Digital Formats: Factors for Sustainability, Functionality, and Quality

Caroline Arms and Carl Fleischhauer
Office of Strategic Initiatives, Library of Congress
Washington, DC, USA

Abstract

The Library of Congress is drafting a decision-support framework pertaining to the preservation of digital content. The framework is presented through a Web site that identifies and documents digital content formats that are promising (or unpromising) for long-term sustainability, together with some explanatory essays. The resource is intended to serve staff who evaluate born digital content for selection for the Library's collections and make provisions to sustain that content.

The initial investigation has outlined two sets of high-level factors that may be used when choosing formats:

- conceptual factors that may affect the sustainability of any digital format
- factors that relate to quality or special functionality that might be desired for certain categories of content

Introduction

The authors are engaged in an ongoing analysis of digital formats. We began with some goals in mind for the Library of Congress as it builds its digital collections:

- to support planning and decision-making,
- to provide an inventory of information about current and emerging formats, and
- to identify and describe the formats that are promising for long-term sustainability, and develop strategies for sustaining these formats.

The results of our analysis are made available on a Web site [http://www.digitalpreservation.gov/formats/]. This online resource is growing as we consider and document additional formats and as new standards are published.

Our focus is on digital content formats that are independent of the physical medium on which they are stored or transported. Content in such formats exists as data files or data streams. Out of scope are audio CDs and DVDs; in scope are MP3 audio files, familiar formats such as TIFF, PDF, and SGML and newer formats such as JPEG2000 and MPEG-4. The intent of our resource is to support human decision-makers. However, we have been working closely with those planning for a Global Digital Format Registry (GDFR).[1] The GDFR effort aims for an active registry that will support the execution of operations on files, to identify, validate, and even transform them. Our work, the proposed GDFR, and the development of the JHOVE toolset[3] for format characterization and validation are intended to be complementary.

Our Web site includes explanatory essays and other discussions, tables representing Library of Congress preferences, and a growing inventory of structured fact sheets, describing individual formats.

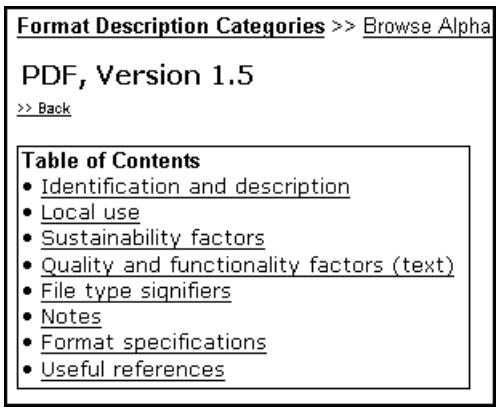

Fig. 1: Contents of sample format description document

Relationships and types for formats

The list of *format description documents* is already long, well over 150. We believe that in order for custodians to preserve content in digital form, they must be able to distinguish between format refinements and variants that are significant to sustainability, functionality, or quality. Formats have versions, subtypes, and dependencies on other formats. TIFF provides a relatively simple example. TIFF *may contain* bitmaps represented by a number of different bitstream encodings: uncompressed, compressed using the lossless LZW algorithm, or, for a bitonal image, compressed using ITU G4 compression. Future migration or transformation of a bitonal G4 TIFF will likely use a different target format than that for a 24-bit uncompressed TIFF. In addition, TIFF has subtypes TIFF/EP (an ISO standard for electronic photography and TIFF/IT (an ISO standard for exchanging prepress

images). A more complex example is Adobe's Portable Document Format (PDF). PDF can act as a relatively straightforward format for paginated text, a wrapper for many different image formats, or a bundling format for complex documents and interactive multimedia.

