Digital Formats for Library of Congress Collections

Factors To Consider When Choosing Digital Formats

DRAFT
Caroline R. Arms
Carl Fleischhauer
Office of Strategic Initiatives

Table of Contents

Introduction

Activity Goals and Objectives

This analysis of digital content formats is intended to contribute to planning at the Library of Congress regarding the expanded acquisition and preservation of digital content. As described in the following section, this November 2003 draft is limited in its coverage. The overall analysis is part of the execution of the Library of Congress Digital Strategic Plan, Goal 2 of which is *Manage and Sustain Digital Content*, which in turn includes *Technical Format Sustainability* as a priority element.

The goals for this analysis are:

- To develop a planning framework and strategies regarding digital content formats, in order to ensure the long-term preservation of digital content by the Library of Congress, and
- To build an inventory of information about current and emerging formats, including the identification of tools and detailed documentation that are needed to ensure that the Library of Congress can manage content created or received in these formats through the content life cycle, and
- To identify and describe the formats that are promising for long-term sustainability, and develop strategies for sustaining these formats including recommendations pertaining to the tools and documentation needed for their management.
- To identify and describe the formats that are not promising for long-term sustainability, and develop strategies for sustaining the content they contain.

The Library has an immediate need for a framework for making decisions. The challenge it faces is shared with other archival institutions, particularly in relation to information about digital formats. For example, the particular needs of the Public Record Office of the UK National Archives are reflected in their PRONOM application. The inventory of information about formats will have some overlap with a key component of PRONOM and also with the proposed Global Registry of Digital Format Representation Information. A proposal and technical design for such a registry is being developed through a planning activity sponsored by the Digital Library Federation, with MIT and Harvard universities playing leading roles. The Library of Congress is participating in that activity and expects its internal analysis and developments to benefit from and contribute to the more global view.

One key objective for the current analysis is the creation of an online resource for Library of Congress staff, with some elements publicly available, that pertains to the acquisition, management, and preservation of digital content. The online resource will help Library staff answer questions like the following:

- If a digital work is subject to mandatory deposit under U.S. Copyright Law, which of the formats in which it is available is preferred by the Library?
- When seeking to acquire a body of digital content with the intention of sustaining it for the long term, which formats are preferred or acceptable and why?

- Which digital formats must be fully supported by systems, automated tools, or workflow associated with the digital content life cycle processes under discussion at the Library, i.e., support for receiving and validating digital content (in the *Get* process), selecting digital content (in the *Select* process), preparing digital content for responsible long-term custody (in the *Prepare/Assemble* process), and establishing strategies for preservation (in the *Sustain* process)?
- Given content in a particular format, does the Library already have a commitment to support content in this digital format? If so, are their more specific technical requirements that apply? What associated metadata of a technical nature is essential? Does LC have an existing workflow process appropriate for receiving and validating digital content in this format? Or are software tools for format validation and metadata extraction available for building a workflow process?
- If a particular digital format is not already categorized as preferred or acceptable for a particular category or subcategory of material, what information or assistance is available to develop a recommendation that a format should be supported or that a process be developed for reformatting to a supported format?

Related set of documents, work in progress

This document, *Factors to Consider When Choosing Digital Formats*, discusses the high-level conceptual factors that affect the sustainability of any digital format. It also introduces additional factors that relate to quality or functionality (beyond normal rendering) that might be desired for certain categories of content, e.g., sound, still images, and text. It concludes with a list of references and a sample template for use when drafting descriptions of specific digital formats.

The quality and functionality factors receive elaborated treatment in a set of *Factors and Formats* documents devoted to specific categories of digital content. These documents also discuss the expectations for normal rendering of each of these categories and offer suggestions for how format preferences might be expressed for different contexts.
A further set of documents—to be available database-style in the online resource—will consist of a set of *Format Description Documents* for digital formats, using the template mentioned above. The format descriptions, which will require regular additions and updates, are intended to be compatible with the proposed global registry of digital formats supported by the Digital Library Federation.

