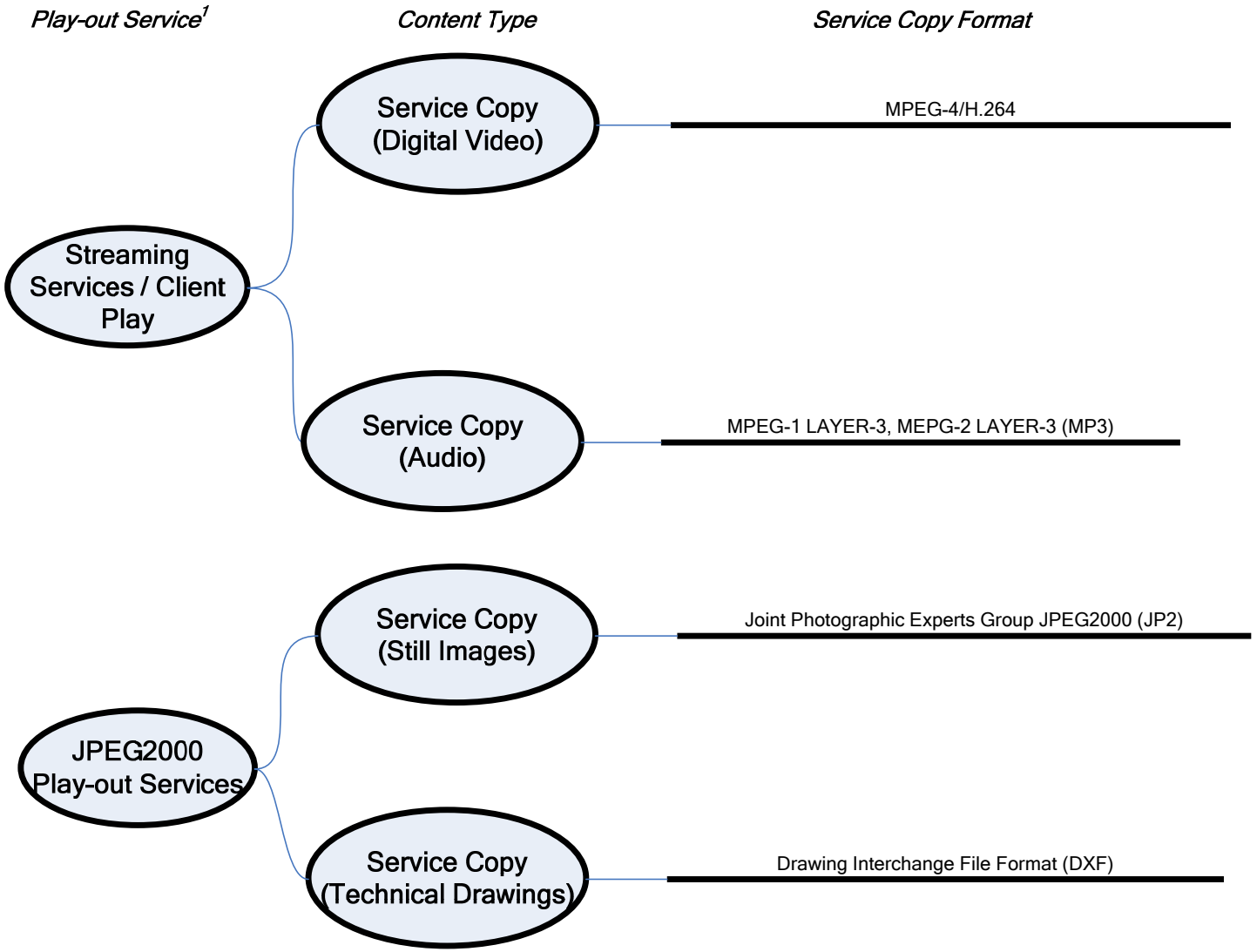
[79] Work is currently being undertaken by LAC's TDR Preservation Working Group to address the "play-out" services.