Table 1: Relationship examples for PDF

Format	Relationship	Related Format
PDF	has subtype	PDF, version 1.3 (July 2000)
PDF	has subtype	PDF, version 1.4 (December 2001)
PDF	has subtype	PDF, version 1.5 (August 2003)
PDF	has subtype	PDF, version 1.6 (November 2004)
PDF	may contain	TIFF, JPEG, JPEG2000, (possibly all at once)
PDF	has subtype	Tagged PDF (can represent logical document structure)
PDF	has subtype	Accessible PDF (tagged + further constraints)
PDF	has subtype	PDF/X (ISO standard 15930, for prepress use)
PDF	has subtype	PDF/A (Proposed ISO standard 19005, for long-term preservation)
PDF 1.4	has earlier version	PDF 1.3
PDF 1.4	has later version	PDF 1.5

The commonly used format name, such as TIFF or PDF, offers insufficient discrimination for preservation purposes. Format names--and as well filename extensions like jpg, pdf, mov, and MIME types--are too generic to distinguish between significantly different subtypes and versions. This fact is reflected in the level of format detail offered by other resources or tools intended to support preservation of digital content, such as PRONOM (an online registry of file formats and their supporting software products from the UK's National Archives)[4], the data model for the Global Digital Format Registry, and its associated JHOVE software.

The scope of formats included and distinguished in our inventory is very broad. It includes not only formats at the level indicated by a file extension (e.g., .tif), but versions developed over time, refinements tailored to a particular use, and variants distinguished by different bitstream encodings, even if in a common wrapper. Also included are format classes, whose familial characteristics are important. The WAVE audio format is an instance of the RIFF format class. File formats for MPEG-4 and Motion JPEG2000 are both based on the ISO Base Media File Format, a newer format class. We also include format descriptions for bitstream encodings that may be incorporated into or used as the basis for various wrapper or bundling formats. Examples are LPCM (the closest equivalent in the audio realm to an uncompressed bitmap) and XML.

Other formats bind together files or objects comprising a single digital work, e. g., text and supporting illustrations or a movie with sound tracks in different languages. These *bundling formats* represent a bundle of files or bitstreams, usually listing the components and their relationships through what is sometimes called *structural metadata*. They often incorporate technical details about each component, since a single work may include a mix of texts, sound, images, etc. Bundling formats may be designed to encapsulate the component data streams or take the form of a separate file that accompanies the set of component files. Some emerging standards that play such a bundling role are intentionally generic; these include METS (Metadata Encoding and Transmission Standard) and MPEG-21. Other bundling formats, such as the Digital Talking Book Format, have a more constrained structure for a specific purpose.

Some observations

New formats are very complex. This is evident in the versions and subtypes for PDF; similar differentiations pertain to JPEG2000 and MPEG-4. The specifications for these and other emerging formats are published in multiple parts with multiple nuances. It is hard to predict which parts will be adopted and hence which subtypes offered to the Library of Congress because appropriate tools are available to creators. Digital works created in these formats are also complex. The auto manufacturer BMW has sponsored short films--famous on the Web—made by prominent directors, featuring well known actors and, of course, starring BMW cars. Several versions of these shorts can be downloaded. The “enhanced” QuickTime version is a particularly complex example. From a single *mov* file, you can switch from the normal soundtrack to a commentary track, display a text transcription, or switch over to what they call a virtual reality presentation that shows off the car in all its splendor. This QuickTime file, like its MPEG-4 counterparts, uses an object based design internally. The player lists all of the file’s elements (in effect, the objects in the file) under the properties setting.

Different formats are employed or favored in different stages of a content item's lifecycle. Albeit a bit of a simplification, it has proved useful to distinguish three states in a publishing or distribution stream:

- Initial: while the author is creating it
- Middle: while the publisher manages and archives it
- End: what is presented or sold to an end-user

Initial state formats are often proprietary and may be limited to the creator's favorite software package. These formats tend to be complex, for example, retaining information about current choices for cropping and

layering components of an image being prepared for advertising purposes. The native format for Adobe Photoshop is an example here. *Middle state formats* are used by industry to send or exchange data, as exemplified by the PDF/X or TIFF/IT files that a designer may employ when submitting digital art to a magazine. These prepress formats use separate layers to support color separation and spot color in ways compatible with printing technology. In other cases, a flattened bitmapped image at high resolution may be used as a master for future repurposing. Middle-state formats may emerge as preferred for archiving within an industry. *Final state formats* are for items in the marketplace and are often transient. A record company might say, "This year, we released the song in RealAudio, next year we'll probably sell it on iTunes as encrypted AAC." Depending on the delivery system, the disseminated files may even be generated dynamically from a master in response to a customer's particular requirements.

