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).

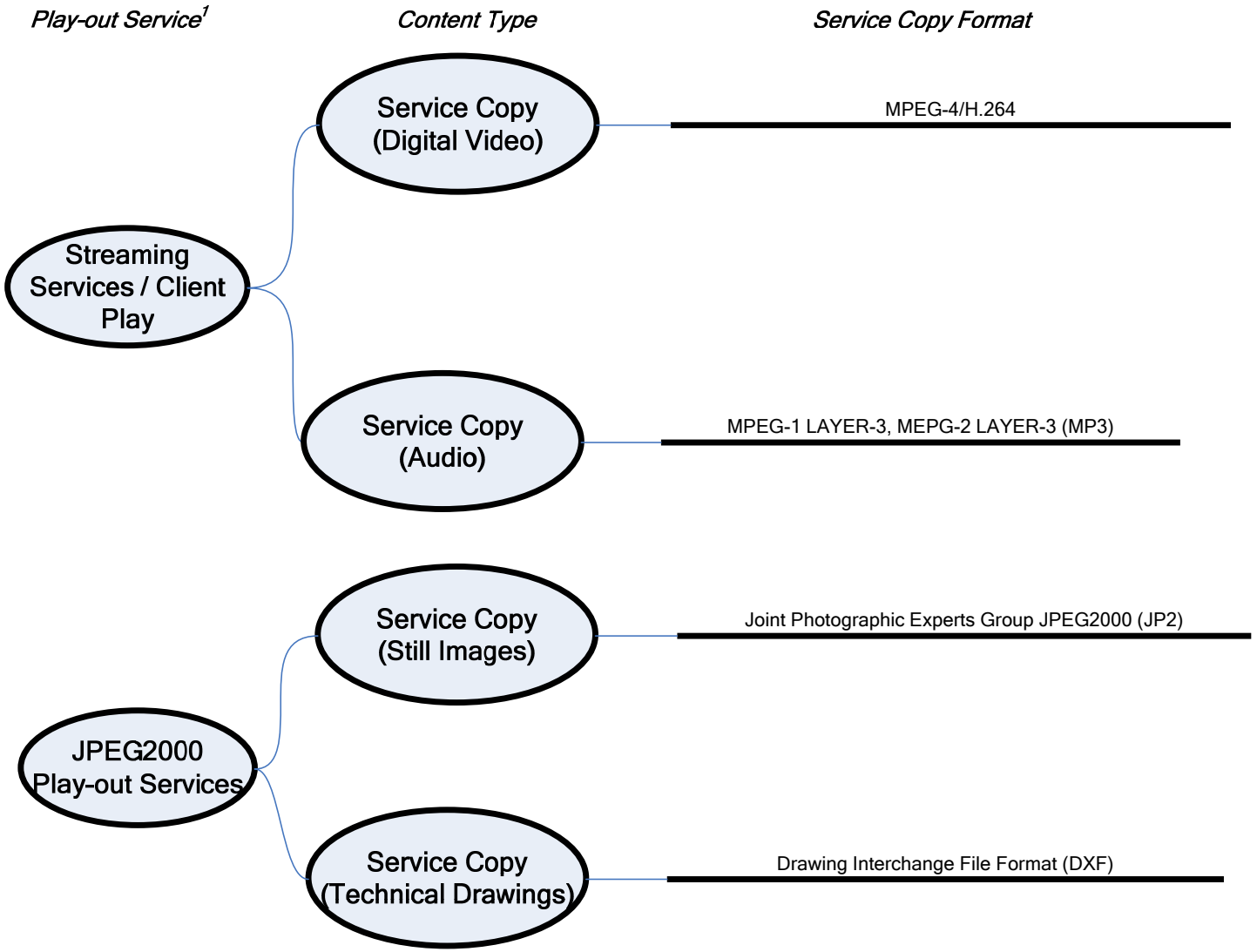
Table 4 summarizes the ratings of LAC's recommended file formats against the criteria identified in Section 1.5.2, whereas Appendix A – Recommended Preservation Format Evaluation provides detailed rating information. Please note that there is no implied order of preference / precedence in the list of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