The November 2003 versions of these documents are a work in progress, incorporating feedback from readers of a June 2003 version circulated for comment to a few knowledgeable readers in and outside the Library of Congress. In order to test a general conceptual approach, the writers of this document have limited their effort to born-digital still images, sound recordings, textual items, and video programs. There are three reasons for this:

- The analog equivalents to these types of items are familiar to the Library of Congress.
- Digital still images, sound recordings, textual items, and video programs have been regularly produced in the reformatting of historical collections carried

out by American Memory, and thus several relevant digital formats are familiar to the Library.

- These types of content (and perhaps a few others) are often the building blocks for more complex digital content, e.g., serials, compilations, multimedia works, etc., and the writers of this document wanted to start their analysis with the simpler forms and then move to the more complex.

Beyond the categories of still images, sound recordings, textual items, and video programs lie the complex content types mentioned above, as well as an array of content categories and digital formats that do not have close counterparts in the analog realm, e.g., interactive programs and executable objects.

The writers of this document propose that one of their next actions will be developing a process to identify additional content categories and format types of interest to the Library and to assign them priorities for analysis.

Some formats are being omitted from this analysis, at least in its early manifestations, since content that uses them is less likely to find their way into Library of Congress collections soon, e.g., the family of specialized formats that represent numeric data. Other omitted examples—within the broad categories of sound and video—are the heavily compressed formats especially designed to serve wireless communication, mobile telephony, teleconferencing, and the like. Although content from these communication modes is rarely if ever added to the Library's collections, some heavily compressed formats are employed for digital recorded books and may be used in emerging forms of moving image presentations, which will find their way to the Library. The example of Audible, Inc., (http://www.audible.com), however, indicates that publishers in these fields are likely to offer multiple formats for a given title, including ones that are moderately rather than heavily compressed, and these higher quality versions would generally be preferred for the Library's collections.

Scope of the Overall Activity

This analysis of digital format sustainability will focus on digital content formats that are independent of the physical medium on which they are stored or transported. Content in such formats has been dubbed "media-independent," "intangible," or "remote" (a cataloger's term), and it exists as data files or data streams. Media-independent digital content is stored and can be transported on media, e.g., CD-R, portable hard disk, and data tape, but this use of media is incidental to the content. In contrast, media-dependent formats are inextricably linked to their physical forms, e.g., audio CDs, DVDs, and digital videotape formats like DigiBeta. The development of preferences for these media-dependent or tangible formats by the Library inevitable raises issues of workflow, management of physical inventory, and media life. Since these issues, while important, are outside the scope of this analysis, media-dependent formats are excluded from this consideration of format preference.[1]

[1] To illustrate the complexity associated with curatorial decision-making regarding based digital content, consider the case of a body of recorded music that can be acquired as audio compact disks or as a set of MP3 files. At the Library of Congress, tens of thousands of newly published sound recordings are added to

In this analysis, the term *format* is used very broadly, to cover information packages can be stored as data files or transmitted as data streams. When considering sustainability, it is not sufficient to look at file formats at the level indicated by file extensions in the Windows operating system (e.g., .mp3) or an Internet MediaType (e.g. text/html). Some file formats are members of a class of related formats, whose familial characteristics are important. (Example: WAVE, a format for audio, is an instance of the RIFF format class.) Some file formats are "wrappers" and their implementations must be distinguished in terms of their underlying bitstream structures. (Example: WAVE files may contain pulse code modulated [PCM] audio.) It is also worth noting that file formats often exist in nuanced variations, either versions that develop through time or versions that are tailored to narrow, specific purposes. (Example: The Aldus Corporation TIFF format version 5.0 was supplanted by Adobe's version 6.0. More recently, the ISO standards body has established TIFF/IT for publishers' preprint requirements and TIFF/EP for electronic photography.) Some file formats have optional features (e.g. for encryption) that will inhibit sustainability if used. For examples, the programs downloaded from Audible, Inc., (http://www.audible.com/), are copyright-secure audio files in a format that "prevents a customer from passing along duplicate digital audio files to another listener." (http://audible.custhelp.com/) Thus the Library's format descriptions must document relationships among formats and how they are used in practice, as well as acknowledging the fact of versioning. A team will be needed to remain abreast of new developments and update the format descriptions over time.

Also included in this broad interpretation of *format* are specifications that represent bundles of files or bitstreams comprising a single digital work (e.g., text and supporting illustrations, or a movie with sound tracks in different languages). These bundling formats usually list the components 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. They often incorporate technical details about each component, since a single object may include a mix of texts, sound, images, etc.

