ISSN 1747-1524

DCC | Digital Curation Manual

Instalment on
"File Formats"

http://www.dcc.ac.uk/resource/curation-manual/chapters/file-formats

---

Stephen Abrams
Harvard University Library
http://hul.harvard.edu/

October 2007
Version 1.0

Legal Notices

Catalogue Entry

Title	DCC Digital Curation Manual Instalment on File Formats
Creator	Stephen Abrams (author)
Subject	Information Technology; Science; Technology--Philosophy; Computer Science; Digital Preservation; Digital Records; Science and the Humanities.
Description	The goal of digital curation is to ensure the appropriate usability of managed digital assets over time. Format is a fundamental characteristic of a digital asset that governs its ability to be used effectively. Without strong format typing a digital asset is merely an undifferentiated string of bits. The information content encoded into an asset's bits can only be interpreted properly and rendered in human-sensible form if that asset's format is known. While it is possible for bits to be preserved indefinitely without consideration of format, it is only through the careful management of format that the meaning of those bits remains accessible over time. This instalment investigates aspects of format description, validation, and characterisation that may assist with long-term curation and usability of data.
Publisher	HATII, University of Glasgow; University of Edinburgh; UKOLN, University of Bath; Council for the Central Laboratory of the Research Councils.
Contributor	Seamus Ross (editor)
Contributor	Michael Day (editor)
Date	29 October 2007 (creation)
Type	Text
Format	Adobe Portable Document Format v.1.3
Resource Identifier	ISSN 1747-1524
Language	English
Rights	© HATII, University of Glasgow

Citation Guidelines

Stephen Abrams, (October 2007), "File Formats", *DCC Digital Curation Manual*, S.Ross, M.Day (eds), Retrieved <date>, from http://www.dcc.ac.uk/resource/curation-manual/chapters/file-formats

About the DCC

The JISC-funded Digital Curation Centre (DCC) provides a focus on research into digital curation expertise and best practice for the storage, management and preservation of digital information to enable its use and re-use over time. The project represents a collaboration between the University of Edinburgh, the University of Glasgow through HATII, UKOLN at the University of Bath, and the Council of the Central Laboratory of the Research Councils (CCLRC). The DCC relies heavily on active participation and feedback from all stakeholder communities. For more information, please visit www.dcc.ac.uk. The DCC is not itself a data repository, nor does it attempt to impose policies and practices of one branch of scholarship upon another. Rather, based on insight from a vibrant research programme that addresses wider issues of data curation and long-term preservation, it will develop and offer programmes of outreach and practical services to assist those who face digital curation challenges. It also seeks to complement and contribute towards the efforts of related organisations, rather than duplicate services.

DCC - Digital Curation Manual

Editors

Seamus Ross
Director, HATII, University of Glasgow (UK)

Michael Day
Research Officer, UKOLN, University of Bath (UK)

Peer Review Board

Preface

The Digital Curation Centre (DCC) develops and shares expertise in digital curation and makes accessible best practices in the creation, management, and preservation of digital information to enable its use and re-use over time. Among its key objectives is the development and maintenance of a world-class digital curation manual. The *DCC Digital Curation Manual* is a community-driven resource—from the selection of topics for inclusion through to peer review. The Manual is accessible from the DCC web site (http://www.dcc.ac.uk/resource/curation-manual).

Each of the sections of the *DCC Digital Curation Manual* has been designed for use in conjunction with *DCC Briefing Papers*. The briefing papers offer a high-level introduction to a specific topic; they are intended for use by senior managers. The *DCC Digital Curation Manual* instalments provide detailed and practical information aimed at digital curation practitioners. They are designed to assist data creators, curators and re-users to better understand and address the challenges they face and to fulfil the roles they play in creating, managing, and preserving digital information over time. Each instalment will place the topic on which it is focused in the context of digital curation by providing an introduction to the subject, case studies, and guidelines for best practice(s). A full list of areas that the curation manual aims to cover can be found at the DCC web site (http://www.dcc.ac.uk/resource/curation-manual/chapters). To ensure that this manual reflects new developments, discoveries, and emerging practices authors will have a chance to update their contributions annually. Initially, we anticipate that the manual will be composed of forty instalments, but as new topics emerge and older topics require more detailed coverage more might be added to the work.

To ensure that the Manual is of the highest quality, the DCC has assembled a peer review panel including a wide range of international experts in the field of digital curation to review each of its instalments and to identify newer areas that should be covered. The current membership of the Peer Review Panel is provided at the beginning of this document.

The DCC actively seeks suggestions for new topics and suggestions or feedback on completed Curation Manual instalments. Both may be sent to the editors of the *DCC Digital Curation Manual* at curation.manual@dcc.ac.uk.

Seamus Ross & Michael Day.

18 April 2005

Biography of the author

Stephen Abrams is the digital library program manager at the Harvard University Library, where he provides technical leadership for strategic planning, design, and coordination for the Library's digital projects, systems, and assets. He was the architect of the JHOVE format identification, validation, and characterization tool; the ISO project leader and document editor for the PDF/A standard (ISO 19005-1); and is directing efforts towards establishing a Global Digital Format Registry (GDFR).

Table of Contents

Introduction and scope

The goal of digital curation is to ensure the appropriate usability of managed digital assets over time. Format is a fundamental characteristic of a digital asset that governs its ability to be used effectively. Without strong format typing a digital asset is merely an undifferentiated string of bits. The information content encoded into an asset's bits can only be interpreted properly and rendered in human-sensible form if that asset's format is known. While it is possible for bits to be preserved indefinitely without consideration of format, it is only through the careful management of format that the meaning of those bits remains accessible over time. However, as Waters and Garret noted in the 1996 report of the Task Force on Archiving of Digital Information, "Rapid changes in the. . .formats for storage. . .threaten to render the life of information in the digital age as, to borrow a phrase from Hobbes, 'nasty, brutish and short.'"[1] Similarly, the Library of Congress planning report, *Preserving Our Digital Heritage: Plan for the National Digital Information Infrastructure Preservation Program* (NDIIPP), stated that "Longevity of digital data and the ability to read those data in the future depend upon standards for encoding and describing, but standards change over time."[2] Clearly then, a deep understanding of format is of primary importance to curation activities.

Format can be used by curation managers as an important organizing principle for their digital collections. A number of phases of curation activity—selection, acquisition, delivery, preservation—potentially include considerations of format. Selection may involve a choice between various versions of content. One strong determinant could be an assessment of the formats most amenable to long-term usability. During acquisition it is important to verify that assets conform to desired formal specifications. On a technical plane this will involve format-based validation and characterization. A digital asset remains usable only so long as it can be delivered—that is to say, rendered—properly, and rendering processes are inherently

[1] Waters, D. and J. Garrett, eds., 1 May 1996, *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*, Commission on Preservation and Access, Research Libraries Group, http://www.rlg.org/en/page.php?Page_ID=20442 [Accessed: 2 January 2007].

[2] *Preserving Our Digital Heritage: Plan for the National Digital Information Infrastructure Preservation Program*, October 2002, Library of Congress, http://www.digitalpreservation.gov/repor/ndiipp_plan.pdf [Accessed: 14 March 2007].

specific to format. Thus, preservation planning and monitoring for incipient obsolescence will be structured around classes of assets grouped by format.

Ideally, formats should be defined expansively, capable of fully representing a wide range of content so as to reduce the number of formats necessary to represent collection content and to simplify curation workflows. The manner in which a format represents a piece of abstract content should be efficient, yet should not entail discarding any significant information content. Formats should be publicly documented in an authoritative and inclusive manner, and supported by a wide range of platform-independent tools for content creation, validation, modification, and rendering. This support should be robust and persistent over long time-spans. Instances of formatted assets should be self-contained and self-documenting to facilitate reverse engineering or digital archaeology should these recourses become necessary.

In terms of the ISO 14721 Open Archival Information System (OAIS) reference model, *information* is the fundamental unit of exchangeable knowledge.[3] Instances of information that share common characteristics can be described abstractly in terms of an information model. A format defines a transformation from an instantiation of an information model to a tangible byte stream. This transformation can be considered in three conceptually independent stages: a semantic encoding that maps portions of the information model to appropriate sets of information structures; a syntactic encoding that maps these structures to a set of data units; and a serialization encoding that maps data units to sequences of bytes. A format is therefore a class defined in terms of the rules that express these three encodings.

For example, consider an image format such as TIFF (Tagged Image File Format).[4] Its abstract information model includes any physical or synthetic visual field capable of being sampled discretely in terms of a regular rectangular grid. This multi-channel grid, known as a "raster", is the fundamental semantic property of the model. Each channel of each sample is syntactically represented by a scaled integer value. Finally, each integer is serialized in either big- or little-endian form.[5]

[3] ISO 14721, 2003, *Space data and information transfer systems – Open archival information system – Reference model*, pre-print available http://public.ccsds.org/publications/archive/650x0b1.pdf [Accessed: 26 December 2004].

[4] *TIFF Revision 6.0*, 3 June 1992, Adobe, http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf [Accessed: 14 March 2007].

[5] Big- and little-endian refer to the sequencing of the bytes that make up a multi-byte representation of a numeric value. For example, a 32 bit (4 byte) integer value can be represented with the most numerically significant byte first (big-endian, historically used by IBM PowerPC and Sun SPARC processors) or with the

(See, for example, Figure 1.) In practice, the formal specifications for most formats freely co-mingle the rules for semantic, syntactic, and serialization encodings. Similarly, most processes that operate on formatted byte streams do not do so with clear sequential demarcation between these three levels. Nevertheless, this tripartite conceptual framework is useful for consideration of the theoretical properties of formats.

Note that this expansive definition of format as an encoding allows it to be applied against a wide range of conceptual entities, some of which fall outside of what would nominally be thought of as file formats. For example, IEEE 754 standard for the binary representation of floating point numbers and, at the macro level, the Unix File System (UFS) can both be considered formats: they each define unambiguous rules for mapping from abstract information to sequences of bytes.[6]

The terminology "file format" has historically been used in discussions of this topic. However, it is more proper to use the term "format" alone to refer to the semantic, syntactic, and serialization encoding rules for byte streams, whether they are encapsulated in persistent physical files or exist in more ephemeral in-memory form or in network transfer contexts. While every file has a format, a particular subset of a file can also have its own unique format designation. For example, a TIFF file can contain an ICC (International Color Consortium) color profile and various types of metadata, such as IIM (Information Interchange Model) or XMP (Extensible Metadata Platform).[7] Thus, a single TIFF file can encompass four or more formats: TIFF itself (defining the overall organization of the entire "file"), along with ICC, IIM, and XMP associated with various segments of the file. Similarly, container formats such as TAR or ZIP are explicitly used to encapsulate many individual files or byte streams, each potentially with an independent format, in a single file.[8] On the other hand, the

least significant byte first (little-endian, used by Intel processors).

[6] IEEE 754, 1985, *Standard for Binary Floating-Point Arithmetic*; McKusick, M. K., W. M. Joy, S. J. Leffler, and R. S. Fabry, August 1984, "A Fast File System for UNIX", *Transactions on Computer Systems*, volume 2, number 3, pp. 181-197.

[7] ICC.1, *Image technology colour management — Architecture, profile format, and data structure*, October 2004, http://www.color.org/ICC1V42.pdf [Accessed: 15 April 2007]; *IPTC—NAA Information Interchange Model, Version 4*, 1 July 1999, International Press Telecommunications Council, http://www.iptc.org/std/IIM/4.1/specification/IIMV4.1.pdf [Accessed: 15 April 2007]; *XMP Specification*, September 2006, Adobe, http://partners.adobe.com/public/developer/en/xmp/sdk/XMPspecification.pdf [Accessed: 15 April 2007].

[8] TAR (tape archive) is codified as part of the POSIX (Portable Operating System Interface) standard, ISO/IEC 9945-1, 2003, *Information technology – Portable Operating System Interface (POSIX) – Part 1: Base Definitions*. ZIP is a file storage and transfer format. *ZIP File Format Specification*, 11 April 2007, PKWARE, Inc.,

Shapefile format often used in Geo-spatial Information System (GIS) applications defines an aggregate logical asset manifest in three separate physical files each with its own format.[9] The Shapefile format can be considered a "content model", a notion defined by the open source Fedora repository project as a class of digital asset instantiated by a set of files associated with a network of structural and functional relationships.[10] This is similar to the PREMIS notion of a representation: a "set of files, including structural metadata, needed for a complete and reasonable rendition of an Intellectual Entity."[11] (PREMIS—Preservation Metadata: Implementation Strategies—is a set of core metadata applicable to digital preservation repositories.) For example, document or page-oriented content is often represented digitally with individual TIFF page image files and OCR files associated with a parent METS file providing structural and intellectual metadata. (METS, the Metadata Encoding and Transfer Standard, is a widely used container format for digital library applications.[12]) Such a multi-part entity can be considered an instance of a single, aggregate-level format, as it can be defined in terms of unambiguous encoding rules.