Figure 11: LAC Format Guidelines - Mapping to Play-out Services[1]



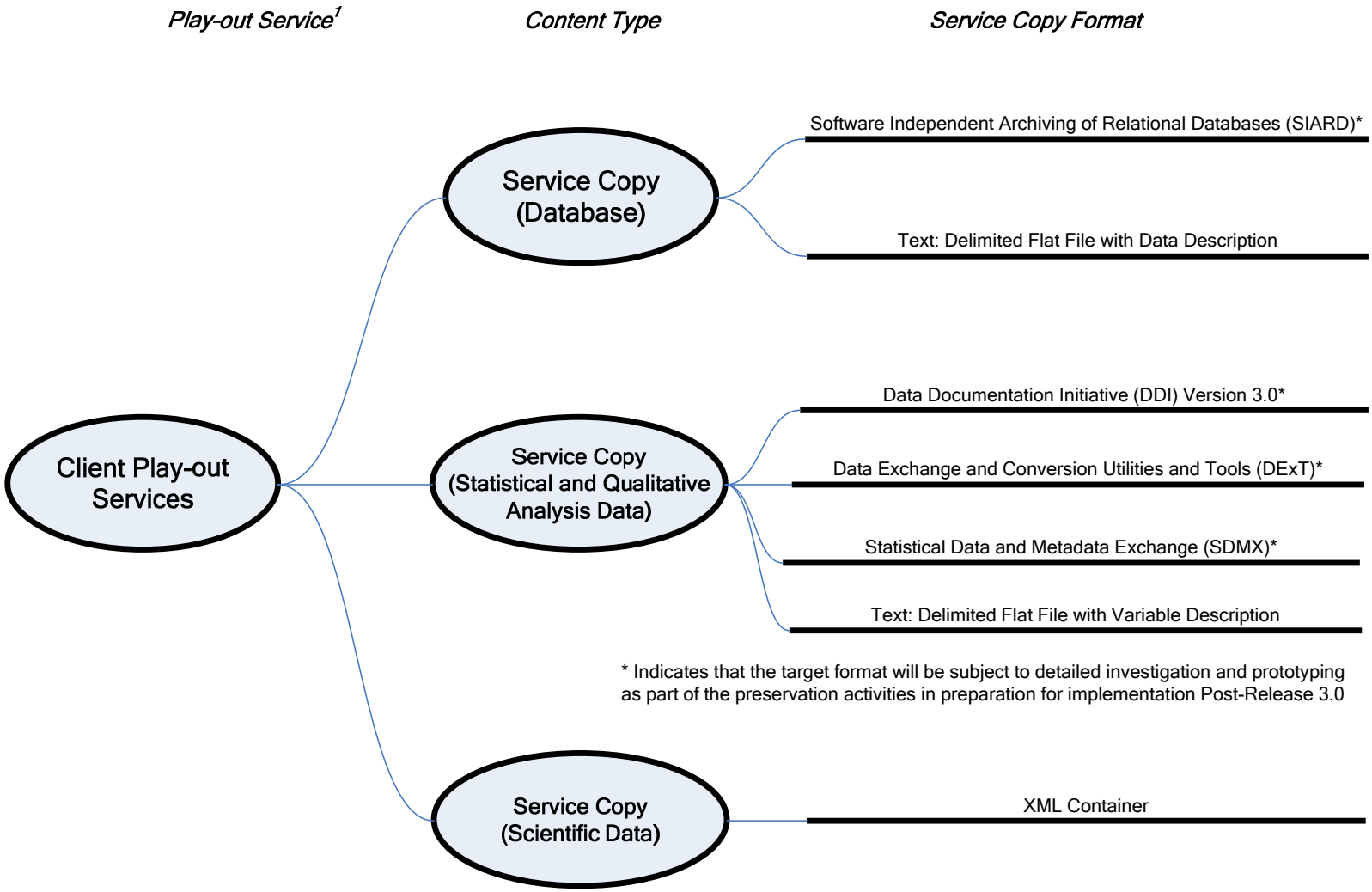
[1] In all instances, the option of a "client" play-out service is always available (e.g., "open with/save as")

Figure 12: LAC Format Guidelines - Mapping to Play-out Services[1]