Some emerging standards that play such a bundling role are intentionally generic; these include METS (Metadata Encoding and Transmission Standard) and MPEG-21, designed, in the words of the MPEG standard's working group, to support content producers and consumers as they "exchange, consume, trade, and otherwise manipulate

---

the collections each year. Compact disks offer higher music quality (no compression) but will require the management of a physical inventory, generation of metadata by copy cataloging from third party sources and, in time, reformatting of the recordings into the form of files. The MP3 files will be of lower quality but arrive "ready for the server" and are likely to contain metadata that can be extracted in an automated process. When should higher quality at higher overall cost be selected, and when is it preferable to follow a path that yields reduced quality but inexpensive management?

This challenge in decision-making will intensify with the proliferation of new media-dependent formats like Super Audio CD (SACD) and DVD-Audio, which offer surround sound and multimedia (DVD-Audio) but which will include technological protection that make preservation reformatting more difficult. Similar issues arise in the case of moving-image DVDs, many of which include added value expressed in interactive programming. DVDs are in wide distribution and generally feature technological protection.

Digital Items in an efficient, transparent, and interoperable way." Other bundling formats have been proposed for more specific purposes or communities. For example, Material Exchange Format (MXF) and Advanced Authoring Format (AAF) have been developed for improved workflow for content authoring and post-production in the motion picture and broadcast industries, while the NARAS Producers and Engineers Wing *Delivery Recommendations* is proposed by the as a standard for recording studios to submit content to record companies, e.g., master recordings destined for CD publication. Bundling formats may be designed to encapsulate the component data streams, to take the form of a separate file that accompanies the set of component files (or allow either option). Two emerging bundling formats exist for fixed media like CD and DVD[2], and if they succeed in the marketplace they may evolve into structures that can be used for media-independent content. Bundling formats are likely to be very important for the Library's preferences for receipt of digital content.

The preceding paragraphs supply some of the reasons why format preferences will often have to specify more than just the name of a file format. Preference statements may also benefit from specific recommendations relating to quality or call for descriptive or technical metadata. In the case of digital photography, for example, the Library may prefer uncompressed or losslessly compressed images to ones that have been compressed, resulting in a reduction of clarity. Files of textual material that do not support searching of the text would be discouraged. Content with embedded or accompanying metadata that follows standards or published guidelines are likely to be less costly for the Library to prepare for sustaining in a digital repository and integrate into systems for providing user access.

This format analysis activity—when at a higher level of completion than today—will identify two or three dozen preferred and acceptable digital formats for the Library of Congress. But many more exist: a recent description of web harvesting in Sweden and Finland reported 440 distinct formats held by those digital archives, although the report found that 44 formats in eight categories were the "most common."[3] Formats are constantly being created and/or evolving, and the Library must be prepared for constant updating of its format preferences, as it must also be prepared to provide technical support for the preferred and acceptable formats. To the degree possible, the Library must also be prepared to transform ("normalize") digital content that it wishes to acquire when this content is offered exclusively in formats other than the preferred or acceptable. It is clear that, even with automated tools, the acquisition of new works—what is called

---

[2] These media-dependent bundling specifications are partly aimed at home users who may create their own "libraries" of images, sound files, and the like. They may also be used in shaping content that is commercially distributed. The structures facilitate the identification and navigation of content, e.g., when a multi-song disk is made on a desktop PC and then played in an automobile or mobile device. The Optical Storage Industry Association has developed specifications for the XML-based MPV(the abbreviation is derived from "MultiPhoto/Video" and "MusicPhotVideo"; http://www.osta.org/mpv/), while Microsoft and Matsushita Electric Industrial Co., Ltd. (the parent of Panasonic) have joined forces to develop HighM.A.T. (High-Performance Media Access Technology; http://www.highmat.com/).

[3] *DAVID: Archiving Websites*, (page 37). Report available from the publications menu at http://www.antwerpen.be/david.

the *Get* process in the Library's digital content life cycle—will be a step requiring intense activity.