Background and developments to date

To date, the most widely used formalism for managing formats is the MIME (Multipurpose Internet Mail Extensions) media type registry operated by the Internet Assigned Name Authority (IANA).[13] MIME types are widely used as the primary means of technical characterization of managed digital assets in digital library applications, preservation and access repositories, and curation workflows. For example, they are used in the Content-type header of the World Wide Web (WWW) Hypertext Transfer Protocol (HTTP) protocol to provide web user agents with the information necessary for

http://www.pkware.com/documents/casestudies/APPNOTE.TXT [Accessed: 24 August 2007].

[9] *ESRI Shapefile Technical Description*, July 1998, ESRI, http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf [Accessed: 15 April 2007].

[10] Payette, S., May 2006, "Formalizing Content Models", *Fedora Content Model Workshop*, Karlsruhe, http://www.fedora.info/presentations/cmodel-intro.ppt [Accessed: 4 March 2007].

[11] *Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group*, May 2005, OCLC, Research Libraries Group, http://www.oclc.org/research/projects/pmwg/premis-final.pdf [Accessed: 18 March 2006].

[12] *Metadata Encoding and Transmission Standard (METS) Official Web Site*, 23 August 2007, Library of Congress, http://www.loc.gov/standards/mets/ [Accessed: 23 August 2007].

[13] *MIME Media Types*, 7 December 2006, The Internet Corporation for Assigned Names and Numbers, http://www.iana.org/assignments/media-types/ [Accessed: 28 December 2006]. See also Freed, N. and J. Klensin, December 2005, *Media type Specifications and Registration Procedures*, RFC 4288, BCP 13, Internet Engineering Task Force, http://www.ietf.org/rfc/rfc4288.txt [Accessed: 2 January 2007].

proper rendering.[14] MIME type is also used as the primary means for explicit technical characterization in the PREMIS data dictionary.

MIME uses a two-level system of identifiers: a top-level media type and a subtype.[15] The media type declares the general type, or genre, of the designated data, which is further refined and particularized by the subtype. For example, the TIFF format would be identified as "image/tiff": the specific TIFF formulation of image data. The following media types are currently defined:

- Application
- Audio
- Example
- Image
- Message
- Model
- Multipart
- Text
- Video

The "application" media type is used for uninterpreted binary data or information intended primarily for processing by an application; the "example" is intended to be referenced in the context of textual examples only; and the "message" and "multipart" types are generally used only in the context of Internet mail.

Although MIME types are useful for the mail-centric purpose for which they were originally developed, in several respects they do not provide all the functions necessary for effective curation activities. The MIME registry is composed of textual documents intended for human comprehension, not machine actionability. Furthermore, the amount of information about the formats in the registry is fairly minimal and its completeness varies greatly. Perhaps most significantly, though, MIME types are defined at a fairly coarse granularity. For example, in many important curation contexts the variant "profiles" of TIFF such as TIFF/EP (ISO 12234-2), TIFF/IT (ISO 12639), GeoTIFF, and DNG (Digital Negative) can be considered to have quite different sets of significant properties, necessitating independent workflows, yet all are defined by the same MIME type, "image/tiff".[16] For

[14] Fielding, R., J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee, June 1999, *Hypertext Transfer Protocol – HTTP/1.1*, RFC 2616, Internet Engineering Task Force, http://www.ietf.org/rfc/rfc2616.txt [Accessed: 14 March 2007].

[15] Freed, M. and N. Borenstein, November 1996, *Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types*, RFC 2046, Internet Engineering Task Force, http://www.ietf.org/rfc/rfc2046.txt [Accessed: 2 January 2007].

[16] ISO 12234-2, *Electronic still-picture imaging – Removable memory – Part 2: TIFF/EP image data format*, 15 October 2001; ISO 12639, *Graphic technology – Prepress digital data exchange – Tag image file format for image technology (TIFF/IT)* , 4 September 2003; Ritter, N. and M. Ruth, 28 December 2000, *GeoTIFF Format Specification*, http://remotesensing.org/geotiff/spec/geotiffhome.html [Accessed: 14 March 2007]; *Digital Negative (DNG) Specification*, February 2005, Adobe, http://www.adobe.com/products/dng/pdfs/dng_spec.pdf [Accessed: 14 March 2007].

purposes of providing a general-purpose rendering environment it may be sufficient to associate a given file with the generic TIFF format. However, in order to migrate that same file to a successor format, it will be important to know that it uses the DNG format, which defines important extensions to baseline TIFF function. Without designing the migration workflow specifically for DNG, important information content may be irretrievably lost. The inability of the MIME scheme to capture the nuance of format granularity is addressed by more recent format registry efforts that will be described subsequently.

A number of formats have been defined in more than one media type, for example, XML (Extensible Markup Language) is registered in both the "text" and "application" types.[17] The existence of the catchall "application" media type has a tendency to encourage this type of cross-categorization. A more extensive classification scheme that avoids an "application"-like category has been proposed by Clausen with the intent of giving a single unambiguous category for every format:[18]

- Configuration and metadata
- Containers
- Data
- Databases
- Document-like
- Moving-image
- Program, i.e., interpretables or executables
- Program supporting
- Sound
- Spreadsheets
- Still-image
- Structured graphics, e.g., CAD/CAM (Computer Aided Design/Computer Aided Manufacturing)

Many other equally valid and useful classification schemes could undoubtedly be constructed. Since the pertinent descriptive characteristics of a format are highly dependent on the context of its use, a format may rightfully be placed into a number of different classes defined by a highly structured classification scheme. This suggests that a repeatable, faceted classification scheme is preferable to a single enumerative scheme from the perspective of capturing the nuance of format characteristics and for simplifying format discovery.

Documentation of formats

The 2000 CLIR report on *Risk Management of Digital Information* stated that "The most difficult aspect of this project was the acquisition of

[17] *Extensible Markup Language (XML) 1.0 (Third Edition)*, 4 February 2004, World Wide Web Consortium, http://www.w3.org/TR/REC-xml/ [Accessed: 22 April 2007].

[18] Clausen, L., May 2004, *Handling File Formats*, State and University Library, Denmark, http://netarchive.dk/publikationer/FileFormats-2004.pdf [Accessed: 15 April 2007].

complete and reliable file format specifications. . . . There is a pressing need to establish reliable, sustained repositories of file format specifications, documentation, and related software."[19] The final report of the University of Leeds Representation and Rendering Project highlighted a number of the similar difficulties found in sources of format information: "Existing sources of file format information fall far short of what is required to successfully tackle the problems of data obsolescence. . . . The accuracy of the majority of available file format information is mediocre at best."[20] The report recommended that efforts be made to collect and preserve format documentation through the establishment of systems of OAIS representation networks. In OAIS, representation information is a more general concept than format and is defined as any information useful for mapping from a content object to more meaningful concepts. A representation network is the set of representation information that fully describes the meaning of a content object, recognizing that this description often necessitates a recursive structure in which a particular piece of representation information may require its own representation information in order to be properly interpreted.

A large quantity of format-related information is now available on the World Wide Web. For example, the *Wotsit* web site provides links to or copies of specification documents for 1,030 formats.[21] The *FILExt* web site provides information on the mapping between file extension and format as well as software tools that can perform various processing operations on the formats.[22] Similar lists of file extensions accompanied by varying degrees of representation information are available from many other sites as well, but the inclusivity, authenticity, and sustainability of these sites is open to question.

Christensen has argued for the creation of sustainable format repositories to manage representation information about formats so that that information will be available for future curation and preservation practitioners.[23] He proposed that

[19] Lawrence, G., W. Kehoe, O. Rieger, W. Walters, and A. Kenney, June 2000, *Risk Management of Digital Information: A File Format Investigation*, Council on Library and Information Resources, http://www.clir.org/pubs/reports/pub93/pub93.pdf [Accessed: 1 August 2006].

[20] *Survey and assessment of sources of information on file formats and software documentation: Final Report*, Representation and Rendering Project, University of Leeds, http://www.jisc.ac.uk/uploaded_documents/FileFormatsreport.pdf [Accessed: 14 September 2006].

[21] *Wotsit.org: The Programmer's Format and Data Resource*, 19 April 2007, http://www.wotsit.org/ [Accessed: 19 April 2007].

[22] *FILExt: The File Extension Source*, 19 April 2007, http://filext.com/ [Accessed: 19 April 2007].

[23] Christensen, N., 2004, "Towards format repositories for web archives", *4th International Web Archiving Workshop*, http://netarchive.dk/publikationer/FormatRepositories-2004.pdf [Accessed: 20 June 2006].

these repositories manage the following information:

- Well-known format identifiers
- Procedures for determining the format of a given digital asset
- Format rendering services, systems, and tools
- Format conversion services, systems, and tools

He also emphasized the necessity for timely review and update of all managed information.

In 2000 the European Commission's Information Society Technologies (IST) programme funded the DIFFUSE project (Dissemination of InFormal and Formal Useful Specifications and Experiences) as a resource for the collection and dissemination of standards and best practices related to digital asset management.[24] Although project funding expired in 2003, the substantial set of information on formats that was collected is now available through the auspices of the Digital Curation Centre (DCC).[25] As with the IANA registry, the information presented on the DIFFUSE portal is intended for human comprehension.

The National Archives (TNA) of the UK has developed the PRONOM format registry to provide a service for both human and machine clients.[26] The PRONOM information model manages the relationships between the technical properties of formats, including classification; signatures; software, hardware, and media dependencies; and external entities such as actors, documentation, intellectual property rights, and identifiers which relate to these properties.[27] PRONOM allows information to be defined at the level of both format "families" (e.g. TIFF) and the individual formats that belong to those families (TIFF 5.0, TIFF 6.0, TIFF/IT, etc.). The registry currently contains information on over 580 formats. Future enhancements are intended to layer added-value preservation planning services on top of the existing registry structure and functionality.

The DCC and the CASPAR (Cultural, Artistic and Scientific knowledge for Preservation, Access, and Retrieval) project are developing a central registry, the DCC RegRep, for sharing OAIS representation information and to promote the use of that information in digital curation

[24] *Diffuse – Home Page*, 29 December 2003, Diffuse Project, http://web.archive.org/web/20031229131742/http://www.diffuse.org/ [Accessed: 28 December 2006]. Note that this URL refers to the archived version of the DIFFUSE project web site that is available only through the Internet Archive.

[25] *Diffuse*, 15 May 2006, Digital Curation Centre, http://www.dcc.ac.uk/diffuse/ [Accessed: 28 December 2006].

[26] *PRONOM*, 15 April 2007, National Archives, http://www.nationalarchives.gov.uk/PRONOM/ [Accessed: 15 April 2007].

[27] Brown, A., 4 January 2005, *PRONOM Information Model*, http://www.nationalarchives.gov.uk/aboutapps/fileformat/pdf/pronom_4_info_model.pdf [Accessed: 8 August 2006].

and preservation activities.[28] The registry data model is extensible and incorporates the PRONOM information model.[29] Although intended for general use, in its current developmental form it emphasizes properties necessary for curation of scientific data. The DCC RegRep is intended to be federated in the future to reduce unnecessary duplication of effort.

The Global Digital Format Registry (GDFR) project, a collaborative effort of Harvard University and OCLC, is also developing an explicitly distributed format registry network composed of independent, but interoperable nodes that will synchronize their holdings through an OAI-based protocol.[30] The GDFR data model is based on that of PRONOM with a number of functional extensions. The model makes provision for documenting the technical characteristics of formats as well as managing their specifications, intellectual property rights, related actors, and external dependencies. Specifications are represented either by actual documents directly managed and distributed in the GDFR network if appropriate rights clearances can be obtained, or if not, by bibliographic citation. These citations may include actionable links where applicable. There is also a mechanism whereby agents (individuals or institutions) can publicly register the fact of their local physical custody of specifications documents, and the arrangements under which these materials can be searched or accessed. TNA and the DCC have stated their intention to pursue interoperation with the GDFR network as it becomes operational.

An early provisional GDFR data model was used as the basis for FOCUS (FOrmat CUration Service), being developed at the University of Maryland Institute for Advanced Computer Study (UMIACS).[31] FOCUS is a component of the larger ADAPT (Approach to Digital Archiving and Preservation Technology) infrastructure.[32] The GDFR data model was also the basis

[28] *Representation Information Registry Repository*, 15 April 2007, Digital Curation Centre, http://registry.dcc.ac.uk/omar/ [Accessed: 15 April 2007].