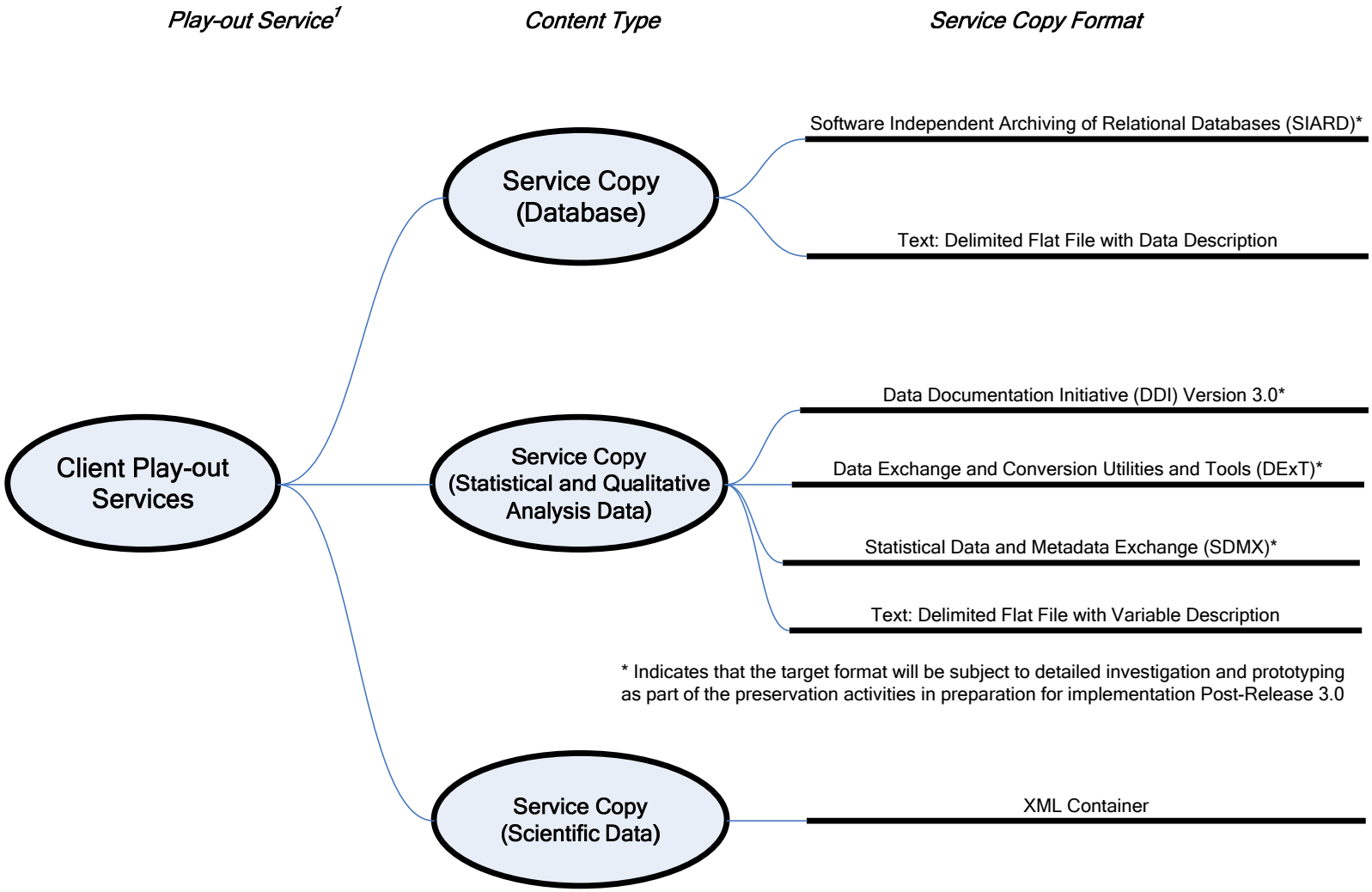
Criterion	Evaluation Basis	Rating
Openness/Transparency	Specifications available from one or more of the following: a) Open membership organization (such as the W3C (World Wide Web Consortium), the OMG (Object Management Group)) b) International standards organization (such as the ISO) c) Industry-based 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	 ✓