Although this discussion of formats is technical, collection policy matters are implied at many turns in the narrative. For example, the following section notes that the Library of Congress frequently collects finished works and occasionally collects works that document the creative process itself, a distinction that suggests the values in play when acquiring digital works. The *Factors and Formats* documents devoted to specific types of content categorize works in ways that will require selection decisions rooted in collections policies, e.g., is *this* digital image one for which color values are so significant that its acquisition format ought to support color management?

Framework for LC Decision-making

As the Library of Congress refines its processes for acquiring media-independent born-digital content, a conceptual technical framework is needed to support decision-making. *Acquisition* in this context is meant broadly, to cover the several ways in which the Library obtains content for its collections, e.g., through the workings of the copyright law; via purchases, exchanges, or licensing; and by donation. Included in this consideration are special projects like the Veterans History activity, which bring to the Library documentary materials produced by organizations across the nation, and the Minerva project, which harvests sites from the World Wide Web.

In the analog realm, much of what the Library collects is published, the final manifestation of a creative process. The acquisition of works in this *final state* will continue in the digital realm. The institution's special collections divisions, however, also collect works in other states. First are exemplars of the creative process, e.g., manuscripts and other draft documents or musical scores, i.e., work in its *initial state*. Although collected only in rare instances, this category may also include raw materials used in the creative process, e.g., the outtakes or leftover footage in a video production or the recorded music tracks that include a musician's mistakes, later expunged from the published manifestation. The Library may also collect works in what might be called a *middle state*, the form that content takes in the hands of a publisher. In some cases, the middle-state form is what is delivered to the publisher, as exemplified by the PDF/X or TIFF/IT files that a designer may employ when submitting digital art, or the proposed *Delivery Recommendations* from the NARAS Producers and Engineers Wing, for multi-track sound recordings fresh from the studio, with associated metadata used to produce the final mix. Middle-state formats are likely to be used by publishers for their own archiving.

For certain categories or subcategories of content, multiple versions in different formats will be desired by the Library in order to manage items through the content life cycle, and this may create a certain tension when preferred formats are identified. For example, the inspection of arriving content, even in a Copyright Office examination activity, requires ease of access and viewing. Similarly, easily accessible formats make possible the provision of digital content to readers in the Library's reading rooms. At the same time, other formats—often richer and with larger files—provide the best option for long-term

preservation. For example, compressed versions of images or sound recordings may be the most facile for access, while their uncompressed counterparts are the most sustainable. In some instances, in order to meet the need for multiple versions, the Library may have to produce more easily accessible versions itself.

Another tension concerning identifying preferred formats may arise in the context of textual items. Here, the Library may be offered formats like PDF that fit the document creators' wish to control the details of layout, font, or other matters of appearance. This desire must be weighed against the long term needs of those researcher-users whose needs will be best answered by formats that express the structure of the document, e.g., chapter headings and section breaks, in ways that makes this structure available for future automated analysis. A document with this type of structure—for example, using XML to identify the structural elements—can be processed to support future discovery, links from references to the associated documents and, more important, research studies carried out by, say, a social scientist looking for paragraphs or chapters that bear on certain topic.

Factors to Consider when Evaluating a Digital Format

Overview of Factors

In considering the suitability of particular digital formats for the purposes of preserving digital information as an authentic resource for future generations, it is useful to articulate important factors that affect choices. The *sustainability factors* listed below apply across digital formats for all categories of information. These factors influence the likely feasibility and cost of preserving the information content in the face of future change in the technological environment in which users and archiving institutions operate. They are significant whatever strategy is adopted as the basis for future preservation actions: migration to new formats, emulation of current software on future computers, or a hybrid approach. Some important considerations, e.g., matters pertaining to the authenticity of a digital item, are attributes of the systems used to manage digital content and not of the content format itself.

For particular genres or forms of expression for content, there will be additional factors relating to the ability to represent significant characteristics of the content, *factors reflecting quality and functionality* that will be expected by future users. For example, significant characteristics of sound are different from those of still pictures, whether digital or not, and not all digital formats for images are appropriate for all genres of still pictures.

Sustainability Factors

Disclosure

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 digital format over the long term is not feasible without an understanding of how the information is represented (encoded) as bits and bytes in digital files.