[29] *DCCRegRepOverall*, 2 September 2005, Digital Curation Centre, http://twiki.dcc.rl.ac.uk/bin/view/Main/DCCRegRepOverall [Accessed: 15 April 2007].

[30] Abrams, S. and A. Stanescu, November 2006, "Global Digital Format Registry (GDFR): An Interim Status Report", *DLF Fall Forum*, Boston, http://www.diglib.org/forums/fall2006/presentations/Abrams-2006-11.pdf [Accessed: 9 November 2006]. Please note that the author is the architect and project manager for GDFR.

[31] Geremew, M., S. Song, and J. JaJa, May 2006, "Using Scalable and Secure Web Technologies to Design a Global Digital Format Registry Prototype: Architecture, Implementation, and Testing", *IS&T Archiving 2006*, Ottawa, http://www.umiacs.umd.edu/research/adapt/focus/documents/Archiving06.pdf [Accessed: 15 April 2007].

[32] *UMIACS ADAPT Project: An Approach to Digital Archiving and Preservation Technology*, 6 September 2006, University of Maryland, http://www.umiacs.umd.edu/research/adapt/ [Accessed: 15 April 2007].

for the FRED (Format Registry Demonstration) system deployed at the University of Pennsylvania.[33] FRED is a component of the TOM (Typed Object Model) project, a distributed brokerage system for format-related processing such as conversion.[34] FRED currently contains information on 20 common formats.

These systems are all intended to provide inclusive coverage of formats independent of intellectual domain. Other registry projects are aiming for deeper and less broadly constituted efforts. For example, the National Geospatial Digital Archive (NGDA) project is developing a format registry focusing on formats used in geospatial applications.[35] The NGDA registry uses a community-based moderated wiki as the mechanism for collecting and managing format information. The registry currently manages preliminary information on 19 common GIS (Geospatial Information Systems) formats.

[33] *Welcome to FRED*, 2 November 2004, University of Pennsylvania, http:/tom.library.upenn.edu/fred/ [Accessed: 15 April 2007].

[34] Ockerbloom, J., 7 April 2005, *The Typed Object Model (TOM)*, University of Pennsylvania, http://tom.library.upenn.edu/ [Accessed: 15 April 2007].

[35] *Format Registry Main Page*, 1 December 2006, National Geospatial Data Archive, http://ngda.library.ucsb.edu/format/index.php/Main_Page [Accessed: 15 April 2007].

Format relationships

Many formats exist within a network of associations with other formats. The most important of these relationships include *extension* and *restriction*. Format *B* is an extension of format *A* if all instances of *A* are also instances of *B*, but no instances of *B* are necessarily instances of *A*. Through the property of substitutability, the parent of the extension format can be used in all contexts of the extension. For example, UTF-8 is an extension of ASCII since all valid ASCII byte streams can be used in the context of any UTF-8-aware process without any loss of ASCII-enabled function.[36] On the other hand, using a valid UTF-8 byte stream in the context of an ASCII-only-aware process could result in some loss of UTF-8-enabled function. The extension relationship is transitive, in other words, the fact that format *C* is an extension of format *B*, which is itself an extension of format *A*, necessarily implies that format *C* is an extension of format *A*.

Restriction is the inverse of extension: format *B* is a restriction of format *A* if all instances of *B* are also instances of *A*, but no instances of *A* are necessarily instances of *B*. For example, PDF/A-1 is a restriction of

[36] ISO/IEC 646, *Information technology – ISO 7-bit coded character set for information interchange*, 1991. For UTF-8, see *The Unicode Standard, Version 5.0*, 2007, Addison-Wesley, Boston.

PDF 1.4, and once again by the property of substitutability, a PDF/A-1 file can be used without loss of function in any PDF 1.4 context, but the converse is not true: a PDF 1.4 file could lose significant function in a strict PDF/A-1 context. Restriction is also a transitive relationship. Additionally, the fact that format *A* is an extension of format *B* implies that format *B* is a restriction of format *A*, and vice versa. Thus, the choice of which relationship to use to define the association is arbitrary. As a best practice, however, the one that is most consistent with the temporal relationship of the associated formats should be used. Since PDF 1.4 was defined prior to PDF/A-1 if makes better sense to say that PDF/A is a restriction of the pre-existing PDF 1.4 format. In other words, the selected relationship should use the temporally antecedent format as its source (". . . is a restriction of PDF 1.4") and the subsequent format as its target ("PDF/A-1 is a restriction of . . .").

Both extension and restriction are specific instances of a more general *modification* relationship. For example, BWF (Broadcast WAVE Format) is a modification of WAVE, itself an extension of RIFF (Resource Interchange File Format), since BWF both extends *and* restricts WAVE by defining an additional Broadcast Audio Extension ("bext") chunk and by only allowing LPCM (linear pulse code modulation) audio.[37]

Additional relationships include *version*, *equivalence*, and *containment*:

- Version is often, but not necessarily, equivalent to extension, but implies some change in baseline function of a previous version of a format within a recognized "familial" context, generally indicated by product identification, e.g. "HTML 4.01 is a new version of 4.0."
- Equivalence indicates the syntactic-level equivalence between variant formats expressing the same semantic information, e.g., "TIFF big-endian is syntactically equivalent to TIFF little-endian."
- Containment is the ability of a format to encapsulate instances of other formatted byte streams. Arms and Fleischhauer distinguish between "wrapper" containers and "bundling" containers.[38] Wrapper formats encapsulate

[37] EBU Technical Specification 3285, *BWF – a format for audio data files in broadcasting*, July 2001 http://www.ebu.ch/CMSimages/en/tec_doc_t3285_tcm6-10544.pdf [Accessed: 19 April 2007]. For RIFF, see *Multimedia Programming Interface and Data Specifications 1.0*, August 1991, IBM, Microsoft; *Microsoft Windows Multimedia Programmer's Reference*, 1991, Microsoft.

[38] *Formats, Evaluations, and Relationships*, Library of Congress, 7 March 2007, http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml#rel [Accessed: 19 April 2007].

> metadata that describe the contained byte streams and the internal relationships between those streams, while bundling formats do not include any such description. QuickTime is a wrapper format while ZIP and TAR are bundling formats.[39]

The restriction and extension relationships are used to characterize formatted assets at appropriate levels of granularity, which, as was previously mentioned, can be significant in certain contexts. To provide a generalized rendering capability it may be sufficient merely to recognize an asset as a WAVE audio file, while the preservation of the full information content of that file could be dependent on knowing that it is in fact a Broadcast Wave File. Any migration workflow for the file must support the full range of BWF extensions over the baseline WAVE features. The containment relationship can be useful in cases where asset aggregation is necessary for purposes of operational efficiency or data exchange. The choice of a particular container format needs to be based on its support for encapsulating the full range of source formats.

Development of preservation friendly formats

With the realization that the formal characteristics of formats play a significant role in the long-term usability of digital assets, the developers of new formats have begun to pay more attention to those characteristics. JPEG 2000 (ISO 15444-1) is representative of a new generation of formats designed to include many attributes that are conducive to long-term preservation.[40] JPEG 2000 is an ISO standard with an unambiguous, community-vetted, public specification document. The format also supports calibrated color management, built-in error resilience features, provision for embedding arbitrary metadata in XML or other forms, and a flexible internal structure of nestable units, or "boxes". These boxes are self-identifying and expose their size so that processors that do not understand a particular box can skip over it without an impact on the parsing, interpretation, and rendering of the remaining portions of the object. The JPEG 2000 standard also defines important information on conformance testing and reference implementation software.

PDF (Portable Document Format) is

[39] *QuickTime File Format*, 2002, Apple, http://developer.apple.com/documentation/QuickTime/QTFF/qtff.pdf [Accessed: 19 April 2007.

[40] ISO/IEC 15444-1, *Information technology – JPEG 2000 coding system – Part 1: Core coding system*, 2000.

widely used to represent electronic documents, but in its current form it offers many features that may prove to be problematic for long-term preservation, such as dependencies on external resources, executable scripting, and encryption-based digital rights management (DRM).[41] To address this concern, the ISO Joint Working Group (JWG) ISO/TC 171/SC 2/WG 5 developed the derivative PDF/A-1 format as an archival profile of PDF 1.4. [42] (Work is now underway on PDF/A-2, which will be a profile of the current PDF 1.6 version.) The term "profile" is used to indicate a specific set of constraints placed on the generic features of the base format. In PDF/A-1 these constraints are defined in terms of two conformance levels: PDF/A-1a defines a format profile useful for the long-term preservation of the static visual appearance of electronic documents; PDF/A-1b defines additional requirements that permit the preservation of higher-level structural and semantic properties. The development process for PDF/A looked at the entire PDF 1.4 feature set and classified each of those features as either required, recommended, restricted, or prohibited on the basis of the impact that feature has on preservation activities.[43] Criteria useful for this purpose will be discussed subsequently.

Among the key determinants of preservation "friendliness" are the distinctions between proprietary and non-proprietary formats and between open and closed formats. Of these, openness, the public availability of complete authoritative specifications, is the most important. PDF, for example, is proprietary, but is fully documented in an open manner.[44] In an extreme case—the complete absence of any extant PDF rendering tools—one could nevertheless fully and properly, if tediously, interpret any PDF document by reference to the published specification. The recent work to promulgate the Office Open XML specification as an ECMA (European Computer Manufacturer's Association), and soon, an ISO standard, is a welcome move away from closed towards open formats by commercial format developers and rights holders.[45] It is

[41] LeFurgy, W., 2003, "PDF/A: Developing a File Format for Long-Term Preservation", *RLG DigiNews*, volume 7, number 6, http://www.rlg.org/legacy/preserv/diginews/v7_n6_feature1.html [Accessed: 22 April 2007].

[42] ISO 19005-1, *Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1)*, 28 September 2005. Please note that the author was the ISO project leader and document editor for PDF/A.

[43] Fanning, B., M. Warfel, S. Abrams, and S. Sullivan, August 2005, "PDF/A: The Development of a Digital Preservation Standard", *Society of American Archivists 69th Annual Meeting*, New Orleans, http://www.archivists.org/conference/neworleans2005/no2005prog-Session.asp?event=1433 [Accessed: 15 April 2007].

[44] *PDF Reference*, *Version 1.7*, November 2006, Adobe, http://www.adobe.com/devnet/pdf/pdf_reference.html [Accessed: 16 April 2007].

[45] ECMA-376, *Office Open XML File Formats*, December 2006, http://www.ecma-international.org/publications/standards/Ecma-376.htm

important for the digital curation community to continue to engage with these proprietary organizations to encourage them to provide the fullest and widest public disclosure of important specification information.

How the topic applies to Digital Curation

Digital curation is concerned with "maintaining and adding value to a trusted body of digital information for current and future use; specifically, . . . the active management and appraisal of data over the life-cycle of scholarly and scientific materials."[46] For purposes of curation the significant format-centric managerial operations are:

- Identification
- Validation
- Characterization
- Assessment

Identification answers the question, "Given an arbitrary byte stream, in what format is it encoded?"; validation answers the question, "Given a byte stream of purported format, is that format association correct?"; characterization answers the question, "Given a byte stream of known format, what are its salient formal qualities?"; and assessment answers the question, "Given a byte stream of known qualities, what are its prospects for long-term usability?" or conversely, "What is its level of risk of obsolescence and information loss?"

Format identification

The Windows operating system maintains a mapping between file extensions, format, and applications that can be used to render or process that format. Prior to version X, the Macintosh operating system used four byte file type and creator codes, often known as "four character codes" (4CC), for a similar purpose. File extensions and type codes are examples of external signatures, indicators of a file's format that exist external to the file itself. However, as external signatures can, in general, be freely manipulated by human agents they cannot be considered a reliable indicator of format. It is more reliable to use some internal characteristic of a formatted file for this purpose.

In the Unix family of operating systems, file identification is performed on the basis of a file's "magic number", its leading two byte sequence. The BSD *file*(1) command uses a database of magic numbers to determine format (conventionally

[Accessed: 16 April 2007]; ISO/IEC DIS 29500, *Information technology – Office Open XML file formats*, 30 March 2007.