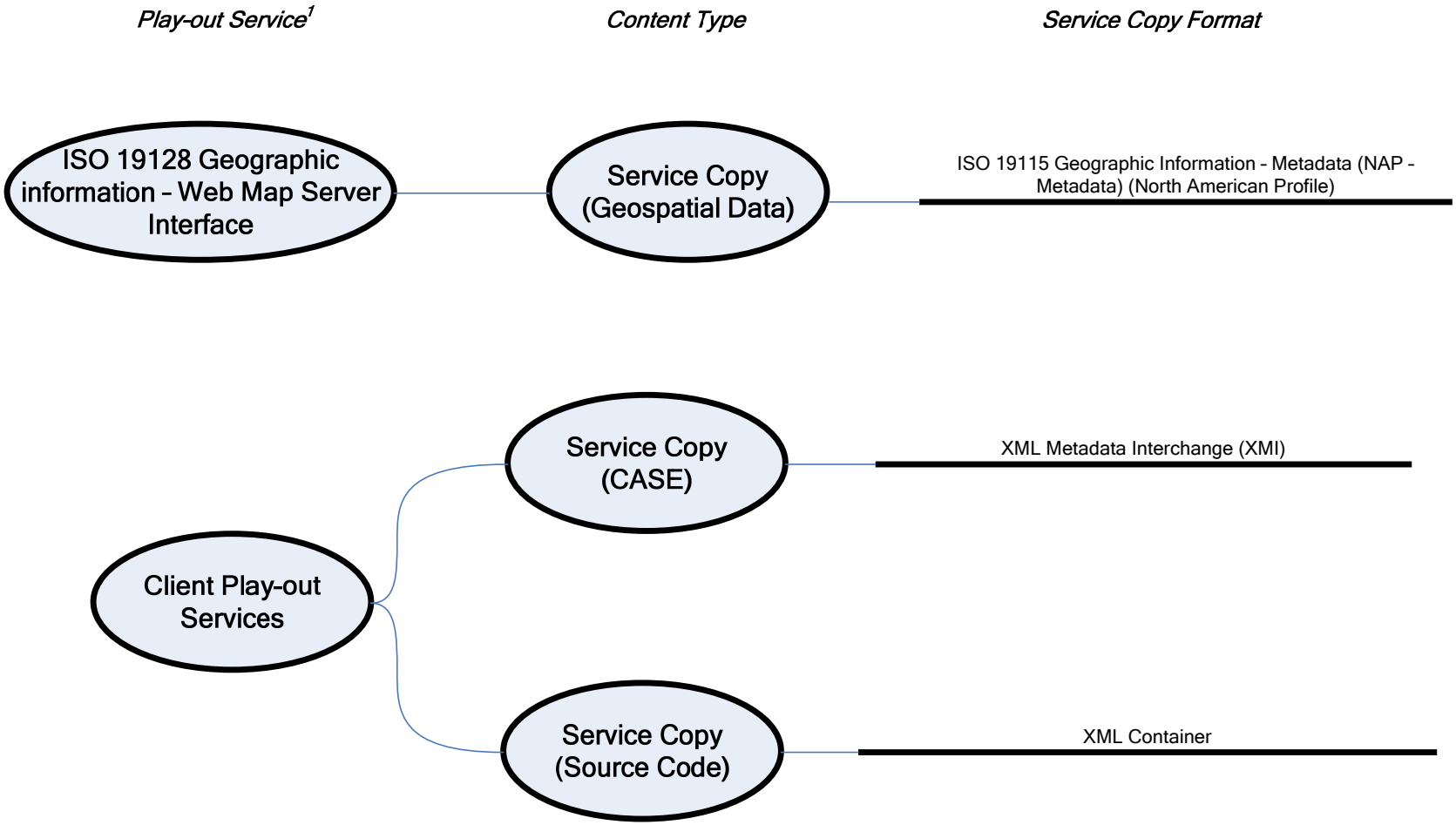
[1] In all instances, the option of a "client" play-out service is always available (e.g., "open with/save as")

Figure 13: LAC Format Guidelines - Mapping to Play-out Services[1]



[1] In all instances, the option of a "client" play-out service is always available (e.g., "open with/save as")

Figure 14: LAC Format Guidelines - Mapping to Play-out Services[1]



[1] In all instances, the option of a "client" play-out service is always available (e.g., "open with/save as")

4.3 Appendix C: Concepts and Definitions

4.3.1 Codecs[80]

A codec encodes a data stream or signal for transmission, storage or encryption, or decodes it for playback or editing. Codecs are used in videoconferencing and streaming media applications. A video camera's analog_to_digital converter (ADC) converts its analog signals into digital signals, which are then passed through a video compressor for digital transmission or storage. A receiving device then runs the signal through a video decompressor, then a digital_to_analog converter (DAC) for analog display. The term codec is also used as a generic name for a video conferencing unit.