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



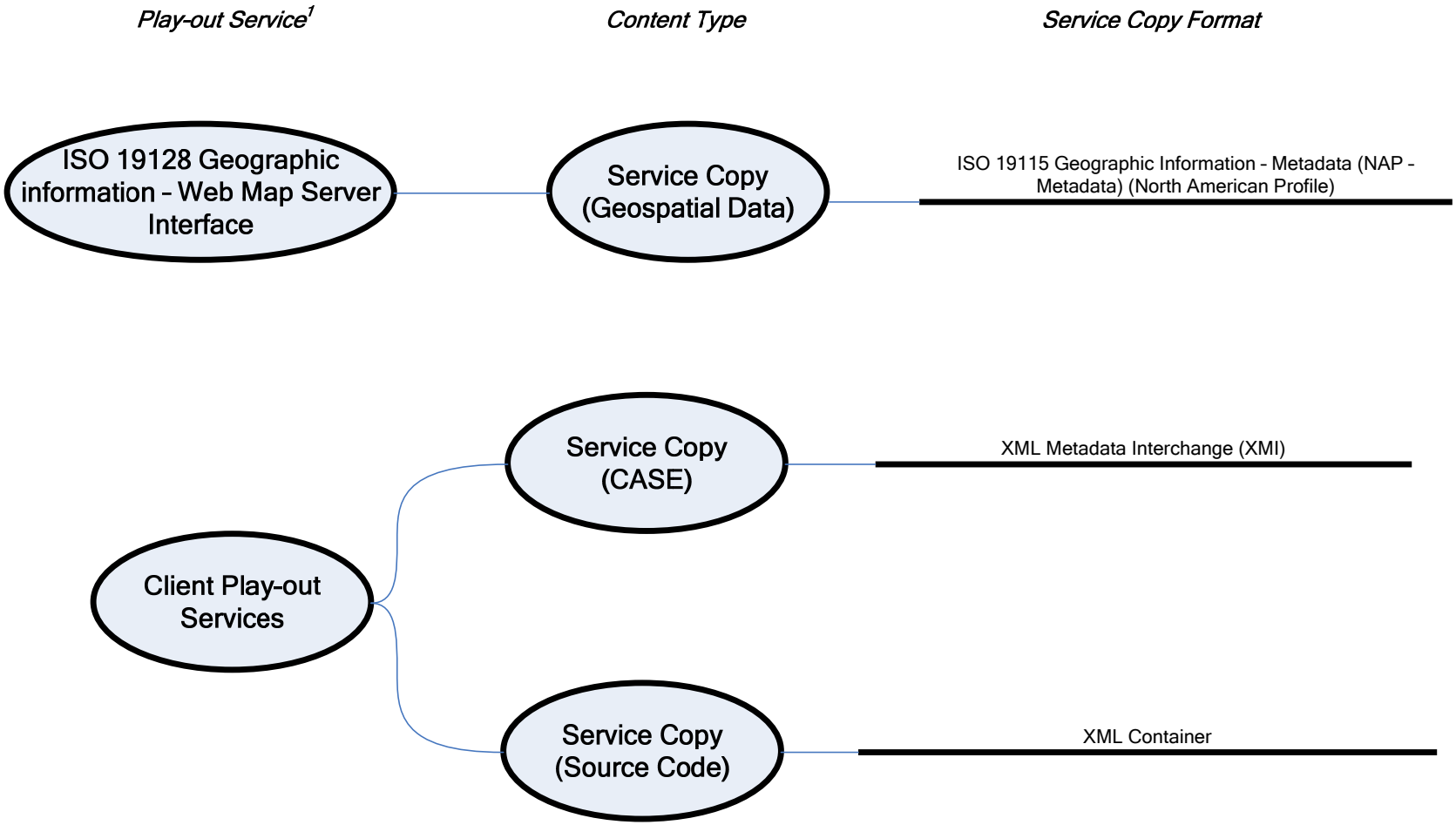
[1] In all instances, the option of a "client" play-out service is always available (e.g., "open with/save as")

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 />, <br />, <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