[46] *About the Digital Curation Centre*, 31 May 2005, Digital Curation Centre, http://www.dcc.ac.uk/about/ [Accessed: 28 December 2006].

found at /etc/magic on Unix systems).[47] While many of these formats are somewhat specific to Unix system functions, a number of widely-used formats are also documented. The *Optima SC* freeware file identifier is based on the *magicdb* database, an extension of the standard *magic*(5) format.[48]
The notion of magic number can be generalized to that of an internal signature, any sequence of bytes that functions, alone or in conjunction with others, as an indicator of the format's identity. TNA has developed DROID, a format identification system based on matching of a large database of signatures managed in the PRONOM registry.[49] DROID uses a flexible regular-expression language to indicate the location and value of various internal signatures. The *TrID* system performs a similar identification function.[50] This tool is capable of automatically determining relevant format signatures by examining collections of formatted files for commonly-appearing patterns. Many other similar format detection systems, both commercial and open source, are available.

Curation and preservation activities cannot be performed effectively without the existence of an authoritative inventory of the digital assets under consideration. Format identification is the initial component in the process of creating such an inventory. Having a robust system for determining the format of digital assets is particularly important in contexts in which assets are routinely unaccompanied by authoritative technical metadata, such as institutional repositories (IR) or web archiving.

Format validation

Signature-based identification is indicative, not definitive. (The open source *Fine Free* version of the BSD *file*(1) command calls itself "a file type guesser".) It is only through a process of validation, parsing the entirety of a bit stream and determining its conformance to published specifications, that a format's identification can be stated definitively. However, the question of what exactly is being validated can be ambiguous. For example, many rendering tools are purposively constructed to be forgiving of

[47] *The Fine Free File Command*, http://www.darwinsys.com/file/ [Accessed: 15 March 2007].

[48] Optima SC, Inc., *File format information, magic database and file identifier*, 30 March 2005, http://www.magicdb.org/ [Accessed: 26 March 2007]; Optima SC, Inc., *File identifier for Windows, DOS and Linux*, 6 January 2006, http://www.optimasc.com/products/fileid/ [Accessed: 26 March 2007].

[49] Brown, A., October 2005, *Automatic Format Detection Using PRONOM and DROID*, National Archives, http://www.nationalarchives.gov.uk/aboutapps/fileformat/pdf/automatic_format_identification.pdf [Accessed: 13 September 2006].

[50] Pontello, M., 26 March 2007,*TrID – File Identifier*, http://mark0.net/soft-trid-e.html [Accessed: 26 March 2007].

content that is not absolutely conformant to specifications. For example, the *Acrobat* PDF renderer is capable of recovering from a large number of formatting violations and still render properly. In such a case, a finding of format invalidity, while technically true, does not necessarily mean that the content is unusable in many practical contexts.

However, in the context of long-term curation activities, it is preferable that strict validation is applied and enforced, and if necessary, curated assets should be regenerated or normalized to enforce full compliance. The aphorism "Forgiving code encourages sloppy practice" is germane in this context. While contemporary tools may be constructed to deal with a wide range of format inconsistencies, the creators of future tools may not have that local knowledge. Thus, assets that are acquired today under an assumption of current usability may be unusable in the future.

A wide range of tools is available for the validation of various formats, mostly specialized to particular formats. In the PDF realm, for example, there are a number of "pre-flight" tools that are used extensively as part of pre-press data exchange workflows. In a more generic context, Harvard University and JSTOR have developed the open source JHOVE (JSTOR/Harvard Object Validation Environment, pronounced "jove") tool for format identification, validation, and characterization.[51] JHOVE uses an extensible plug-in mechanism to recognize and distinguish between 11 audio, document, image, and text format families (AIFF, ASCII, GIF, HTML, JPEG, JPEG 2000, PDF, TIFF, UTF-8, WAVE, XML) in over 70 profile variations. In its current form, JHOVE generally assumes implicitly that a single digital asset is always manifest in a single file with a single format. Of course there are many important cases where this assumption does not hold. For example, as previously noted a TIFF file can contain an ICC color profile and various types of independently-formatted metadata while a Shapefile asset is always manifest in three independently-formatted files. Additional work is planned to enhance JHOVE to support a more sophisticated data model in which an asset can be manifest in an arbitrary number of files or nested byte streams and formats.

The validation capabilities of JHOVE have been extended to meet specific project profiles for the National Digital Newspaper Project (NDNP) at the Library of Congress.[52] The NDNP project was

[51] *JHOVE – JSTOR/Harvard Object Validation Environment*, 29 March 2006, Harvard University, http://hul.harvard.edu/jhove/ [Accessed: 15 April 2007]. Please note that the author was architect and project manager for JHOVE.

[52] Littman, J., May 2006, "A Technical Approach and Distributed Model for Validation of Digital Objects",

concerned firstly that the image files were well-formed and valid with respect to their format, and secondly that they conform to a more restricted set of technical properties than otherwise allowed by the format specification. For example, while the TIFF format permits the use of a wide number of color spaces and compression schemes, the NDNP project performed additional validation to ensure that the TIFF images were always defined using a grayscale color space and were uncompressed. These project profiles have been formally codified as human-readable rules documents.[53] This type of profile-based validation, especially useful to detect systemic technical workflow failure, is an important acceptance criterion for the acquisition of digital material, whether produced by internal or external vendors.

Characterization

The National Library of New Zealand has developed a Metadata Extraction Tool that can characterize 15 audio, video, document, and *Microsoft Office* formats.[54] The tool uses an extensible architecture of pluggable adaptors that automatically extract preservation-related metadata from digital files, outputting that metadata in a standard XML format suitable for uploading into a preservation metadata repository. A refactored version of the tool has recently been released under an open source license.

While JHOVE and the New Zealand extraction tool use specific software adaptors each customized for a single format, the European Union-funded PLANETS Project is developing a characterization scheme that requires only a single generic processing system known as XCL, which is applicable to any format that can be described in terms of two XML-based languages. XCDL is a characterization definition language that can abstractly express the form of arbitrary digital content. XCEL is a characterization extraction language that describes how to retrieve data from a given file.[55] While the XCL approach still requires extensive knowledge of a format's

D-Lib Magazine, volume 12, number 5, http://www.dlib.org/dlib/may06/littman/05littman.html [Accessed June 5, 2006].

[53] *Technical Specifications – Profiles and Schemas: National Newspaper Preservation Program*, 31 August 2006, Library of Congress, http://www.loc.gov/ndnp/techspecs.html [Accessed: 16 April 2007].

[54] *Metadata Extraction Tool*, 16 July 2007, National Library of New Zealand, http://meta-extractor.sourceforge.net/ [Accessed: 23 August 2007].

[55] Heydegger, V., J. Neumann, J. Schnasse, and M. Thaller, 31 October 2006, *PLANETS: Basic design for the extensible characterization language*, http://lehre.hki.uni-koeln.de/planets/documents/deliverables/PlanetsPC2D1D2-end.pdf [Accessed: 23 August 2007].

specification, that knowledge is used to produce characterization documents, not software modules, which may streamline the ability to support a wider range of formats.

The schema used to characterize a format should be selected on the basis of its ability to express fully the significant properties of that format. JHOVE for example, reports still image metadata in terms of the NISO Z39.87 data dictionary and its associated MIX schema, audio metadata in terms of the evolving AES-098B schema, and text metadata in terms of the METS text extension schema.[56] Significant properties, a concept that was first articulated in the CAMiLEON project, are those aspects of a digital asset that carry its important or essential meaning.[57] The UK Arts and Humanities Data Service (AHDS) has published useful guides to significant properties for a number of format genres.[58] The InSPECT (Investigating the Significant Properties of Electronic Content over Time) project is developing comprehensive sets of significant properties for raster image, email, structured text, and audio formats.[59] The project will also result in a tested methodology for determining the significant properties of formats, an increased understanding of the results of preservation strategies, an assessment of the performance of various formats in preserving the properties of various formats, and suggestions for enhancements to the PRONOM registry to implement project outcomes. An additional JISC Invitation to Tender (ITT) has recently been announced for a follow-on study of the significant properties of four additional format genres: e-learning objects, software, vector images, and moving images.[60]

[56] NISO Z39.87, *Data Dictionary – Technical Metadata for Digital Still Images*, 2006, http://www.niso.org/standards/resources/Z39-87-2006.pdf [Accessed: 19 April 2007]; *Metadata for Images in XML Standard (MIX)*, 7 March 2007, Library of Congress, http://www.loc.gov/standards/mix/ [Accessed: 19 April 2007]; AES-X098B, *Administrative and structural metadata for audio objects*; *METS Extenders: Metadata Encoding & Transmission Standard (METS) Official Web Site*, 13 September 2007, Library of Congress, http://www.loc.gov/standards/mets/mets-extenders.html [Accessed: 19 April 2007].

[57] Wheatley, P., September 2001, "Migration – a CAMiLEON discussion paper", *Ariadne*, volume 29, http://www.ariadne.ac.uk/issue29/camileon/ [Accessed: 24 August 2007]. See also Hedstrom, M. and C. Lee, May 2002, "Significant properties of digital objects: definitions, applications, implications", *Proceedings of the DLM-Forum 2002*, Barcelona, 218-227, http://ec.europa.eu/transparency/archival_policy/dlm_forum/doc/dlm-proceed2002.pdf [Accessed: 15 April 2007]; and Heslop, H., S. Davis, and A. Wilson, 2002, *An Approach to the Preservation of Digital Records*, National Archives of Australia, http://www.naa.gov.au/recordkeeping/er/digital_preservation/Green_Paper.pdf [Accessed: 15 April 2007].

[58] *InSpect project*, 25 January 2007, Arts and Humanities Data Service, http://ahds.ac.uk/about/projects/inspect/index.htm [Accessed: 15 April 2007].

[59] *Investigating the Significant Properties of Electronic Content Over Time (INSPECT)*, 1 April 2007, Joint Information Systems Committee, http://www.jisc.ac.uk/whatwedo/programmes/programme_rep_pres/INSPECt.aspx [Accessed: 15 April 2007].

[60] *Significant Properties ITT*, 26 March 2007, Joint Information Systems Committee, http://www.jisc.ac.uk/fundingopportunities/funding_calls/2007/03/significant_properties_itt.aspx [Accessed: 15 April 2007].

Assessment

The intent of any preservation risk assessment process is to correlate the *probability* of risk of loss of information content with the *impact* of that risk. Whenever possible activities and objects whose characteristics are determined to fall into the low probability/low impact range should be substituted for those in the high probability/high impact range. Stanescu has identified the main factors contributing to preservation risk as software, hardware, associated organizations, the nature of the preservation archive, specific preservation plans, and format.[61]

The Groupe Pérennisation des Informations Numériques (PIN) has defined three high-level criteria for evaluating formats in terms of their suitability for preservation:[62]

- The ability to represent the complexity and fullness of meaning of the underlying information content.
- Conformance to public standards.
- The ease with which a formatted asset can be modified without loss of its information content.

It is also important that a format is selected for preservation purposes only if the organization making that selection is familiar with all aspects of that format. The selection process should have a general preference for formats that have an abundance of processing tools; that have simple, rather than complex, specifications; that can be validated for conformance to those specifications automatically; and that can facilitate the automated extraction of embedded metadata.

Arms and Fleischhauer have developed an analytical framework based on factors for format sustainability, functionality, and quality (SFQ).[63] The sustainability factors are applicable across format genres, including:

- Disclosure, the degree to which complete authoritative specifications are publicly available.
- Adoption, the degree to which the format is in widespread use. Adoption may be the most important factor in format preservation as it indicates that there is a significant economic incentive to maintain the

[61] Stanescu, A., January 2005, "Assessing the Durability of Formats in a Digital Preservation Environment: The INFORM Methodology", *OCLC Systems and Services*, volume 21, number 1, pp. 61-81.

[62] Huc, C., et al., 11 May 2004, *Criteria for evaluating data formats in terms of their suitability for ensuring information long term preservation*, Version 5, Groupe Pérennisation des Informations Numériques http://www.ssd.rl.ac.uk/ccsdsp2/mon04/long_term_preservation_criteria.doc [Accessed: 26 March 2007].

[63] Arms, C., and C. Fleischhauer, 2005, "Digital Formats: Factors for Sustainability, Function, and Quality", *IS&T Archiving 2005 Conference*, Washington, DC, http://memory.loc.gov/ammem/techdocs/digform/Formats_IST05_paper.pdf [Accessed June 5, 2006].

usability of the format.

- Transparency, the degree to which the format is open to analysis by non-format aware tools. Unencrypted, uncompressed, textually-encoded formats are more transparent than encrypted, compressed, binary-encodings.
- Self-documentation, the degree to which a formatted asset can contain its own description.
- External dependencies, the degree to which the use of a formatted asset requires the existence of external resources.
- Impact of patents, the degree to which the use of a format is encumbered by intellectual property claims.
- Technical protection mechanisms, the degree to which constraints on the use of a formatted asset are enforceable through technical means such as encryption.