4.3.2 Compression

In discussing file types, it is important to distinguish between file formats and compression algorithms. A codec (**co**mpressor/**dec**ompressor) is an algorithm that is installed on a PC and is capable of compressing or decompressing data in order for the file to be played by a particular software package. Codecs are usually applied to still imagery, moving imagery, or audio files, and can be further subdivided into lossy or lossless compression methods.

A ***lossless*** compression technique discards no information. It looks for more efficient ways to represent data, while making no compromises in accuracy. In contrast, ***lossy*** compression accepts some degradation in the data in order to achieve smaller file sizes. Because of this degradation in quality, lossy compression should be avoided for archival master images.

A codec is usually wrapped by a container, and most can be used to compress several different data formats. Conversely, most data formats can use different codecs. For example, an AVI file ‘container’ can use the DivX video codec, the MP3 audio codec or it could use the Indeo video codec or the PCM audio codec.

In some cases the codec and format are inherently linked, as is the case with proprietary software such as Real Player and Windows Media Player. In the case of MP3, AAC, or JPEG (and others) the codec can stand alone, or be wrapped in a different file format.

Most image compression techniques are independent of file formats, however some (JPEG, PNG) are inherently linked with a file format.

Library and Archives Canada is in the process of completing more extensive research in terms of compression techniques and algorithms and will provide more comprehensive advice in the future.

The most common compression algorithms for imagery consist of:

Run-Length Encoding (RLE)

> Run-length encoding is a simple lossless data compression method. It is well suited to standard palette-based images, but does not work well on continuous-tone images such as photographs.
>
> RLE is used in fax machines, and is relatively efficient because most faxed documents are primarily composed of white space with occasional interruptions of black.

Lempel-Ziv (LZ)

> The Lempel-Ziv (LZ) compression methods were developed by Abraham Lempel and Jacob Ziv in the late 1970’s, and are among the most popular algorithms today for lossless storage.

[80] http://en.wikipedia.org/wiki/Codec

Deflate

Deflate is a variation on LZ which is optimized for decompression speed and compression ratio, although compression can be slow. DEFLATE is used in PKZIP, gzip and PNG.

LZ-Reneau (LZR)

The LZR method forms the basis of the Zip compression scheme.

LZW Lempel-Ziv-Welch

LZW is a universal lossless data compression algorithm created by Abraham Lempel, Jacob Ziv, and Terry Welch in 1984, and was the first widely used universal data compression method. LZW is effective on 1-bit (monochrome) to 24-bit (True Colour) images.

Several file formats utilize LZW for compression, although it is most widely known as the compression algorithm for GIF files. More recently, an implementation of this algorithm is found within the popular Adobe Acrobat software to create PDF files.

CCIT Comité Consultatif International Téléphonique et Télégraphique

The CCITT, now known as ITU-T (ITU Telecommunication Standardization Sector) is the leading publisher of telecommunication technology, regulatory and standard information.

CCITT T.4 (commony referred to as Group 3 compression) is the universal protocol for sending fax documents across telephone lines. There are two levels of resolution: 203 by 98 and 203 by 196, and he compression protocol specifies a maximum transmission rate of 9 600 baud.

CCITT T.6 (commonly referred to as Group 4 compression) is a protocol for sending fax documents over ISDN networks. The Group 4 protocol supports images with a resolution of up to 400 dpi.

Both Group 3 and Group 4 are most commonly used by the TIFF file format. Several other ITU-T transmission standards exist; for more information go to: http://www.itu.int/home/index.html

Huffman Encoding

Huffman encoding is a lossless compression method developed in 1952 by David Huffman, and is one of the oldest and most established compression algorithms. Huffman encoding is designed to work best on images that have a lot of repetition, and is often used as a final compression stage in combination with more modern compression schemes such as JPEG, Deflate, and CCITT Group 3.

4.3.3 Character Sets

Character sets refer to the system for encoding a sequence of characters in an eight bit byte, or octet. Traditionally, character sets and character encoding are considered to be synonymous terms.

Recommended

American Standard Code for Information Interchange (ASCII) [ISO/IEC 8859-1:1998 (Latin-1)]

LAC supports the use of the ISO/IEC 8859-1:1998 ASCII character set for encoding. The standard defines a set of 256 characters where each character is defined using an 8-bit byte.