A spectrum of disclosure levels can be observed for digital formats. 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 this sustainability factor is not approval by a recognized standards body, but the existence of complete documentation, preferably subject to external expert evaluation. The existence of tools from various sources is valuable in its own right and as evidence that specifications are adequate. The existence and exploitation of underlying patents is not necessarily inconsistent with full disclosure but may inhibit the adoption of a format, as indicated below. In the future, deposit of full documentation in escrow with a trusted archive would provide some degree of disclosure to support the preservation of information in proprietary formats for which documentation is not publicly available. Availability, or deposit in escrow, of source code for associated rendering software, validation tools, and software development kits also contribute to disclosure.

Adoption

Adoption refers to the degree to which the format is already used by the primary creators, disseminators, or users of information resources. This includes use as a master format, for delivery to end users, and as a means of interchange between systems. If a format is widely adopted, it 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, including those intended for professional use, and the existence of many competing products for creation, manipulation, or rendering of digital objects in the format. In some cases, the existence and exploitation of underlying patents may inhibit adoption, particularly if license terms include royalties based on content usage. A format that has been reviewed by other archival institutions and accepted as a preferred or supported archival format also provides evidence of adoption.

Transparency

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 and more susceptible to digital archaeology; development of rendering software for new technical environments or conversion software based on the "universal virtual computer" concept proposed by Raymond Lorie will be simpler.[4]

---

[4] For examples of Lorie's treatment of this subject, see his "Long Term Preservation of Digital Information" in E. Fox and C. Borgman, editors, *Proceedings of the First ACM/IEEE Joint Conference on Digital Libraries* (JCDL'01), pages 346-352, Roanoke, VA, June 24-28 2001, http://doi.acm.org/10.1145/379437.379726; and The UVC: a Method for Preserving Digital Documents: Proof of Concept (no date), http://www.kb.nl/kb/ict/dea/ltp/reports/4-uvc.pdf.

Transparency is enhanced if textual content (including metadata embedded in files for non-text content) is encoded in 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. For numeric data, standard representations exist for signed integers, decimal numbers, and binary floating point numbers of different precisions (e.g., IEEE 754-1985 and 854-1987, currently undergoing revision).

Many digital formats used for disseminating content employ encryption or compression. 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. Archival repositories must certainly accept content compressed using publicly disclosed and widely adopted algorithms that are either lossless or have a degree of lossy compression that is acceptable to the creator, publisher, or primary user as a master version.

The transparency factor relates to formats used for archival storage of content. Use of lossless compression or encryption for the express purpose of efficient and secure transmission of content objects to or from a repository is expected to be routine.

Self-documentation

Digital objects that are *self-documenting* are likely to be easier to sustain over the long term and less vulnerable to catastrophe than data objects that are stored separately from all the metadata needed to render the data as usable information or understand its context. A digital object that contains basic descriptive metadata (the analog to the title page of a book) and incorporates technical and administrative metadata relating to its creation and early stages of its life cycle will be easier to manage and monitor for integrity and usability and to transfer reliably from one archival system to its successor system. Such metadata will also allow scholars of the future to understand how what they observe relates to the object as seen and used in its original technical environment. The ability of a digital format to hold (in a transparent form) metadata beyond that needed for basic rendering of the content in today's technical environment is an advantage for purposes of preservation.

The value of richer capabilities for embedding metadata in digital formats has been recognized in the communities that create and exchange digital content. This is reflected in capabilities built in to newer formats and standards (e.g., TIFF/EP, JPEG2000, and the Extended Metadata Platform for PDF [XMP]) and also in the emergence of metadata standards and practices to support exchange of digital content in industries such as publishing, news, and entertainment. Archival institutions should take advantage of, and encourage, these developments. The Library of Congress will benefit if the digital object files it receives include metadata that identifies and describes the content, documents the creation of the digital object, and provides technical details to support rendering in future

technical environments. 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. Some elements will also be extracted for use in the Library's catalog and other systems designed to help users find relevant resources.

Many of the metadata elements that will be required to sustain digital objects in the face of technological change are not typically recorded in library catalogs or records intended to support discovery. The OAIS Reference Model for an Open Archival Information System 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). Digital formats in which such metadata can be embedded in a transparent form without affecting the content are likely to be superior for preservation purposes. Such formats will also allow metadata significant to preservation to be recorded at the most appropriate point, usually as early as possible in the content object's life cycle. For example, identifying that a digital photograph has been converted from the RGB colorspace, output by most cameras, to CMYK, the colorspace used by most printing processes, is most appropriately recorded automatically by the software application used for the transformation. By encouraging use of digital formats that are designed to hold relevant metadata, it is more likely that this information will be available to the Library of Congress when needed.