The authors hypothesize that the best formats from a preservation perspective will be those in the middle state. These are likely to have higher quality than final-state formats and may also be the focus of developing archiving approaches by industry. However, to seek middle state digital formats would represent a change in the Library's most widespread current practice, which is to select final state works, the best editions as authorized by copyright law. Implementation of a middle state preference by the Library will require negotiation with creators.

Factors to consider when choosing formats

In considering digital formats for the Library's collections, two types of factors come into play: *sustainability factors* and *quality and functionality factors*.

Sustainability factors apply across digital formats for all categories of information. We have identified seven factors that influence the feasibility and cost of preserving content. We believe that these factors will be significant whether preservation strategies entail future migration to new formats, emulation of current software on future computers, a hybrid of migration and emulation, or normalization on receipt.

Seven sustainability factors

1. *Disclosure* refers to the degree to which complete specifications and tools for validating technical integrity exist and are accessible to those creating and sustaining digital content. Preservation of content in a given format is not feasible without an understanding of how the information is encoded as bits and bytes in digital files. A spectrum of disclosure levels exists. 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 (and preservation of) complete documentation.

Examples:
- TIFF, well documented, many third-party tools
- MrSID, proprietary compression, only partially documented
- JPEG2000 Part 1, open standard, fully documented

2. *Adoption* refers to the degree to which the format is already used by the primary creators, disseminators, or users of information resources. A format that is widely adopted is less likely to become obsolete rapidly, and tools for migration and emulation are more likely to emerge from industry without specific investment by archival institutions. Evidence of wide adoption of a digital format includes bundling of tools with personal computers, native support in web browsers or market-leading content creation tools, and the existence of many competing products for creation, manipulation, or rendering of content in the format. Declared support of a format by other archival institutions is also relevant.

Examples:
- TIFF uncompressed, widely recommended as master for color or grayscale bitmapped images
- JP2 (JPEG2000 Part 1), increasingly adopted, including in medical and geospatial fields
- JPEG2000 (other parts), in early stages of adoption. JPM (JPEG2000 Part 6) looks promising for bitonal images of text.

3. *Transparency* refers to the degree to which the digital representation is open to direct analysis with basic tools, including 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.

Transparency is enhanced if textual content (including metadata embedded in files for non-text content) employs standard character encodings (e.g., UNICODE in the UTF-8 encoding) and stored in natural reading order. For preserving software programs, source code is much more transparent than compiled code. For non-textual information, standard or basic representations are more transparent than those optimized for more efficient processing, storage, or bandwidth. Examples of direct forms of encoding include, for raster images, an uncompressed bit-map and, for sound, pulse code modulation with linear quantization.

Encryption is incompatible with transparency; compression inhibits transparency. However, for practical reasons, some digital audio, images, and video may never be stored in an uncompressed form, even when created, and archival repositories will certainly accept content compressed using publicly disclosed and widely adopted algorithms.

Examples:

- TIFF uncompressed, straightforward encoding, reverse engineering can be envisaged even if specifications lost.
- JPEG2000, part 1, compression encoding is complex but other factors, e.g., adoption, may reduce likelihood of society losing understanding of the compression algorithm and outweigh this seeming shortcoming

4. *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 relating to creation and the early stages of the life cycle will be easier to manage over the long term than data objects that are stored separately from the metadata needed to render or understand them.