Extended Binary Coded Decimal Interchange Code (EBCDIC)

EBCDIC is an encoding schema that is used by IBM mainframe computers. The character set was developed in the 1960s and similar to ASCII, it uses an 8 bit binary code to represent up to 256 characters. The character set comes in six slightly different forms, but it is still being used today

on IBM mainframes. Detailed information on EBCDIC can be found in the IBM publication *IBM Character Data Representation Architecture, Reference and Registry, SC09-2190-00*, December 1996.

Unicode Version 3.0 UTF-8 [ISO/IEC 10646-1:2000]

LAC supports the Unicode version 3.0 standard that defines a multi-octet character set called the Universal Character Set (UCS). Unicode 3.0 UTF-8 (UCS Transformation Format - 8) provides a unique number for up to 49,194 characters, regardless of the platform, program or language. Unicode 3.0 has been updated by later versions of the standard. These updates do not replace the bulk of the existing material of Unicode 3.0. These revisions add characters, correct or extend the character properties in the Unicode Character Database or have significance for the interpretation of some aspects of the standard. The Unicode standard is recommended by LAC because it provides the default UCS encoding schema for HTML, SGML, XHTML and XML.

4.3.4 Well formed/Well formedness[81]

In web page design, and generally for all markup languages such as SGML, HTML, and XML, a well-formed element is one that is either

- opened and subsequently closed,
- an empty element, which in that case must be terminated,
- properly nested so that it does not overlap.

For example, in HTML: word is a well-formed element, while word is not, since the bold element is not closed. In XHTML, empty elements (elements that inherently have no content) should be closed by putting a slash at the end of the opening tag, e.g. <img />,
, <hr />, etc. In HTML, there is no closing tag for such elements, and no slash is added to the opening tag.

Furthermore, if an element has any attributes, each attribute value must be properly quoted.

In a well-formed document,

- all elements are well-formed, and
- a single element, known as the root element, contains all of the other elements in the document.

For example, the code below is not well-formed HTML, because the em and strong elements overlap:

```
<!-- WRONG! NOT WELL-FORMED HTML! -->
<p>Normal <em>emphasized <strong>strong emphasized</em> strong</strong></p>
```

```
<!-- Correct: Well-formed HTML. -->
<p>Normal <em>emphasized <strong>strong emphasized</strong></em>
<strong>strong</strong></p>
<p>Alternatively <em>emphasized</em> <strong><em>strong emphasized</em>
strong</strong></p>
```

In XML, the phrase well-formed XML document is often used to describe a text that follows all the syntactic rules labelled as well-formedness rules in the XML specification: strictly speaking the phrase is tautological, since a text that does not follow these rules is not an XML document. The rules for well-

[81] http://en.wikipedia.org/wiki/Well-formed_element

formedness go beyond the requirements mentioned above to nest tags properly and to quote attribute values: they include, for example, rules about the characters that can appear in names and elsewhere, the syntax of comments, processing instructions, entity references, and CDATA sections, and many other similar details. Sometimes the adjective well-formed is used to contrast with valid: a valid XML document is one that is not only well-formed, but also conforms to the grammar defined in its own DTD (Document Type Definition).

4.3.5 Document Validity

Valid XML[82]: A valid XML document respects the rules dictated by a DTD or XML schema. In addition, extended tools are available such as OASIS CAM standard specification that provide contextual validation of content and structure that is more flexible than basic schema validations.

Valid Web Pages[83]: Most pages on the World Wide Web are written in computer languages (such as HTML) that allow Web authors to structure text, add multimedia content, and specify what appearance, or style, the result should have.

As for every language, these have their own grammar, vocabulary and syntax, and every document written with these computer languages are supposed to follow these rules. The (X)HTML languages, for all versions up to XHTML 1.1, are using machine-readable grammars called DTDs, a mechanism inherited from SGML.

However, just as texts in a natural language can include spelling or grammar errors, documents using Markup languages may (for various reasons) not be following these rules. The process of verifying whether a document actually follows the rules for the language(s) it uses is called validation, and the tool used for that is a validator. A document that passes this process with success is called valid.

With these concepts in mind, we can define "markup validation" as the process of checking a Web document against the grammar (generally a DTD) it claims to be using.

[82] http://en.wikipedia.org/wiki/XML_validation
[83] http://validator.w3.org/docs/help.html#validation_basics