External dependencies

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 microphone or 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.

This factor is primarily relevant for categories of digital content beyond those considered in more detail in this document, for which static media-independent formats exist. It is however worth including here, since dynamic content is likely to become commonplace as part of electronic publications. The challenge of sustaining dynamic content with such dependencies is more difficult than sustaining static content, and will therefore be much more costly.

Technical protection mechanisms

To preserve digital content and provide service to users and designated communities decades hence, 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. Content for which a trusted repository takes long-term responsibility must not be protected by technical mechanisms

such as encryption, implemented in ways that prevent custodians from taking appropriate steps to preserve the digital content and make it accessible to future generations.

No digital format that is inextricably bound to a particular physical carrier is suitable as a format 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 expected of a trusted repository.

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. In most cases, exploitation of the technical protection mechanisms is optional. Hence this factor applies to the way a format is used in business contexts for particular bodies of content rather than to the format.

The embedding of information into a file that does not affect the use or quality of rendering of the work will not interfere with preservation, e.g., data that identifies rights-holders or the particular issuance of a work. The latter type of data indicates that this copy of this work was produced for an specific individual or other entity, and can be used to trace the movement of this copy if it is passed to another entity.

Quality and Functionality Factors

Quality and functionality factors vary by genre and form of expression
Other factors that affect the choice of digital format reflect considerations of quality and functionality expected by current or future users. These factors will vary by genre or form of expression for content. For example, significant characteristics of, say, a musical recording will include whether it is in mono, stereo, or intended to convey the effect of "surround sound." For still pictures, the degree to which exact color fidelity is needed will vary. For maps and graphs, it may be sufficient to distinguish colors reliably, while for works by, say, an illustrator, the ability to reproduce the precise tones may be important. Other forms of digital content, such as numeric datasets, online newspapers, electronic books, etc., present different sets of significant characteristics. Choices about which characteristics are essential to represent and preserve will have to be made, and recorded as metadata, at different stages of the life cycle for digital content. One point at which such a choice may occur is when the Library expresses a preference for one digital format over another for a particular item or body of material. For example, the format that is preferred for images resulting from digital photography (the equivalent of photographic negatives) may not be the same as that for images that merge line art, manipulated photographs, and computer graphics and are intended for printing as a poster or brochure.

In addition to the choice of format, there will be detailed considerations (e.g., spatial resolution for images, sampling frequency for sound, choice of XML Schema to use for marking up text) that come into play when establishing the preferred technical specifications for content added to the Library's collections.

The *Factors and Formats* documents for various categories and subcategories of digital content—to the degree that they have been completed at this writing—describe the following quality and functionality factors:

Sound recordings
- Fidelity (support for high audio resolution)
- Sound field (support for multichannel audio)
- Functionality beyond normal sound rendering (notation-based representations, e.g., MIDI)

Still images
- Clarity (support for high image resolution)
- Color maintenance (support for color management)
- Functionality beyond normal image rendering (vector graphics, 3-D models, etc.)

Textual content
- Integrity of document structure and navigation
- Integrity of layout, font, and other design features
- Integrity of rendering for mathematics, chemical formulae, diagrams, etc.

Video content
- Clarity (support for high image resolution)
- Fidelity (support for high audio resolution)
- Sound field (support for multichannel audio)
- Functionality beyond normal video rendering (encoded animation, frame-accurate editing, additional sound tracks)

Selecting a Format

Balancing the factors

In practice, preferences among digital formats will be based on a balance among the factors listed above: disclosure, adoption, transparency, self-documentation, external dependencies, technical protection, quality, and functionality. Sometimes these 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 example, the developers of the JPEG2000 format based on wavelet compression are said to have tested the published specification by giving it to several programmers independently and asking them to program a compliant viewer based only on the specification. 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.