The value of richer capabilities for embedding metadata in digital formats has been recognized in the communities that create and exchange digital content. Such capabilities are built in to newer formats and standards (e.g., JPEG2000, and the Extended Metadata Platform for PDF [XMP]), and are reflected in emerging metadata standards and practices for exchange of digital content in industries such as publishing, news, and entertainment. This development is illustrated by the progression from the original JPEG standard, which contained very scant metadata, to the EXIF JPEG used in some digital cameras, which combines JPEG compression with richer metadata, and now to the JPEG2000 standard. Part 2 of JPEG2000 allows for any metadata to be embedded in metadata 'boxes' and specifically incorporates the extensive DIG35 metadata schema.

For operational efficiency of a repository system used to manage and sustain digital content, some of the metadata elements are likely be extracted into a separate metadata store or into catalogs or other systems designed to help users find relevant resources.

Many of the metadata elements required to sustain digital objects are not typically recorded in library catalogs or records intended to support discovery. The OAIS Reference Model recognizes the need for supporting information (metadata) in several categories: representation (to allow the data to be rendered and used as information); reference (to identify and describe the content); context (for example, to document the purpose for the content's creation); fixity (to permit checks on the integrity of the content data); and provenance (to document the chain of custody and any changes since the content was originally created).

5. *External dependencies* refers to the degree to which a particular format depends on particular hardware, operating system, or software for rendering or use and the predicted complexity of dealing with those dependencies in future technical environments. Some forms of interactive digital content, although not tied to particular physical media, are designed for use with specific hardware, such as a joystick. Scientific datasets built from sensor data may be useless without specialized software for analysis and visualization, software that may itself be very difficult to sustain, even with source code available.

Examples:

- Adobe eBooks require a Microsoft Passport or Adobe ID account to allow copying
- Open eBook format is free of external dependencies

6. *Impact of patents*. Refers to the 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 formats are often low or nil, the existence of patents may slow the development of open source encoders and decoders and prices for commercial software for transcoding content in obsolescent formats may incorporate high license fees. When license terms include royalties based on use (e.g., a royalty fee when a file is encoded or each time it is used), costs could be high and unpredictable. It is not the existence of patents that is a potential problem, but the terms that patent-holders might choose to apply.

The core components of emerging ISO formats such as JPEG2000 and MPEG-4 are associated with "pools" that offer licensing on behalf of a number of patent-holders. The license pools simplify licensing and reduce the likelihood that one patent associated with a format will be exploited more aggressively than others. The progression in the MPEG realm is interesting. MPEG-1 required no licenses. The MPEG-2 license pool requires toolmakers to license the technology (and pass through the associated cost) for each copy they sell of a product that can make MPEG-2 files. MPEG-4 goes a step further; pay-per-view fees (or their equivalent) are required each time a user plays an MPEG-4 and this requirement has put a brake on the adoption of MPEG-4.

7. *Technical protection mechanisms*. This refers to the implementation of mechanisms such as encryption that prevent the preservation of content by a trusted repository. To preserve digital content and provide service to future users, custodians must be able to replicate the content on new media, migrate and normalize it in the face of changing technology, and disseminate it to users at a resolution consistent with network bandwidth constraints. Long-term retention will be difficult if not impossible for content protected by technical mechanisms that prevent custodians from taking appropriate steps to preserve it.

No digital format inextricably bound to a particular physical carrier is suitable for long-term preservation; nor is an implementation of a digital format that constrains use to a particular device or prevents the establishment of backup procedures and disaster recovery operations.

Some digital content formats have embedded capabilities to restrict use in order to protect the intellectual property. Use may be limited, for example, for a time period, to a particular computer or other hardware device, or require a password or active network connection. Since the exploitation of these technical

protection mechanisms within a format is typically optional, this factor applies to the way a format is used in business contexts rather than to the format itself.

Examples:
- Sound recordings from Audible.com will only play with software and/or devices from Audible.
- MP3 files play anywhere.

