Digital Formats
Factors for Sustainability, Functionality, and Quality

Caroline R. Arms
Library of Congress, Office of Strategic Initiatives

Long Term Knowledge Retention: Archival and Representation Standards
NIST, March 15, 2006

The Library of Congress

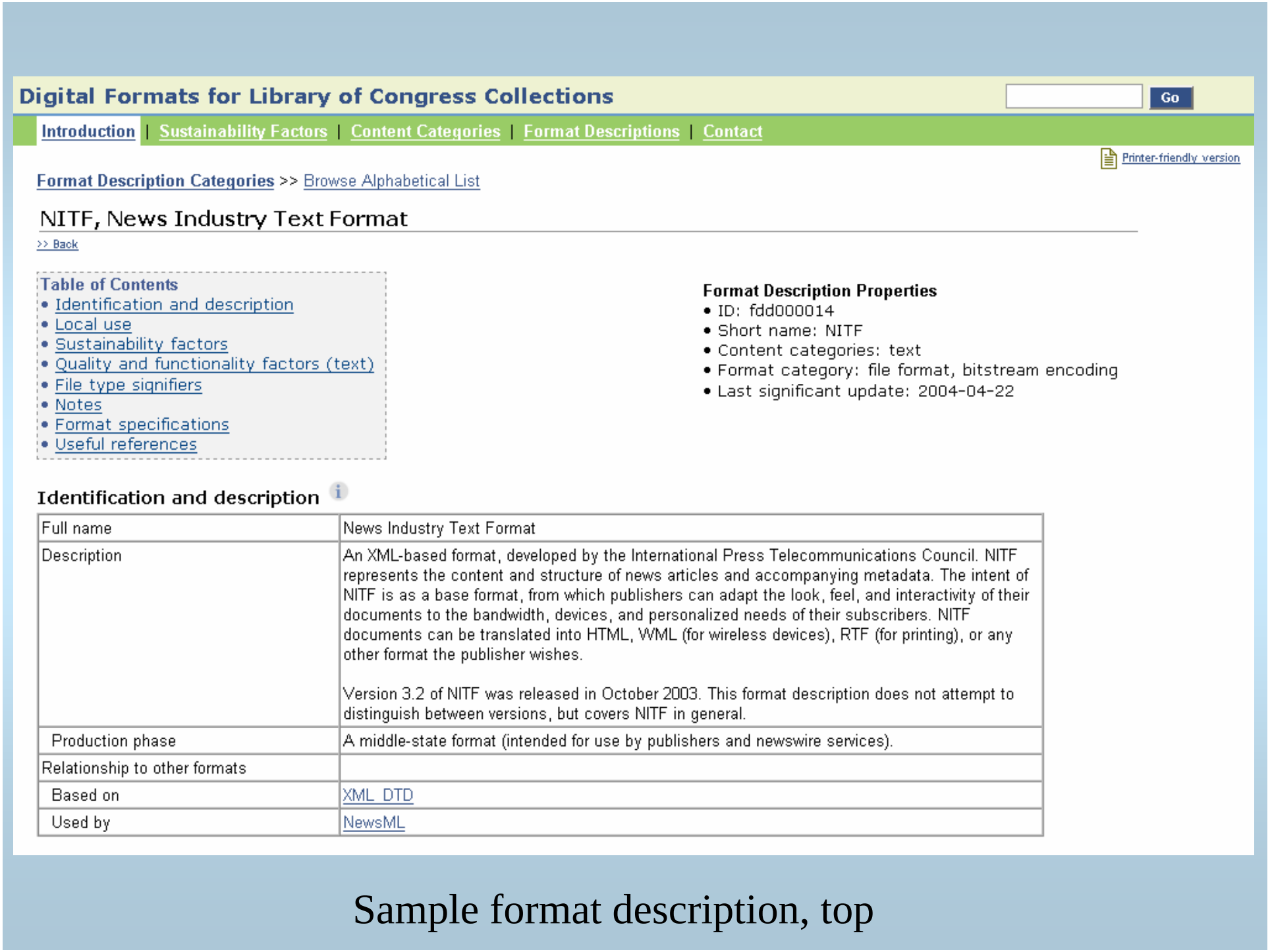
- Collects materials created by others
- Acquires the majority though copyright registration/deposit
  - adds approximately 10,000 items to the collections daily.
- Has very limited rights to make copies of or transform content it receives though copyright deposit during the period of copyright protection
- Can require under Section 408 of the Copyright Act (Mandatory Deposit) that publishers deposit the "best" edition they publish, but cannot set a quality threshold
- Collects much more than books
  - 29 million books and other printed materials, 2.7 million recordings, 12 million photographs, 4.8 million maps, 5 million music items and 58 million manuscripts.