Also important to the selection of acceptable formats is the channel by which digital content may be received. For content that will be received through the Copyright Office,

it is important that the list of acceptable formats include formats that can be conveniently provided by those wishing to register material for copyright or from whom the Library of Congress will expect deposit. For this channel, adoption may be the key factor, leading to acceptance of content in formats that provide less quality or functionality than would be sought in direct negotiations with a source of digital content. For example, for visual materials registered for copyright as digital images, the formats supported by digital cameras aimed at both professional and consumer markets must be considered. Similarly, for recorded sound, the formats used for widespread online distribution through downloading must be acceptable.

Summary

If archival institutions like the Library of Congress are to make responsible decisions about collecting content in digital form, it is important that staff with expertise in particular content areas can be aware of the impact of decisions about the digital formats in which content in their area may be acquired or received. The sustainability factors presented here are proposed as a simplified way to look at a very complex technical problem.

The writers hope that the framework of high-level sustainability factors presented here can provide a basis for decision-making for a considerable period. Issues of quality and functionality for particular categories of content or modes of expression (introduced in accompanying *Factors and Formats* documents) may also prove valuable for the longer term, although new modes of expression, developed for the interactive, networked environment of today, will require consideration of very different aspects of functionality. Meanwhile, the list of formats that are preferred or acceptable today must be reviewed frequently, given the rapid development of new standards for digital content and pace of adoption to support the enhanced quality and functionality that tomorrow's creator will generate and tomorrow's user will expect .

Sources and Related Resources

Most of the resources listed here pertain to (1) the assessment of digital formats, (2) efforts to provide gateways and/or listings of format documentation, and (3) activities that entail preservation as it pertains to formatted digital content. Specific instances of format documentation proper will be provided in the *Format Description Documents*, for which a sample template is provided as Appendix A of this document.

Readers of this version of this document are encouraged to recommend additional items to the writers.

I. Resources with a central concern with digital format issues

Australia National Archives, *XML Data Formats: Requests for Comment*
National Archives of Australia document pertaining to wrappers for "a number of data formats for converting the digital records it receives from Commonwealth agencies."
URL: www.naa.gov.au/recordkeeping/preservation/digital/xml_data_formats.html

Diffuse Standards and Specifications List
From the European Commission sponsored *Diffuse* project, a source for standards documents and specifications, including data representation.
URL: www.diffuse.org/standards.html

DLF Global Registry for Digital Format Representation Information
An effort to build a registry that "will maintain persistent, unambiguous bindings between public identifiers for digital formats and representation information for those formats."
URL: hul.harvard.edu/formatregistry/

DSpace statement concerning format support
From the project at MIT, a listing of formats categorized as *supported*, *known*, and *unsupported*.
URL: http://dspace.org/mit/policies/format.html

Graphics File Formats, 2nd Edition
Reasonable overviews of many formats albeit lacking in the detail for certain types of digital archeology.
Citation: James D. Murray and William vanRyper (Sebastopol, CA: O'Reilly & Associates, 1994).

Graphics File Formats FAQ (Part 3 of 4): Where to Get File Format Specifications
Web Site created by James D. Murray (jdm@ora.com) that offers links to about 200 sites for graphics file formats. When consulted in April 2003, the last modified tag reported 20Jan97.

URL: isc.faqs.org/faqs/graphics/fileformats-faq/part3/preamble.html

Leeds, University of, *Survey and Assessment of Sources of Information on File Formats and Software Documentation*

Report from the Representation and Rendering Project at the University of Leeds (UK, n.d., ca. 2003). Describes the publicly available sources of information on file formats and software, with some comments on its quality and completeness.
URL: www.jisc.ac.uk/uploaded_documents/FileFormatsreport.pdf

NIST National Software Reference Library

From the National Institute of Standards and Technology, a project to collect of software and to incorporate file profiles computed from this software into a reference data set that can be used by law enforcement, government, and industry to identify files found on a computer.
URL: www.nsrl.nist.gov

PRONOM Digital Format Database

From the Public Records Office of the United Kingdom, a database system provides information about file formats and the application software needed to open them.
URL: http://www.pro.gov.uk/about/preservation/digital/pronom/default.htm

Tripwire File Signature Database (FSDB)

A planned "repository of file metadata derived from published software allowing customers to identify, authenticate and assure the integrity of files. It will provide the capability to enhance proactive management of change through granular file dependency structure."
URL for press release: http://www.tripwire.com/fsdb/press_release.pdf