Functionality and quality factors are those that are dependent upon the significant properties of a particular format genre. For example, in the image genre these factors might include the ability to support high image resolution and calibrated color management workflows.

A similar set of assessment criteria has been developed by TNA.[64] While repeating the importance of disclosure, this recommendation also stresses the *stability* of specifications. Formats whose specifications undergo frequent changes, or changes that are not backward compatible, should be avoided for preservation purposes. TNA also states a preference for formats that can enforce bit-level integrity through built-in mechanisms for checksums and error-correcting codes. Under an assumption that a migration strategy will be used for long-term preservation, formats should also be evaluated with respect to their support for mechanisms for authenticity, processability, and presentation.

The format assessment guidance from the Danish Royal Library and the State and University Library of Århus requires specific compatibility with OAIS (ISO 14721).[65] Other important aspects of this guidance include:

- Support for all important Internet protocols
- Support for embedded metadata
- Support for recoding of access limitations
- Support for future transformation
- Support for compression

[64] Brown, A., 19 June 2003, *Selecting File Formats for Long-Term Preservation*, National Archives, http://www.nationalarchives.gov.uk/documents/selecting_file_formats.pdf [Accessed: 13 September 2006].

[65] Christensen, S., July 2004,*Archival Data Format Requirements*, Royal Library, Denmark, http://netarkivet.dk/publikationer/Archival_format_requirements-2004.pdf [Accessed: 13 September 2006].

- Efficiency in processing

As was noted previously, the use of compression *lowers* the degree of transparency and should therefore be avoided. However, compression may become necessary in the context of the economic and operational cost of archiving ever increasing amounts of data. Nevertheless, compression should be used only after careful consideration of its potential impact on long-term preservation activities.

In addition to these proposals for generic assessment criteria applicable for all formats, more focused guidance has been proposed for specific format genres. For example, Folk and Choi have analyzed the criteria necessary for geospatial formats.[66] Folk and Barkstrom have performed a similar analysis for scientific and engineering formats.[67] Due to the large size of many data sets in this realm, their recommendations include many criteria specific to the organization, size, and raw I/O efficiency of formatted assets. In order to facilitate the free exchange of data amongst an international community of scholars, they also emphasize support for multi-language implementation of processing software. They also suggest the desirability of a format being able to encapsulate the software necessary to render itself, an extreme form of the general self-containment property.

Topic in action

Standardization

As mentioned previously, one of the attributes of preservation formats is conformance to standards. A number of recent formats have been explicitly developed through an accredited standard process, such as JPEG 2000 (ISO 15444). Others, such as PDF/A (ISO 19005-1), TIFF/EP (ISO12234-2), TIFF/IT (ISO 12639), Open Document Format (ODF, ISO/IEC 26300), and Office Open XML (ISO/IEC DIS 29500), were existing formats, or newly developed variants, that were subsequently promulgated through an accredited standards process.[68]

In contrast to these *de jure* standards, many popular formats fall into the category of *de facto* standardization on the basis of ubiquity. Although of potentially broad applicability, these

[66] Folk, M., and V. Choi, 8 January 2004, *Scientific formats for geospatial data preservation: A study of suitability and performance*, National Center for Supercomputing Applications, National Archives and Records Administration, http://www.ncsa.uiuc.edu/NARA/Sci_fmts_and_geodata_HDF.pdf [Accessed: 13 September 2006].

[67] Folk, M., and B. Barkstrom, May 2003, "Attributes of File Formats for Long-Term Preservation of Scientific and Engineering Data in Digital Libraries", *JCDL '03*, Houston, http://www.ncsa.uiuc.edu/NARA/Sci_Formats_and_Archiving.doc [Accessed: 13 September 2006].

[68] ISO/IEC 26300, *Information technology – Open Document Format for Office Applications (OpenDocument) v1.0*, 2006.

standards are generally the result of parochial community interest. For example, the Broadcast WAVE Format (BWF) was developed by the European Broadcasting Union to simplify the interchange of broadcast media. The International Color Consortium specified the ICC.1 color profile format.[69] Members of the geospatial data community have collaborated on an extension to the TIFF image format to create GeoTIFF. TIFF itself remains a proprietary format. While ISO did define JPEG as a compression algorithm, the format associated with that algorithm, Still Picture Image File Format (SPIFF, ISO/IEC 10918-1) is not widely supported.[70] The JPEG File Interchange Format (JFIF), developed by a commercial vendor, is the widely supported form of JPEG-compressed data.[71]

All of these formats, whether proprietary or not, are open, rather than closed, with complete published specification documents. The OpenRAW forum was founded in an effort to encourage similar behavior from camera manufacturers.[72] Many professional, and some consumer, cameras produce their highest quality images in proprietary "raw" image formats that can be processed by a limited range of largely proprietary tools.[73] For the most part, these raw formats are undocumented and therefore have poor prospects for long-term usability. The *Adobe DNG* (Digital Negative) format is an effort, although one directed by a commercial interest, at establishing an open format for the interchange of raw image data. From the point of view of preservation disclosure, Adobe has been responsive in publishing the specifications for many of its formats and offering perpetual, no-cost licensing for technology covered by many of its intellectual property rights claims.

In a related effort regarding camera vendors, the RLG Automatic Exposure initiative is aimed at encouraging those vendors to make full use of the existing features of standard formats regarding embedded technical and

[69] ICC.1, *Image technology colour management – Architecture, profile format, and data structure*, October 2004, International Color Consortium, http://www.color.org/ICC1V42.pdf [Accessed: 15 April 2007].

[70] ISO/IEC 10918-1, *Information technology – Digital compression and coding of continuous-tone still images: Requirements and guidelines*, 1994.

[71] Hamilton, E., 1 September 1992, *JPEG File Interchange Format* Version 1.02, C-Cube Microsystems, http://www.w3.org/Graphics/JPEG/jfif3.pdf [Accessed: 15 April 2007].

[72] *OpenRAW: Digital Image Preservation Through Open Documentation*, 2 April 2007, http://www.openraw.org/ [Accessed: 21 April 2007].

[73] Bates, M., S. Manuel, S. Loddington, and C. Oppenheim, May 2006, *Digital Lifecyles and File Types: Final Report, JISC Digital Repositories Programme: Rights and Rewards in Blended Institutional Repositories Project*, Joint Information Systems Committee, http://rightsandrewards.lboro.ac.uk/files/resourcesmodule/@random43cbae8b0d0ad/1148047621_DigitalLifecyclesV2.pdf [Accessed: 19 May 2006].

administrative metadata.[74] Image formats such as TIFF, JPEG, and JPEG 2000 have the capability to embed metadata in various forms, such as DIG35, Exif (JEITA CP-3451), NISO Z39.87/MIX, or XMP.[75] The automatic population of this metadata at the point of capture is a desirable attribute for preservation images as it fulfills the self-documentation property.

While standardization is obviously better than non-standardization, the mere existence of standards does not necessarily mean that they will be widely implemented. For example, the JPEG 2000 image format is an ISO standard, but it is nevertheless not widely supported by the current generation of web browsers, although less preservation-friendly formats such as GIF (Graphics Interchange Format), JPEG, and PNG (ISO/IEC 15948) are supported.[76] The recent announcement that native support for JPEG 2000 in the *Firefox* browser will be implemented as part of the *Google Summer of Code* is therefore a welcome development.[77]

Ingest workflow

The initial point at which consideration of format is important is the point of creation of the digital asset. As discussed previously, assets should be created with strict conformance to the specifications of their formats. This conformance should be validated both prior to and following ingestion of assets into curatorial or preservation systems. (For an example workflow distributed between client and server agents, see Figure 2.) The client-side validation and characterization ensures that errors are detected as far upstream in the production process as possible, the point at which they are most easily correctable, in the most authoritative manner, and with the least effort. The result of successful validation and characterization is the production of a compliant Submission Information Package (SIP). Server-side validation and characterization ensures that repository-specific practices

[74] *Automatic Exposure – Technical Metadata*, 2006, Research Libraries Group, http://www.rlg.org/en/page.php?Page_ID=2681 [Accessed: 21 April 2007].

[75] *I3A Standards – Initiatives – DIG35*, 21 April 2007, International Imaging Industry Association, http://www.i3a.org/i_dig35.html [Accessed: 21 April 2007]; JEITA CP-3451, *Exchangeable image file format for digital still cameras: Exif Version 2.2*, April 2002, Japan Electronics and Information Technology Industries Association, http://www.exif.org/Exif2-2.PDF [Accessed: 21 April 2007].

[76] *Graphics Interchange Format*, version 89a, 31 July 1990, Compuserve Inc., http://www.w3.org/Graphics/GIF/spec-gif89a.txt [Accessed: 21 April 2007]; ISO/IEC 15948, *Information technology – Computer graphics and image processing – Portable Network Graphics (PNG): Functional specification*, 10 November, 2003, http://www.w3.org/TR/PNG/ [Accessed: 21 April 2007].

[77] *Firefox – Rediscover the Web*, 20 April 2007, Mozilla, http://www.mozilla.com/en-US/firefox/ [Accessed: 20 April 2007]; *Google code – Summer of Code – Application Information: JPEG 2000 Support for Firefox*, 20 April 2007, Google, http://code.google.com/soc/mozilla/appinfo.html?csaid=C7B9CCBBF96648B3 [Accessed: 20 April 2007].

regarding SIP packaging and asset format are followed and that a proper inventory of managed digital assets can be generated. This validation step is absolutely necessary in cases where client-side checking is not performed, or if the criteria for that checking are not identical with the server-side checks. Even if it is identical, server-side checking is still justified to ensure compliance and to detect systemic flaws that may be introduced into client-side systems.

The Library of Congress organized the Archive Ingest and Handling Test (AIHT) to investigate issues that might arise in the large-scale exchange of data between heterogeneous preservation repositories.[78] Although the participating institutions used many of the same tools, variations introduced by the versioning of those tools led to divergent interpretations of the formats and the validation conformance of the exchanged assets. This experience points out the importance of bi-lateral submission agreements for both the form of submitted content as well as the conformance regime.

[78] Shirkey, C., December 2005, "Conceptual Issues from Practical Tests", *D-Lib Magazine*, volume 11, number 12, http://www.dlib.org/dlib/december05/shirky/12shirky.html [Accessed: 21 April 2007]. Please note that the author was a participant in the AIHT project.

Preservation strategies

An in-depth discussion of preservation strategies is provided in other chapters of this manual. However, there are some salient points regarding format in the selection of a particular preservation strategy. The previously enumerated assessment criteria provide general guidance on the selection of the formats in which to preserve content. Depending upon the nature of the archival repository, however, the potential for such a choice may not exist. Institutional repositories, for example, generally are required to accept all submitted assets regardless of their format. Nevertheless, best practice guidance with regard to the selection of preservation-friendly formats serves an important purpose in educating the creators of digital content to be cognizant of the repercussions of their choices.

In addition to utilizing formats that allow the capture of the richest possible set of information, Kunze suggested using "desiccated" formats to create easily-preserved, low-technology derivatives that support the minimal representation of essential information.[79] For textual or document-based content, the desiccated form would be ASCII; for still images, a simple uncompressed

[79] Kunze, J., November 2005, "Web Archiving Service (WAS)", *DLF Forum*, http://www.diglib.org/forums/fall2005/presentations/kunze-2005-11.pdf [Accessed: 21 April 2007].

bit-map. The intent of this recommendation is not to countenance the discarding of information content, but to lower the technological requirements for the "copy of last resort." In the event, hopefully unlikely, that the more fully-featured version of the content can no longer be preserved, the presumably longer-lived desiccated version would continue to persist in usable form and provide some approximation of the original information content. In other words, desiccated formats are used to capture a set of significant properties that have been pared down to the barest minimum. The rational for desiccated formats is similar to Uneson's injunction that preservation formats should focus on representation, not presentation.[80]

The question of the timing of format obsolescence and the periodicity with which concomitant preservation intervention will be required has been examined by Rusbridge, who suggested that such obsolescence will occur at a much slower pace than was previously thought.[81] He found that the ubiquity of, and dependence on, digital information is so fundamentally entwined with all aspects of contemporary economic, intellectual, and entertainment pursuits that some form of that information will have a tendency to persist. The two caveats he raises in this regard are to disruptive technological change and extended periods of time.

The negative impact of format obsolescence can be mitigated through the use of emulation strategies. Emulation focuses on the obsolescence of systems for processing formatted assets, not the assets themselves, or their formats. In theory, an organization implementing an emulation strategy can forgo any consideration of format assessment and selection and the collection and management of format representation information, other than selecting formats supported by tools for which emulators are available.