Quality and functionality factors

Quality and functionality factors pertain to the ability of a format to represent the significant characteristics required or expected by current and future users of a given content item. These factors will vary for particular genres or forms of expression. For example, significant characteristics of sound are different from those for still pictures, whether digital or not, and not all digital formats for images are appropriate for all genres of still pictures.

To date, our analysis of functionality and quality factors focuses on four familiar content categories: still images, sound, textual materials, and video. Ahead lie categories whose future use is less analogous to Library of Congress experience, including Web sites and datasets. The latter will likely have to be treated in subcategories, such as geospatial data, social science surveys, etc.

As we looked at these factors, we found it useful to develop the concept of *normal rendering*, a baseline for the behavior of content when presented to a user, e.g., images that permit zooming or sounds that can be played, stopped, and restarted. Certain formats offer *functionality beyond normal rendering*, and these may be needed to serve the needs of users with special interests in certain content types. For example, some users will prefer that vector-based images like those used for architectural drawings remain malleable (editable) so that the full functionality, e.g, to view only selected types of elements or to change scale for drawing elements independently of labels, can be retained. This contrasts with freezing the drawings as bit maps, which is also possible.

The following outline lists the quality and functionality factors we use for still image formats.
- *Normal rendering* for still images includes on-screen viewing and printing to paper; and the ability to zoom in to study detail and the ability to produce publication quality output
- *Clarity* (support for high still image resolution) - the degree to which "high resolution" content may be represented within this format. Quality tends to correlate to pixel counts and bit depth. Vector formats offer “clean edges” and “geometric precision.” Implementations that eschew or minimize compression loss will be preferred.
- *Color maintenance* (support for color management) relates to the degree to which the color gamut represented in a given image can be managed, with an eye on inputs and outputs. Formats that allow ICC profiles to be embedded will be preferred.
- *Support for graphic effects and typography* is usually associated with vector graphics formats or formats that support bit-mapped and vector layers. Desirable features are support for the use of shadows, filters or other effects as applied to fill areas and text, levels of transparency, and use of fonts and patterns.
- *Functionality beyond normal image rendering* would include support for 3-D models, layers, or special treatment for regions of interest.

Balancing the factors

In practice, preferences among digital formats will be based on finding a balance among all the factors, for sustainability, quality, and functionality. Sometimes the factors compete. For example, some formats adopted widely for delivery of content to end users are proprietary or apply lossy compression for transmission over low-bandwidth networks. Disclosure can substitute for transparency. For content of high cultural value and for which a special functionality has particular significance, the ability of a format to support that functionality may outweigh the sustainability factors.

Curator’s view

Discussions with curators and other decision-makers are often facilitated by reducing complexities to a tabular comparison. For example, the rough and ready table below illustrates how one might use the factors to score some formats for bitmapped images. The first seven rows are the sustainability factors; the latter pair are quality and functionality factors. The table compares five formats or format subtypes. Most rows use a three-point scale: plus (+), period (.), and minus (-), with the plus sign indicating the most favorable score.

Table 2: Scorecard for bitmapped image formats

	TIFF (unc.)	EXIF-TIFF	JPEG	JP2	MrSID
Disclosure	+	+	+	+	.
Adoption	+	+	+	.	.
Transparency	+	+	-	-	-
Self-documentation	-	+	-	+	-
External dependencies	n/a	n/a	n/a	n/a	n/a
Patents	+	+	.	+	-
Tech. protection (possibility)	N	N	N	N	?
Clarity	+	+	-	+	+
Color maintenance	.	.	-	+	.

Some still image items acquired by the Library of Congress will warrant higher functionality and quality than others. For example the original artwork of a cartoonist, a digital snapshot submitted as part of an oral history project seeking community submissions, and a documentary nature photograph may warrant different balances of the factors. We have attempted to categorize some types of still images likely to be added to the

collections and for which the significant characteristics that must be preserved are potentially different.