Analysis of Digital Formats

- Provide inventory of information about formats
- Identify and describe the formats promising for long-term sustainability
- Web site: http://www.digitalpreservation.gov/formats/

Formats descriptions as of July 2004

AAC_MP2 (Advanced Audio Coding, MPEG-2)
AAC_MP4 (Advanced Audio Coding, MPEG-4)
AAC_ADIF (Advanced Audio Coding, MPEG-2, Audio Data Interchange Format)
AAC_M4A (Advanced Audio Coding, MPEG-2, m4a File Format)
AIFF (Audio Interchange File Format)
AIFF_LPCM (AIFF File Format with LPCM Audio)
ASF (Advanced Systems Format)
AudCom (Audible.Com File Format)
AudCom_MP3 (Audible.Com MP3)
AVI (Audio Video Interleaved)
AVI_MJPEG (AVI, MJPEG Codec)
AVI_Indeo (AVI, Indeo Codec)
AVI_Cinepak (AVI, Cinepak Codec)
AVI_DivX (AVI, DivX Codec)
Cinepak (video codec)
DLS, Downloadable Sounds Format
DivX_5, Version 5 (video codec)
ID3 (ID3 Metadata for MP3)
ID3v1 (ID3, version 1)
ID3v2 (ID3, version 2)
IFF (Electronic Arts Interchange File Format 1985)
Indeo_3, Version 3 (video codec)
Indeo_5, Version 5 (video codec)
LPCM
MIDI_SD, MIDI Sequence Data
MJPEG (Motion JPEG)
MODS, Module Music Format (Mods)
MP3_ENC (MP3 Encoding)
MP3_FF (MP3 File Format)
MPEG-1
MPEG-2
MPEG-2_SP, Simple Profile
MPEG-2_MP, Main Profile
MPEG-2_422, 4:2:2 Profile
MPEG-4
MPEG-4_V, Visual Coding (Part 2)
MPEG-4_V_SP, Visual Coding, Simple Profile
MPEG-4_V_SSP, Visual Coding, Simple Scalable Profile
MPEG-4_V_ASP, Visual Coding, Advanced Simple Profile
MPEG-4_V_CP, Visual Coding, Core Profile
MPEG-4_V_MP, Visual Coding, Main Profile
MPEG-4_V_SStP, Visual Coding, Simple Studio Profile
MPEG-4_AVC, Advanced Video Coding (Part 10)
MPEG-4_AVC_BP, Advanced Video Coding, Baseline Profile
MPEG-4_AVC_MP, Advanced Video Coding, Main Profile
MPEG-4_AVC_EP, Advanced Video Coding, Extended Profile
NITF, News Industry Text Format
Ogg, Ogg File Format
Ogg_Vorbis, Ogg Vorbis Audio Format
PCM
PDF, Portable Document Format
PDF/A, PDF for Preservation
Quicktime
QTA_MP3, QuickTime Audio, MP3 Codec
QTA_AAC, QuickTime Audio, AAC Codec
QTV_Apple, QuickTime Video, Apple Codec
QTV_Cinepak, QuickTime Video, Cinepak Codec
QTV_Sorenson, QuickTime Video, Sorenson Codec
QTV_MJPEG, QuickTime Video, Motion JPEG Codec
QTV_MPEG, QuickTime Video, MPEG-1 Codec
RealAudio_10, Version 10
RealAudio_RA, RealAudio Codec
RealAudio_AAC, AAC Codec
RealAudio_LL, Lossless Codec
RealAudio_MC, Multichannel Codec
RealVideo_10, Version 10
RMID, RIFF-based MIDI File Format
Sorenson_3, Version 3 (video codec)
SMF, Standard MIDI File Format
SVG, Version 1.1
TIFF, Revision 6.0 and earlier
Vorbis, Vorbis Audio Codec
WAVE
WAVE_LPCM
WAVE_LPCM_BWF
WMA, Windows Media Audio
WMA_WMA9, Windows Media Audio File with WMA9 Codec
WMA_WMA9_PRO, Windows Media Audio File with WMA9 Professional Codec
WMAWMA9_LL, Windows Media Audio File with WMA9 Lossless Codec
WMA9, Windows Media 9 Audio Codec
WMA9_PRO, Windows Media 9 Professional Audio Codec
WMA9_LL, Windows Media 9 Lossless Audio Codec
WMV, Windows Media Video
WMV_WMV9, Windows Media Video with WMV9 Codec
WMV_WMV9_PRO, Windows Media Video with WMV9 Professional Codec
WMV9, Windows Media 9 Video Codec
WMV9_PRO, Windows Media 9 Professional Video Codec
XMF, eXtensible Music Format
XML