The other main strategic choice for preservation is migration, the periodic transformation from incipiently obsolete formats to contemporaneously viable formats. Migration can be implemented in two forms: early migration ("just in case") or late migration ("just in time"). Under an early migration scheme all instances of an affected format are transformed following a determination that the format is approaching obsolescence. This scheme has the advantage of homogeneity: all instances of format

[80] Uneson , M., 1 September 2005, "Tomorrow's File Endings: On Archiving Principles and Archiving Formats", *ScieCom Info*, volume 2, http://www.sciecom.org/sciecominfo/artiklar/uneson_05_2.pdf [Accessed: 2 January 2007].

[81] Rusbridge, C., February 2006, "Excuse Me. . . Some Digital Preservation Fallacies?" *Ariadne*, issue 46, http://www.ariadne.ac.uk/issue46/rusbridge/ [Accessed: 21 April 2007].

A are converted to format *B* at the same time, with the same workflow. Its primary disadvantage is the difficulty in making the determination concerning obsolescence and the chance that the migration is performed earlier than is absolutely necessary, perhaps with less than efficient or reliable tools resulting in a less than desirable outcome. Another significant disadvantage is the potential for the accumulation of incremental information loss through the repeated migration activity: format *A* to *B*, *B* to *C*, *C* to *D*, and so forth.

Late migration postpones the migration activity until it is known to be necessary, and then the activity is applied on an as needed basis at the point of request for a particular asset. Furthermore, in the form propounded by the CAMiLEON project, the transformation is always applied against the original form of the content: from format *A* to format *B*, *A* to *C*, *A* to *D*, etc., to avoid accumulation error.[82] This approach has the advantage that the minimal work is performed, and that by putting off that work until absolutely necessary, the greatest degree of technological knowledge is available to implement the transformation. The LOCKSS (Lots of Copies Keeps Stuff Safe) project has developed a proof-of-concept system for this approach using GIF-to-PNG conversion for harvested web content.[83] Late migration is also the principle underlying the Multivalent browser, which uses an extensible set of media adaptors to provide contemporaneous support for behaviors primarily applicable to document-centric content whose formats are invariant over time.[84]

Since 2004 Cornell University has operated a File Format and Media Migration Pilot Service (FFMM) to help maintain the continuing viability of faculty-produced data stored in obsolete, or obsolescing, formats and media.[85] The provisional service was developed locally in reaction to the perception that commercial data conversion and recovery services were prohibitively expensive. Format migration was performed following extensive manual analysis of data to determine source and prospective target formats

[82] Mellor, P., P. Wheatley, and D. Sergeant, September 2002, "Migration on Request, a Practical Technique for Preservation", *6th European Conference on Digital Libraries*, Rome, http://www.si.umich.edu/CAMILEON/reports/migreq.pdf [Accessed: 21 April 2007].

[83] Rosenthal, D., T. Lipkis, T. Robertson, and S. Morabito, January 2005, "Transparent Format Migration of Preserved Web Content", *D-Lib Magazine*, volume 11, number 1, http://www.dlib.org/dlib/january05/rosenthal/01rosenthal.html [Accessed: 21 April 2007].

[84] Phelps, T. and P. Watry, September 2005, "A No-Compromises Architecture for Digital Document Preservation", *9th European Conference on Digital Libraries*, Vienna, http://multivalent.sourceforge.net/Research/Live.pdf [Accessed: 21 April 2007].

[85] Entlich, R. and E. Buckley, 15 October 2006, "Digging Up Bits of the Past: Hands-on With Obsolescence", *RLG DigiNews*, volume 10, number 5, http://www.rlg.org/en/page.php?Page_ID=20987 [Accessed: 16 November 2006].

and to identify appropriate processing tools.

Notification and recommendation systems

To remove the human component from migration workflows, Ferreira et al. have proposed the development of automated migration tools.[86] The resulting CRiB (Conversion and Recommendation of Digital Object Formats) system is based on a service-oriented architecture (SOA) and is composed of separate Service Registry, Format Detector, Format Evaluator, Migration Advisor, Migration Broker, and Object Evaluator services.[87] The Format Evaluator uses a built-in Format Knowledge Base, but could be modified to interoperate with external services. The Migration Advisor develops alternative migration paths for the source format identified by the Detector. These paths are based on services known to the Registry. The Object Evaluator performs post-transformation quality assessments.

The PANIC (Preservation services Architecture for New media and Interactive Collections) system developed by the Australian Distributed Systems Technology Center (DTSC) provides similar automated processing by dynamically constructing workflows from preservation web services described in terms of a machine-processable ontology.[88] PANIC is composed of four main constituents: an Invocation component, including sub-components for obsolescence detection, service discovery, service selection, and service invocation; a Notification component, including services and standards registries; a Discovery component by which users interact with the system; and a Provider component responsible for performing the requested actions. PANIC can provide services for both migration and emulation of formatted assets.

The National Library of Australia developed the AONS (Automated Obsolescence Notification System) framework to perform automated analysis of preservation risk. AONS retrieves format information from PRONOM and recommendations from the Library of Congress Sustainability of Digital Formats

[86] Ferreira, M., A. Baptista, and J. Ramalho, July 2006, "A Foundation for Automatic Digital Preservation", *Ariadne*, volume 48, http://www.ariadne.ac.uk/issue48/ferreira-et-al/ [Accessed: 21 April 2007].

[87] *CRiB: Conversion and Recommendation of Digital Object Formats*, 16 February 2007, University of Minho, http://crib.dsi.uminho.pt/ [Accessed: 16 February 2007].

[88] Hunter, J., and S. Choudhury, June 2004, "A Semi-Automated Digital Preservation System Based on Semantic Web Services", *Joint Conference on Digital Libraries*, Tucson, http://www.itee.uq.edu.au/~eresearch/papers/2004/jcdl2004.pdf [Accessed: 21 April 2007].

web site.[89] AONS interoperates with the DSpace open source repository and prototypical work has been starting on integration with the Fedora/Fez repository.[90] A second generation system, AONS II, is now under development and is intended to be structured in a more repository-agnostic manner to facilitate interoperability with a wide variety of repository systems.[91] The design of the proposed system includes six modular services: a Format service, which manages a format information database based on information gleaned from a number of external sources; an Obsolescence service for assessing preservation risk; an Action service for inter-module message passing; a Local Crawl service used to harvest data from a remote repository and determine the formats of the harvested assets; a Collection service to enable interoperability in local and federated service environments, and a Web Interface service to provide human interfaces to the system.

[89] *Sustainability of Digital Formats: Planning for Library of Congress Collections*, 20 October 2006, Library of Congress, http://www.digitalpreservation.gov/formats/ [Accessed: 28 December 2006].

[90] Curtis, J., 29 September 2006, *AONS Systems Documentation*, National Library of Australia, http://www.apsr.edu.au/ [Accessed: 20 December 2006]; *Welcome to DSpace*, 21 April 2007, MIT, Hewlett-Packard, http://www.dspace.org/ [Accessed: 21 April 2007]. *Fedora*, 21 April 2007, Fedora Project, http://www.fedora.info/ [Accessed: 21 April 2007].

[91] Walker, M., *AONS II Technical Architecture*, 16 February 2007, National Library of Australia, http://wiki.nla.gov.au/download/attachments/7723/AONSII_TechnicalArchitecture.doc [Accessed: 21 April 2007].

Next steps

The curation, repository, and preservation communities are at the brink of major improvements in the understanding and management of formats in preservation and curation contexts. Systems, practices, and formats are now being designed with long-term preservation as an explicit consideration and requirement. This will have a significant impact on the range of options available to curation practitioners and the probability of favorable outcomes of preservation actions. In order for this to happen, however, it is important that deep and broad knowledge of format-related information is collected, managed in a sustainable manner, and made available to practitioners at the point at which it is necessary. Interoperable format registries such as DCC RegRep, GDFR, and PRONOM will provide the technical environment for the preservation and dispersion of this material. However, the value of these registries lies in their acquiring comprehensive coverage of authentic information for the widest range of formats. It is not feasible at an economic, administrative, or technical level for the collection of this information to be completely centralized. It is therefore incumbent upon the collective membership of the curation community to participate in the process of documenting the panoply of formats in past, current, or future use. The various

international centers of expertise organized around content and format-genres, both at the individual and institutional level, are best positioned to provide this vital service. The aggregation function provided by these registries will serve to multiply the value of the supplied information by pooling the distributed expertise of disparate preservation practitioners and making it available for use by a much wider community.

In addition to representation information about formats themselves, it is also important that a continuing effort is made to develop best practice guidelines regarding the selection of formats in specific curation contexts, and the technical profiles that will be suggested or required for that use. These recommendations should be focused on increasing the use of preservation-friendly formats. At the same time the quality of the processing behavior exhibited by the tools used to manipulate formatted assets should be directed towards rigorous conformance to the stated specifications of the underlying formats. This may necessitate continual engagement with tool vendors and developers to reiterate the importance of this conformance to long-term preservation efforts and the deleterious consequences of non-conformance, leading in the worse case to the irretrievable loss of significant information content.

Best practice recommendations are dependent on assumptions concerning the significant properties of digitally-encoded assets. The JISC-funded projects looking at the determination of properties for a number of commonly-used content genres and the concomitant development of a formal process for making such determinations are important first steps in this direction. The wider community should become involved in the review and refinement of these determinations and methodological frameworks. As the boundaries of content genres, and the behavioral expectations concerning the use of that content, continue to expand over time and entirely new genres continually evolve, this will be an ongoing process that needs to be institutionalized in the routine policies and practices of curation programs and organizations.

At the same time it is important to extend the notion of format-like typing beyond the current byte stream and file level to the aggregate object level, with the same necessity for well-developed analysis of significant properties, best practice recommendations, and rigorous conformance criteria. Moving the target of analysis and planning to the aggregate content model level will provide significant advantages in terms of operational efficiency, a major concern of repository and preservation workflows in the face of ever increasing scale in the number

and size of assets under management.

Many of these functions are being deployed or are under active development by various international institutions operating in the curation arena. As the growth in the management of digital assets, and the increasingly sophisticated nature of those assets, continues at increasing rates, the mutual advantage of more focused cooperation and interoperation between these institutions becomes more attractive and, to some degree, necessary. It is reasonable to assume that various centers of expertise concerning curation activities will develop along the lines of content and format genres. In the case of audio and moving image content, for example, an understanding of the technical details of the relevant formats alone is not sufficient. It is also necessary to monitor ongoing developments in server-side delivery platforms and client-side rendering platforms. Additionally, preservation workflows may encompass the use of highly-specialized playback and editing equipment. This level of expertise, and the hardware and software environment concomitant to it, may exceed the financial and staff capabilities of smaller curation institutions and programs.

In a decentralized world, however, where digital assets can flow freely to those points within a global network at which appropriate processing can take, local expertise may not be necessary. Instead, effective and efficient content curation can become an inherently multi-institutional process. Such a distributed environment is dependent upon increasing technical capacities for the transfer of significant bodies of digital materials, agreement on the packaging and description of that material, and perhaps most problematically, the development of appropriate costing models and multi-lateral business arrangements. Although this emerging curation environment is complex and will undoubtedly be slow to evolve, it appears to be necessary given the heterogeneity and complexity of the formats that form the technical underpinnings of the collections to which curation responsibilities are attached. It is unlikely that any individual institution with genre-spanning acquisition policies will be able to develop sufficiently deep and broad local expertise to curate the variety of content placed under management without significant interaction with external partners.

Future developments

As appropriate analytical frameworks become well articulated for the characterization and assessment of formatted assets at varying conceptual levels, both byte stream and object, the curation

community can expect to make subsequent strides in the automation of workflows that implement complete and increasingly sophisticated chains of management and preservation activities. Such a workflow might comprise the following actions taken in response to the acceptance of curatorial responsibility for an asset of unknown provenance and formal attributes: byte stream-level format identification, validation, and characterization; a similar set of actions applied at the object level; and finally, preservation risk assessment incorporating locally-configured rules and heuristics based on institutional determinations of technical capabilities and levels of tolerance. As it becomes necessary over time, subsequent intervention recommendation and notification, workflow generation and invocation, and post-intervention quality assurance and re-characterization should be obtainable as automated features of future curation systems. This movement away from current manual workflows to more highly automated workflows will become necessary for curation programs to apply more than cursory attention to the ever growing body of digital content submitted for active management.