Table 3: Categories of still image (bitmapped)

I1	Pictorial expression of high value. Examples: Works by graphic artists, photographers, advertisers for whom the designated community has high interest in the artist's intent.
I2	Images for which the artist's pictorial intent is less significant but color or tonality is significant. Examples: documentary photographs of nature, fashion, architecture; newspaper "file" photos; Landsat images
I3	Images for which spatial resolution is important, but color depth and precise color accuracy are not important. Examples: maps, graphs, technical drawings, Vector graphics "frozen" as bit-maps
I4	Pictorial expression of lower artistic value, such as: routine output of a portrait studio; images with significance as the expression of everyday life ("snapshots"); interesting-but-not-artistically valuable images associated with oral histories.
I5	Images incidental to Web harvesting, including animations consisting of only a few frames

For each of the categories we have proposed, although not fully vetted with colleagues, we are developing a short list of preferred and acceptable formats. For the top two categories, for example, our current, admittedly conservative, preference is for TIFF with no compression, although lossless JPEG2000 is acceptable, especially if color management data is included. If the image was created in a digital camera, we would prefer TIFF/EP; for graphic art, for, say, a magazine, TIFF/IT or PDF/X would be preferred. For the third category, the general preferences are similar, but color management is less needful. For the fourth category, the stakes are lower, and lossy compressed formats are certainly acceptable. For the fifth category, the Library will take what is available.

We do not necessarily expect the preferences to remain static. For example, we foresee that we will cling less firmly to uncompressed TIFF as a preferred image format as we overcome our reticence about JPEG2000. We are aware of its many advantages in terms of functionality and support for metadata and color management. As adoption of JPEG2000 grows, the balance is shifting.

Conclusion

This activity is in its infancy and we are eager for it to grow. During 2005, we will describe many more formats and hope to add new categories. We are very much aware that we are generalists about formats and welcome review and commentary by specialists. Our Website offers an online form for comments. Meanwhile, we hope to maximize our synergy with the Global Digital Format Registry and JHOVE, seeing our role as offering information to custodians of digital content and their role as tools that assist those custodians in their work.

References

1. Stephen L. Abrams and David Seaman. Towards a Global Digital Format Registry. WLIC:69th IFLA General Conference and Council, 128-E (2003) http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf
2. Global Digital Format Registry (GDFR). http://hul.harvard.edu/gdfr/
3. JHOVE. JSTOR/Harvard Object Validation Environment. http://hul.harvard.edu/jhove/
4. PRONOM File Format Registry. http://www.nationalarchives.gov.uk/pronom/
5. Reference Model for an Open Archival Information System (OAIS) http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html

Biographies

Carl Fleischhauer holds a BA degree from Kenyon College and an MFA from Ohio University. His work experience includes film and video production at West Virginia University (1969-1976); folklife field research, publications, and exhibitions at the American Folklife Center at the Library of Congress (1976-1990); coordination of the Library's American Memory program for online access to historical collections (1990-1998); and continuing service to collection-digitizing and digital preservation efforts at the Library of Congress in the National Digital Library Program and the Office of Strategic Initiatives (1998-present); the latter activity has included extensive participation in planning for development of the National Audio-Visual Conservation Center to be located in Culpeper, Virginia. Fleischhauer's publications include long playing records and audio compact discs of folk-music field recordings, a laser videodisc about a cattle ranch in Nevada, and books on the FSA-OWI photographic project and bluegrass music.

Caroline Arms has a BA degree from Oxford University (UK) and an MBA from Dartmouth College. At the Library of Congress, she has played a technical role in managing and providing access to digital content, including integrating twenty-seven collection from other institutions into American Memory and making descriptive records for Library of Congress collections harvestable by others. Earlier in her career, she ran the Microcomputer and Media Center at the Falk Library of the Health Sciences at the University of Pittsburgh and was the first Director of Computing at the Amos Tuck School of Business Administration at Dartmouth College. In the late 1980s, she edited *Campus Networking Strategies* and *Campus Strategies for Libraries and Electronic Information.*