Sample format description, top

Local use

LC experience or existing holdings	None
LC preference	Suggested as a preferred format for textual news feeds, individual news articles, etc..

Sustainability factors

Disclosure	Open standard. NITF was developed by the International Press Telecommunications Council, an independent international association of news agencies and publishers.
Documentation	NITF specifications can be found at http://www.nitf.org/specifications.php
Adoption	Substantial adoption by newswire services. See http://www.nitf.org/users.php.
Licensing and patent claims	None
Transparency	Human-readable XML. Well-documented DTD.
Self-documentation	The DTD includes descriptive elements relevant to news articles, such as title, byline, dateline, headline. Also in the DTD are elements for topical subjects, publication details, and revision history.
External dependencies	None
Technical protection considerations	None

Quality and functionality factors (text)

Normal rendering	Good support.
Integrity of structure	The NITF DTD represents the structure of an individual news article.
Integrity of layout	Not intended to define precise rendering, but to support rendering using stylesheets for different disseminations, via PDAs, RSS, web browsers, etc. Includes HTML-compatible support for tables and lists.
Integrity of rendering of equations, etc.	Not supported.
Beyond normal rendering	Supports links to external media objects, such as images, audio, or video. DTD also supports embedding of media objects (in binary format).

File type signifiers

Tag type	Value	Note
Filename Extension	xml	Common practice for XML document instances is to use the .xml extension. The particular XML Schema or DTD should be declared within the document.

Sample format description, middle

Format specifications

URLs

- http://www.nitf.org/IPTC/NITF/3.2/dtd/nitf-3-2.dtd. NITF DTD (Document Type Definition)
- http://www.nitf.org/IPTC/NITF/3.2/documentation/nitf-documentation.html Documentation for NITF

Print

Useful references

URLs

- http://www.nitf.org/
- http://www.iptc.org/

Print

Last updated Friday, 04-Jun-2004 11:20:07 EDT

Sample format description, bottom

Formats: Types & Relationships

- file formats
  - at the level indicated by file extensions, e.g., .mp3
  - as indicated by Internet MediaType (aka MIME type), e.g. text/html
  - versions introduced over time
  - refinements are tailored to specific purposes, e.g., TIFF-EP for electronic photography
- class of related formats whose familial characteristics are important
  - e.g., the WAVE audio format is an instance of the RIFF format class
- "wrappers" distinguished in terms of their underlying bitstreams
  - e.g., WAVE files may contain linear pulse code modulated [LPCM] audio (like a CD) or highly compressed audio as used for digital telephony.
- bundling formats bind together files comprising a single digital work
  - e.g., text and supporting illustrations, or a movie with sound tracks in different languages

Simple Example: TIFF

- Wrapper for different bitstreams
- Simple, but extensible method for embedding metadata

may contain	Uncompressed bitmap, LZW compressed bitmap, bitonal Group IV (bitstreams)
has subtype	TIFF/EP (for electronic photography)
has subtype	TIFF/IT (for prepress applications)
has subtype	DNG (Adobe’s proposed format for digital negatives)

More Complex Example -- PDF

Much more than text

A file format, a wrapper, a bundling format, all in one

Complexity of relationships

has subtype	v.1.3 (July 2000, 696 pages)
has subtype	v.1.4 (December 2001, 978 pages)
has subtype	v.1.5 (August 2003, 1172 pages)
has subtype	v.1.6 (November 2004, 1236 pages)
may contain	TIFF, JPEG, JPEG2000, etc., etc., etc. (all at once)
has subtype	Tagged PDF (can represent logical document structure)
has subtype	Accessible PDF (tagged + further constraints)
has subtype	PDF/X (ISO standard, for pre-press use, e.g., submission of graphics to magazine publishers)
has subtype	PDF/A (Under development as ISO standard, for archiving)