Most of these processes can be seen as being added-value extensions or new services layered on top of existing, or rapidly evolving, format registries; validation, characterization, and transformation tools; and service-oriented brokerage systems. Due to the lack of clear empirical evidence as to the relative efficacy of emulation versus migration-based preservation strategies, these future services should make free provision for both strategic directions. Of course, the availability of these two parallel paths for intervention is desirable in its own right. A heterogeneity of approaches, especially if utilized on a common body of managed content, will increase the prospects of successful outcomes by reducing the impact of systemic failures of intervention design or implementation. Emulation will require continuing investigation into the curatorially-significant properties of content and format genres, and in particular, the manner in which these genres rely on sophisticated input and output modalities and user behavioral expectations. Comprehensive support for these format features may prove more difficult to provide in emulation environments as opposed to more prosaic numerical processing functions. Migration strategies will continue to rely on the ongoing development of preservation formats with increasing capacity to represent rich sets of content that can be used as migration targets without loss of information content.

Conclusions

Considerations of format are important in most of the high-level functions of an archival system as defined by the OAIS (ISO 14721) reference model. The formats of SIP components need to be identified and validated as part of Ingest processing. The formats of those components, and any derivative AIP (Archival Information Package) components, need to be stored as part of the Data Management function to characterize those components in a curatorially-significant manner. This format information is used to respond appropriately to Access requests, which may encompass additional AIP-to-DIP (Dissemination Information Package) transformations. The stored representation information is also an important input to the risk assessment and intervention activities that are part of the Preservation Planning function.

Within the Archival Storage function, a digitally-encoded asset must minimally possess the following attributes in order to maintain its usability over time: *viability*, that is, it must be retrievable from its storage medium; and *fixity*, that is, its bits must be unchanged from their accepted form. In archival records management contexts, the asset must also possess *authenticity*, that is, it must be the identical information that was originally acquired, or derived through verifiable processes and exchanges of physical and intellectual custody. Taken together, these attributes constitute "bit level" preservation, which can be performed without any notion of what those bits are meant to represent. Bit level preservation, however, is generally not sufficient to provide usability, which is inextricably tied up with the meaning of the bits. Usability therefore requires two additional attributes: *interpretability*, the ability for the semantic meaning underlying the bits to be recovered; and *renderability*, the ability to represent that meaning in directly human-sensible, i.e. analog, form. (In essence, the distinction between the two is that interpretability provides usability in *theory*, while renderability provides usability in *practice*.) Strong format typing is the fundamental property that permits preservation efforts to move beyond the bit level and enable the interpretability and renderability that provide full usability of digitally-encoded content.

An increasingly rich set of preservation-friendly formats, tools that can process these formats, and best practices concerning the selection, creation, use, and preservation of these formats is evolving within the digital curation community. However, the ubiquity, variety, and sophistication of new

content forms are growing at perhaps an even greater pace. As much of this new content is expressed in new formats, or new applications of existing formats, it is incumbent upon curation practitioners to understand the central position that format plays in preservation and access environments. By taking affirmative steps to monitor the important developments concerning the evolution of formats, practitioners can incorporate the consequences of those developments into their local policies, practices, and systems.

References

Print

AES-X098B, *Administrative and structural metadata for audio objects*.

IEEE 754, 1985, *Standard for Binary Floating-Point Arithmetic*.

ISO 12234-2, 2001, *Electronic still-picture imaging – Removable memory – Part 2: TIFF/EP image data format*.

ISO 12639, 2003, *Graphic technology – Prepress digital data exchange – Tag image file format for image technology (TIFF/IT)*.

ISO 19005-1, 2005, *Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1)*.

ISO/IEC 646, 1991, *Information technology – ISO 7-bit coded character set for information interchang*e.

ISO/IEC 9945-1, 2003, *Information technology – Portable Operating System Interface (POSIX) – Part 1: Base Definitions*.

ISO/IEC 10918-1, 1994, *Information technology – Digital compression and coding of continuous-tone still images: Requirements and guidelines*.

ISO/IEC 15444-1, 2000, *Information technology – JPEG 2000 coding system – Part 1: Core coding system*.

ISO/IEC 26300, 2006, *Information technology – Open Document Format for Office Applications (OpenDocument) v1.0*.

ISO/IEC DIS 29500, 2007, *Information technology – Office Open XML file formats*.

Lesk, M., 1990, *Image Formats for Preservation and Access: A Report of the Technology Assessment Advisory Committee to the Commission on Preservation and Access*, Commission on Preservation and Access.

McKusick, M. K., W. M. Joy, S. J. Leffler, and R. S. Fabry, August 1984, "A Fast File System for UNIX", *Transactions on Computer Systems*, volume 2, number 3, pp. 181-197.

Microsoft Windows Multimedia Programmer's Reference, 1991, Microsoft, Redmond.

Mohlenrich, J., ed., 1993, *Preservation of electronic formats & electronic formats*

for preservation, Highsmith.

Multimedia Programming Interface and Data Specifications 1.0, August 1991, IBM, Microsoft.

Stanescu, A., January 2005, "Assessing the Durability of Formats in a Digital Preservation Environment: The INFORM Methodology", *OCLC Systems and Services*, volume 21, number 1, pp. 61-81.

The Unicode Standard, Version 5.0, 2007, Addison-Wesley, Boston.

Online

About the Digital Curation Centre, 31 May 2005, Digital Curation Centre, http://www.dcc.ac.uk/about/ [Accessed: 28 December 2006].

Automatic Exposure – Technical Metadata, 2006, Research Libraries Group, http://www.rlg.org/en/page.php?Page_ID=2681 [Accessed: 21 April 2007].

Bates, M., S. Manuel, S. Loddington, and C. Oppenheim, May 2006, *Digital Lifecyles and File Types: Final Report, JISC Digital Repositories Programme: Rights and Rewards in Blended Institutional Repositories Project*, Joint Information Systems Committee, http://rightsandrewards.lboro.ac.uk/files/resourcesmodule/@random43cbae8b0d0ad/1148047621_DigitalLifecyclesV2.pdf [Accessed: 19 May 2006].

Bennett, J., 1997, *A Framework of Data Types and Formats, and Issues Affecting the Long-Term Preservation of Digital Material*, Joint Information Systems Committee, http://www.ukoln.ac.uk/services/papers/bl/jisc-npo50/bennet.html [Accessed: 19 May 2006].

Brown, A., October 2005, *Automatic Format Detection Using PRONOM and DROID*, National Archives, http://www.nationalarchives.gov.uk/aboutapps/fileformat/pdf/automatic_format_identification.pdf [Accessed: 13 September 2006].

Brown, A., 4 January 2005, *PRONOM Information Model*, National Archives, http://www.nationalarchives.gov.uk/aboutapps/fileformat/pdf/pronom_4_info_model.pdf [Accessed: 8 August 2006].

Brown, A., 19 June 2003, *Selecting File Formats for Long-Term Preservation*, National Archives, http://www.nationalarchives.gov.uk/documents/selecting_file_formats.pdf [Accessed: 13 September 2006].

Christensen, S., July 2004, *Archival Data Format Requirements*, Royal Library, Denmark, http://netarkivet.dk/publikationer/Archival_format_requirements-2004.pdf [Accessed: 13 September 2006].

Clausen, L., May 2004, *Handling File Formats*, State and University Library, Denmark, http://netarchive.dk/publikationer/FileFormats-2004.pdf [Accessed: 15 April 2007].

CRiB: Conversion and Recommendation of Digital Object Formats, 16 February 2007, University of Minho, http://crib.dsi.uminho.pt/ [Accessed: 16 February 2007].

Curtis, J., 29 September 2006, *AONS Systems Documentation*, National Library of Australia, http://www.apsr.edu.au/ [Accessed: 20 December 2006].

Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group, May 2005, OCLC, Research Libraries Group, http://www.oclc.org/research/projects/pmwg/premis-final.pdf [Accessed: 18 March 2006].

DCCRegRepOverall, 2 September 2005, Digital Curation Centre, http://twiki.dcc.rl.ac.uk/bin/view/Main/DCCRegRepOverall/ [Accessed: 15 April 2007].

Diffuse, 15 May 2006, Digital Curation Centre, http://www.dcc.ack.uk/diffuse/ [Accessed: 28 December 2006].

Diffuse – Home Page, 29 December 2003, Diffuse Project, http://web.archive.org/web/20031229131742/http://www.diffuse.org/ [Accessed: 28 December 2006].

Digital Negative (DNG) Specification, February 2005, Adobe, http://www.adobe.com/products/dng/pdfs/dng_spec.pdf [Accessed: 14 March 2007].

EBU Technical Specification 3285, July 2001, *BWF – a format for audio data files in broadcasting*, http://www.ebu.ch/CMSimages/en/tec_doc_t3285_tcm6-10544.pdf [Accessed: 19 April 2007].

ECMA-376, *Office Open XML File Formats*, December 2006, http://www.ecma-international.org/publications/standards/Ecma-376.htm [Accessed: 16 April 2007].

Entlich, R. and E. Buckley, 15 October 2006, "Digging Up Bits of the Past: Hands-on With Obsolescence", *RLG DigiNews*, volume 10, number 5, http://www.rlg.org/en/page.php?Page_ID=20987 [Accessed: 16 November 2006].

ESRI Shapefile Technical Description, July 1998, ESRI,

http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf [Accessed: 15 April 2007].

Extensible Markup Language (XML) 1.0 (Third Edition), 4 February 2004, Word Wide Web Consortium, http://www.w3.org/TR/REC-xml [Accessed: 22 April 2007].

Fedora, 21 April 2007, Fedora Project, http://www.fedora.info/ [Accessed: 21 April 2007].

Ferreira, M., A. Baptista, and J. Ramalho, July 2006, "A Foundation for Automatic Digital Preservation", *Ariadne*, volume 48, http://www.ariadne.ac.uk/issue48/Ferreira-et-al [Accessed: 21 April 2007].

Fielding, R., J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee, June 1999, *Hypertext Transfer Protocol – HTTP/1.1*, RFC 2616, Internet Engineering Task Force, http://www.ietf.org/rfc/rfc2616.txt [Accessed: 14 March 2007].

FILExt: The File Extension Source, 19 April 2007, http://filext.com/ [Accessed: 19 April 2007].

The Fine Free File Command, http://www.darwinsys.com/file/ [Accessed: 15 March 2007].

Firefox – Rediscover the Web, 20 April 2007, Mozilla, http://www.mozilla.com/en-US/firefox/ [Accessed: 20 April 2007].

Folk, M., and V. Choi, 8 January 2004, *Scientific formats for geospatial data preservation: A study of suitability and performance*, National Center for Supercomputing Applications, National Archives and Records Administration, http://www.ncsa.uiuc.edu/NARA/Sci_fmts_and_geodata_HDF.pdf [Accessed: 13 September 2006].

Format Registry Main Page, 1 December 2006, National Geospatial Data Archive, http://ngda.library.ucsb.edu/format/index.php/Main_Page/ [Accessed: 15 April 2007].

Freed, N. and N. Borenstein, November 1996, *Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types*, RFC 2046, Internet Engineering Task Force, http://www.ietf.org/rfc/rfc2045.txt [Accessed: 2 January 2007].

Freed, N. and J. Klensin, December 2005, *Media Type Specifications and Registration Procedures*, RFC 4288, BCP 13, Internet Engineering Task Force, http://www.ietf.org/rfc/rfc4288.txt [Accessed: 2 January 2007].

Google code – Summer of Code – Application Information: JPEG 2000 Support for Firefox, 20 April 2007, Google, http://code.google.com/soc/mozilla/

appinfo.html?csaid=C7B9CCBBF96648B3 [Accessed: 20 April 2007].

Hamilton, E., 1 September 1992, *JPEG File Interchange Format* Version 1.02, C-Cube Microsystems, http://www.w3.org/Graphics/JPEG/jfif3.pdf [Accessed: 15 April 2007].

Heslop, H., S. Davis, and A. Wilson, 2002, *An Approach to the Preservation of Digital Records*, National Archives of Australia, http://www.naa.gov.au/recordkeeping/er/digital_preservation/Green_Paper.pdf [Accessed: 15 April 2007].

Heydegger, V., J. Neumann, J. Schnasse, and M. Thaller, 31 October 2006, *PLANETS: Basic design for the extensible characterization language*, http://lehre.hki.uni-koeln.de/planets/documents/deliverables/PlanetsPC2D1D2-end.pdf [Accessed: 23 August 2007].

Huc, C., et al., 11 May 2004, *Criteria for evaluating data formats in terms of their suitability for ensuring information long term preservation*, Version 5, Groupe Pérennisation des Informations Numériques, http://www.ssd.rl.ac.uk/ccsdsp2/mon04/long_term_preservation_criteria.doc [Accessed: 26 March 2007].