II. Resources with a secondary concern with digital format issues

Building an Electronics Records Archive at the National Archives and Records Administration: Recommendations for Initial Development (Pre-publication Draft, 2003)

Report of study by a committee under the auspices of the National Research Council of the National Academies; see especially the section titled "Data Types and Obsolescence," pp. 5-3 to 5-5.
URL: http://books.nap.edu/books/0309089476/html/index.html

DAVID: Archiving Websites

From the city archives of Antwerp, Belgium, an overview of challenges related to collecting websites. Pages 37-42 discuss some issues associated with formats.
URL: available from the *reports* subsection of the *publications* menu at http://www.antwerpen.be/david/

Dutch National Archives, *From Digital Volatility to Digital Permanence: Preserving email*

From the Digital Preservation Testbed project of the Dutch National Archives, this 2003 document is part of a larger report in progress that will also cover text documents and spreadsheets. It discusses email in terms of authenticity (as an official record) and assess the various preservation strategies that may be applied.
URL: www.digitalduurzaamheid.nl/index.cfm?paginakeuze=185&categorie=2

Journal Archiving and Interchange Document Type Definition (DTD)
From the National Center for Biotechnology Information (NCBI) of the National Library of Medicine (NLM), created with the intent of providing a common format in which publishers and archives can exchange journal content.
URL: http://dtd.nlm.nih.gov/

KB/IBM. *Authenticity in a Digital Environment*
From IBM and the National Library of the Netherlands (KB), this report (December 2002) discusses a framework for defining what is meant by an authentic digital object. Includes an approach for analyzing content in terms of its makeup; for example, see pp. 16-18.
URL: http://www.kb.nl/kb/ict/dea/ltp/reports/2-authenticity.pdf

KB/IBM. *Preservation Requirements in a Deposit System*
From IBM and the National Library of the Netherlands (KB), this report (n.d., ca. 2002) presents the requirements for the preservation subsystem of the Digital Information Archiving System (DIAS) under development at the KB. Appendix C (p. 35) lists 35 recognized file types and subtypes to be implemented in the first release of DIAS.
URL: http://www.kb.nl/kb/ict/dea/ltp/reports/3-preservation.pdf

Appendix A. Template for Digital Format Description Documents

Information about formats and related elements

Library decision-making about preferred digital formats will require the provision of specific information about individual formats and their characteristics. In order to meet this need, the writers of this document have created a template for a series of *Digital Format Description Documents* that will provide moderately detailed information and citations for a variety of file formats, file-format classes, bitstream structures, and the encoding algorithms used to compress the file or bitstream. At this writing, we have proposed that an XML Schema be developed for the *Digital Format Description Documents*, to allow dissemination in various forms, including presentation as a searchable collection from a web site and in a form appropriate for printing.

These descriptions will not only serve as a resource for selection and acquisition at the Library but will also support "sustaining" digital content, providing an initial stage for compiling the Representation Network that is part of the OAIS information architecture. The *Format Description Documents* are intended to be human-readable, but the plan is that identification of formats and the expression of relationships between formats will be consistent with the proposed Global Registry for Digital Format Representation Information (http://hul.harvard.edu/formatregistry/). The creation and management of this set of reference documents by the Library should be associated with acquiring and maintaining copies of relevant format-specific documentation at the Library.

The template example that follows has been tailored for sound, i.e., the functionality and quality factors noted are those for that type of content. Descriptions of actual formats, bitstream structures and encoding algorithms, i.e., filled in templates, are presented in separate documents. The total universe of *Format Description Documents* may reach one or two hundred; only a handful have been prepared at this writing.

DRAFT template for format descriptions for sound

Short Name

Identification and Description

Full name

Description

Relationship to other formats

- Subtype of
- Has subtype
- May contain
- Used by

Local use
LC experience or existing holdings

Sustainability factors
Disclosure

- Standardization
- Other documentation

Adoption

- Licensing and patent claims

Transparency

Self-documentation

External dependencies

Technical protection considerations

Factors for Sound (quality and functionality)
Fidelity (support for high audio resolution)

Sound field (support for multi-channel audio)

Functionality beyond normal sound rendering

Useful References

URLs

Print