Complexity Increasing

- New standards have portmanteau nature
  - Many parts, many options, as already noted in the case of PDF
- JPEG2000
  - Part 1. .jp2 (core lossless and lossy compression schemes for continuous tone, replacement for JPEG)
  - Part 2. .jpx (extensions, including more capabilities for embedding metadata)
  - Part 6. .jpm (multi-layer images, can embed other bitstream encodings, including bitonal)
- MPEG-4
  - Many *profiles* for different contexts, also *advanced video coding*
- Which parts of these standards will be widely adopted?

Two Types of Evaluation Factors

- **Sustainability factors** for all formats
  - influence feasibility and cost of preserving content in the face of future change
- **Quality and functionality factors** that vary by content category
  - reflect considerations that will be expected by future users

Sustainability: Disclosure

- Disclosure refers to the degree to which complete specifications and tools for validating technical integrity exist and are accessible.
- Preservation of content in a given digital format is not feasible without an understanding of how the information is encoded.
- Non-proprietary, open standards are usually more fully documented and more likely to be supported by tools for validation than proprietary formats.
- However, what is most significant for sustainability is not approval by a recognized standards body, but the existence of complete documentation.

Sustainability: Adoption

- Adoption refers to the degree to which the format is already used by the primary creators, disseminators, or users of information resources. A widely adopted format is less likely to become obsolete rapidly, and tools for migration and emulation are more likely to emerge without specific investment by archival institutions.
- Examples for bitmapped images:
  - TIFF uncompressed, widely recommended as master
  - PDF/X, increasingly required for submission to magazines, etc.
  - JPEG2000 Part 1, increasingly adopted

Sustainability: Transparency

- Degree to which the digital representation is open to direct analysis with basic tools, such as human readability using a text-only editor.
- Digital formats in which the underlying information is represented simply and directly will be easier to migrate to new formats, more susceptible to digital archaeology, and allowing easier development of rendering software.
- Examples:
  - Uncompressed raster image bitstream easy to interpret or reverse engineer
  - Lossy compressed image bitstream requires algorithm to decode

Sustainability: Self-documentation

- Self-documentation. Digital objects that contain basic descriptive metadata (the analog to the title page of a book) as well as technical and administrative metadata will be easier to manage over the long term than data objects that do not incorporate the metadata needed to render or understand them.
- Some metadata elements will likely be extracted to support discovery and collection management.
- Examples:
  - JPEG (.jpg) image files contain very scant metadata
  - EXIF JPEG combines JPEG compression with richer metadata
  - JPEG2000 (.jpx) image files may contain metadata ‘boxes’ and can include an extensive DIG35 record
    - OpenGIS Consortium just approved and released specifications for a standard way to embed geospatial metadata into JPEG2000 images

Sustainability: External Dependencies

- Degree to which a particular format depends on particular hardware, operating system, or specialized software for rendering or use and the predicted complexity of dealing with those dependencies in future technical environments.
- Some interactive digital content is designed for use with specific hardware, such as a joystick.

Sustainability: Impact of Patents

- Degree to which the ability of archival institutions to sustain content in a format will be inhibited by patents.
- Although the costs for licenses to decode current standard formats are often low, the development of open source decoders will be inhibited. Tools to transcode content in these formats when they become obsolete may be more costly to develop.
- It is not the existence of patents that is a potential problem, but the terms that patent-holders might choose to apply.
- Examples:
  - MrSID patent exploited through licensing terms with fees depending on transaction volume

Sustainability: Tech Protection Mechanisms

- Refers to the *implementation* of mechanisms such as encryption that prevent the preservation of content by a trusted repository.
- Preservation of the digital content requires replicating it on new media, migrating and normalizing it in the face of changing technology. Protection mechanisms may also prevent the dissemination of content to authorized users.
- Exploitation of technical protection mechanisms is generally optional; their use depends in a particular context may depend on business decisions.

Quality and Functionality Factors

- Vary according to content type, e.g., text, image, sound
- Pertain to current and future usefulness, e.g., for scholarship or repurposing
- Identification of factors reflects consideration of what are likely to be significant or essential features of some content items
  - Surround sound for audio
  - Color maintenance for still images
  - Logical structure for text documents
- User expectations establish what might be called "normal rendering" for a given genre or form of expression for content.

Quality/functionality for text

- Normal rendering for text:
- For users (display):
  - Convenient linear reading on screen,
  - the ability to
    - print sections of the document to paper,
    - excerpt quotations as text strings,
    - search for words within a document.
  - Rendering of any text item must reflect the intent of the author in representing the individual characters, paragraph structure, lists, headings, and indicators of emphasis.
- For applications (index/search)
  - Text indexable to support searching across a corpus of documents.

Additional normal rendering functionality

- Important for many text categories
- Support for integrity of document structure and navigation
  - Essential for directories, dictionaries, encyclopedias
  - Important for e-mail, news feeds
  - Desirable for almost any text
- Support for integrity of layout, font, and other design features
  - Essential for posters, advertisements, brochures
  - Often irrelevant for articles, technical reports
- Support for rendering for mathematics, formulae, diagrams, etc.
  - Crucial if present

Finding the balance

- Preferences will be based on balancing the factors.
- Factors compete; trade-offs will come into play
- No uniformly “best” format for text
- Depends on genre of text work
  - Encyclopedia
  - Scholarly article
  - Work conceived as hypertext.
- Selection of acceptable formats will have to consider how digital content will be acquired by LC.
  - For copyright registration, JPEG must be acceptable, because many digital cameras do not produce uncompressed images
  - For some content of high cultural value, for example the working files of a composer of electronic music, particular functionality may outweigh sustainability factors

Content States in a Production Process

- A simplified view of a publishing or distribution stream sees a content item as existing in three states,
- Different formats are often associated with these three states, appropriate to the task at hand.
  - Initial: author creates
  - Middle: publisher manages and archives
  - Final: end user receives

Middle-state Formats and Sustainability

- The Library of Congress has typically collected final-state formats

- Best formats for long term may be middle-state formats.
  - Likely to have higher quality than final-state formats,
  - May incorporate metadata useful to support preservation
  - Archiving and preservation practices may emerge from industry.

Not an isolated effort

- Our analysis is intended for human readers
- Plan to exploit synergy with efforts at building automated, systems for managing information about formats and tools that can validate, characterize, and transform content in those formats
  - Global Digital Formats Registry
    - http://hul.harvard.edu/gdfr/
  - JHOVE, toolkit for object characterization and validation
    - http://hul.harvard.edu/jhove/
  - Projects funded through LC's National Digital Information Infrastructure and Preservation Program (NDIIPP)

Global Digital Format Registry (GDFR)

The concept of representation format permeates all technical areas of digital repository architecture and operation. Policy and processing decisions regarding ingest, storage, access, and preservation are frequently, if not uniformly, conditioned on a format-specific basis. Proper interpretation of otherwise opaque digital content streams is dependent upon knowledge of how typed content is represented. The current IANA MIME media type registry does not capture format-specific information at an appropriate level of granularity, or in sufficient level of detail, for many digital repository activities. The international digital library and archival co
formats for the purposes of fulfilling their mission. This r
rigorous in validity, public in discovery and delivery, an

Home | News | Use Cases | Data Model | Service Moc

File: http://hul.harvard.edu/gdfr/home.html
Modified: 2003-12-24

JHOVE

Home | | | | |

Format-Specific Digital Object Validation

1 Introduction

eas of digital repositories. Policy and
vation are frequently conditioned on a
repositories need to be able to automate

Digital Formats for Library of Congress Collections Go

Introduction | Sustainability Factors | Content Categories | Format Descriptions | Contact

Format Description Categories >> Browse Alphabetical List

Format Descriptions

Still Image
- SVG 1.1
- TIFF 6
- All still image format descriptions

Sound
- WAVE
- MP3 FF
- All sound format descriptions

Generic
- ASF
- RIFF
- All generic format descriptions

Textual
- NITF
- XML
- All text format descriptions

Moving Image
- MPEG-4 FF 2
- AVI
- All moving image format descriptions

Thank you . . .

http://www.digitalpreservation.gov/formats/