I3A Standards – Initiatives – DIG35, 21 April 2007, International Imaging Industry Association, http://www.i3a.org/i_dig35.html [Accessed: 21 April 2007].

ICC.1, *Image technology colour management – Architecture, profile format, and data structure*, October 2004, International Color Consortium, http://www.color.org/ICC1V42.pdf [Accessed: 15 April 2007].

InSpect project, 25 January 2007, Arts and Humanities Data Service, http://ahds.ac.uk/about/projects/inspect/index.htm [Accessed: 15 April 2007].

Investigating the Significant Properties of Electronic Content Over Time (INSPECT), 1 April 2007, Joint Information Systems Committee, http://www.jisc.ac.uk/whatwedo/programmes/programme_rep_pres/INSPECt.aspx [Accessed: 15 April 2007].

IPTC—NAA Information Interchange Model, Version 4, 1 July 1999, International Press Telecommunications Council, http://www.iptc.org/std/IIM/4.1/specification/IIMV4.1.pdf [Accessed: 15 April 2007].

ISO 14721, 2003, *Space data and information transfer systems – Open archival information system – Reference model*, http://public.ccsds.org/publications/archive/650x0b1.pdf [Accessed: 26 December 2004].

ISO/IEC 15948, 2003, *Information technology – Computer graphics and image processing – Portable Network Graphics (PNG): Functional specification*, http://www.w3.org/TR/PNG/ [Accessed: 21 April 2007].

JEITA CP-3451, *Exchangeable image file format for digital still cameras: Exif Version 2.2*, April 2002, Japan Electronics and Information Technology Industries Association, http://www.exif.org/Exif2-2.PDF [Accessed: 21 April 2007].

JHOVE – JSTOR/Harvard Object Validation Environment, 29 March 2006, Harvard University, http://hul.harvard.edu/jhove/ [Accessed: 15 April 2007].

Lawrence, G., W. Kehoe, O. Rieger, W. Walters, and A. Kenney, June 2000, *Risk Management of Digital Information: A File Format Investigation*, Council on Library and Information Resources, http://www.clir.org/pubs/reports/pub93/pub93.pdf [Accessed: 1 August 2006].

LeFurgy, W., 2003, "PDF/A: Developing a File Format for Long-Term Preservation", *RLG DigiNews*, volume 7, number 6, http://www.rlg.org/legacy/preserv/diginews/v7_n6_feature1.html [Accessed: 22 April 2007].

Littman, J., May 2006, "A Technical Approach and Distributed Model for Validation of Digital Objects", *D-Lib Magazine*, volume 12, number 5, http://www.dlib.org/dlib/may06/littman/05littman.html [Accessed June 5, 2006].

McGovern, N. Y., A. R. Kenney, R. Entlich, W. R. Kehoe, and E. Buckley, April 2004, "Virtual Remote Control: Building a Preservation Risk Management Toolbox for Web Resources", *D-Lib Magazine*, volume 10, number 4, http://www.dlib.org/dlib/april04/mcgovern/04mcgovern.html [Accessed July 25, 2006].

Metadata Extraction Tool, 16 July 2007, National Library of New Zealand, http://meta-extractor.sourceforge.net / [Accessed: 23 August 2007].

Metadata Encoding and Transmission Standard (METS) Official Web Site, 23 August 2007, Library of Congress, http://www.loc.gov/standards/mets/ [Accessed: 23 August 2007].

Metadata for Images in XML Standard (MIX), 7 March 2007, Library of Congress, http://www.loc.gov/standards/mix/ [Accessed: 19 April 2007].

METS Extenders: Metadata Encoding & Transmission Standard (METS) Official Web Site, 13 September 2007, Library of Congress, http://www.loc.gov/standards/mets/mets-extenders.html [Accessed: 19 April 2007].

MIME Media Types, 7 December 2006, Internet Corporation for Assigned Names

and Numbers, http://www.iana.org/assignments/media-types/ [Accessed: 28 December 2006].

NISO Z39.87, *Data Dictionary – Technical Metadata for Digital Still Images*, 2006, http://www.niso.org/standards/resources/Z39-87-2006.pdf&std_id=731 [Accessed: 19 April 2007].

Ockerbloom, J., 7 April 2005, *The Typed Object Model (TOM)*, University of Pennsylvania, http://tom.library.upenn.edu/ [Accessed: 15 April 2007].

OpenRAW: Digital Image Preservation Through Open Documentation, 2 April 2007, http://www.openraw.org/ [Accessed: 21 April 2007].

PDF Reference, Version 1.7, November 2006, Adobe, http://www.adobe.com/devnet/pdf/pdf_reference.html [Accessed: 16 April 2007].

Pontello, M., 26 March 2007,*TrID – File Identifier*, http://mark0.net/soft-trid-e.html [Accessed: 26 March 2006].

Preserving Our Digital Heritage: Plan for the National Digital Information Infrastructure Preservation Program, October 2002, Library of Congress, http://www.digitalpreservation.gov/repor/ndiipp_plan.pdf [Accessed: 14 March 2007].

PRONOM, 15 April 2007, National Archives, http://www.nationalarchives.gov.uk/PRONOM/ [Accessed: 15 April 2007].

QuickTime File Format, 2002, Apple, http://developer.apple.com/documentation/QuickTime/QTFF/qtff.pdf [Accessed: 19 April 2007].

Representation Information Registry Repository, 15 April 2007, Digital Curation Centre, http://registry.dcc.ac.uk/omar/ [Accessed: 15 April 2007].

Ritter, N. and M. Ruth, *GeoTIFF Format Specification*, 28 December 2000, http://remotesensing.org/geotiff/spec/geotiffhome.html [Accessed: 14 March 2007].

Rosenthal, D., T. Lipkis, T. Robertson, and S. Morabito, January 2005, "Transparent Format Migration of Preserved Web Content", *D-Lib Magazine*, volume 11, number 1, http://www.dlib.org/dlib/january05/rosenthal/01rosenthal.html [Accessed: 21 April 2007].

Rusbridge, C., February 2006, "Excuse Me. . . Some Digital Preservation Fallacies?" *Ariadne*, issue 46, http://www.ariadne.ac.uk/issue46/rusbridge/ [Accessed: 21 April 2007].

Shirkey, C., December 2005, "Conceptual Issues from Practical Tests", *D-Lib*

Magazine, volume 11, number 12, http://www.dlib.org/dlib/december05/shirky/12shirky.html [Accessed: 21 April 2007].

Significant Properties ITT, 26 March 2007, Joint Information Systems Committee, http://ww.jisc.ac.uk/fundingopportunities/funding_calls/2007/03/significant_properties_itt.aspx [Accessed: 15 April 2007].

The State of Digital Preservation: An International Perspective, July 2002, Council on Library and Information Resources, http://www.clir.org/pubs/reports/pub107/pub107.pdf [Accessed: 2 January 2007].

Survey and assessment of sources of information on file formats and software documentation: Final report, Representation and Rendering Project, University of Leeds, http://www.jisc.ac.uk/uploaded_documents/FileFormatsreport.pdf [Accessed: September 14, 2006].

Sustainability of Digital Formats: Planning for Library of Congress Collections, 20 October 2006, Library of Congress, http://www.digitalpreservation.gov/formats/ [Accessed: 28 December 2006].

Technical Specifications – Profiles and Schemas: National Newspaper Preservation Program, 31 August 2006, Library of Congress, http://www.loc.gov/ndnp/techspecs.html [Accessed: 16 April 2007].

TIFF Revision 6.0, 3 June 1992, Adobe, http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf [Accessed: 14 March 2007].

UMIACS ADAPT Project: An Approach to Digital Archiving and Preservation Technology, 6 September 2006, University of Maryland, http://www.umiacs.umd.edu/research/adapt/ [Accessed: 15 April 2007].

Uneson, M., 1 September 2005, "Tomorrow's File Endings: On Archiving Principles and Archiving Formats", *ScieCom Info*, volume 2, http://www.sciecom.org/sciecominfo/artiklar/uneson_05_2.pdf [Accessed: 2 January 2007].

Walker, M., *AONS II Technical Architecture*, 16 February 2007, National Library of Australia, http://wiki.nla.gov.au/download/attachments/7723/AONSII_TechnicalArchitecture.doc [Accessed: 21 April 2007].

Waters, D. and J. Garrett, eds., 1 May 1996, *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*, Commission on Preservation and Access, Research Libraries Group,

http://www.rlg.org/en/page.php?Page_ID=20442 [Accessed: 2 January 2007].

Welcome to DSpace, 21 April 2007, MIT, Hewlett-Packard, http://www.dspace.org/ [Accessed: 21 April 2007].

Welcome to FRED, 2 November 2004, University of Pennsylvania, http:/tom.upenn.edu/fred/ [Accessed: 15 April 2007].

Wheatley, P., September 2001, "Migration – a CAMiLEON discussion paper", *Ariadne*, volume 29, http://www.ariadne.ac.uk/issue29/camileon/ [Accessed: 24 August 2007].

Wotsit.org: The Programmer's Format and Data Resource, 19 April 2007, http://www.wotsit.org/ [Accessed: 19 April 2007].

XMP Specification, September 2006, Adobe, http://partners.adobe.com/public/developer/en/xmp/sdk/XMPspecification.pdf [Accessed: 15 April 2007].

ZIP File Format Specification, 11 April 2007, PKWARE, Inc., http://www.pkware.com/documents/casestudies/APPNOTE.TXT [Accessed: 24 August 2007]

Fora

Abrams, S., September 2005, "Digital Formats and Preservation", *International Conference on Preservation of Digital Objects*, Göttingen, http://rdd.sub.uni-goettingen.de/conferences/ipres05/download/Digital Formats And Preservation – Stephen Abrams.pdf [Accessed: 14 April 2007].

Abrams, S. and A. Stanescu, November 2006, "Global Digital Format Registry (GDFR): An Interim Status Report" *DLF Fall Forum*, Boston, http://www.diglib.org/forums/fall2006/presentations/Abrams-2006-11.pdf [Accessed: 9 November 2006].

Arms, C., and C. Fleischhauer, 2005, "Digital Formats: Factors for Sustainability, Function, and Quality", *IS&T Archiving 2005 Conference*, Washington, DC, http://memory.loc.gov/ammem/techdocs/digform/Formats_IST05_paper.pdf [Accessed June 5, 2006].

Aschenbrenner, A., July 2004, "File Format Features and Significant Properties", *International Conference on Preservation of Digital Objects*, Beijing, http://rd.sub.uni-goettingen.de/conferences/ipres04/aschenbrenner/aschenbrenner_pres1.ppt [Accessed: 17 April 2007].

Christensen, N., 2004, "Towards format repositories for web archives", *4th International Web Archiving Workshop*, http://netarchive.dk/publikationer/FormatRepositories-2004.pdf [Accessed: 20 June 2006].

Fanning, B., M. Warfel, S. Abrams, and S. Sullivan, August 2005, "PDF/A: The Development of a Digital Preservation Standard", *Society of American Archivists 69th Annual Meeting*, New Orleans, http://www.archivists.org/conference/neworleans2005/no2005prog-Session.asp?event=1433 [Accessed: 15 April 2007].

Folk, M., and B. Barkstrom, May 2003, "Attributes of File Formats for Long-Term Preservation of Scientific and Engineering Data in Digital Libraries", *JCDL '03*, Houston, http://www.ncsa.uiuc.edu/NARA/Sci_Formats_and_Archiving.doc [Accessed: 13 September 2006].

Formats, Evaluations, and Relationships, Library of Congress, 7 March 2007, http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml#rel [Accessed: 19 April 2007].

Geremew, M., S. Song, and J. JaJa, May 2006, "Using Scalable and Secure Web Technologies to Design a Global Digital Format Registry Prototype: Architecture, Implementation, and Testing", *IS&T Archiving 2006*, Ottawa, http://www.umiacs.umd.edu/research/adapt/focus/documents/Archiving06.pdf [Accessed: 15 April 2007].

Hedstrom, M. and C. Lee, May 2002, "Significant properties of digital objects: definitions, applications, implications", *Proceedings of the DLM-Forum 2002*, Barcelona, 218-227, http://ec.europa.eu/transparency/archival_policy/dlm_forum/doc/dlm-proceed2002.pdf [Accessed: 15 April 2007].

Huc, C., May 2004, *Methodology for data format evaluation for the long term preservation*, http://www.ssd.rl.ac.uk/ccsdsp2/mon04/methodology_for_format_evaluation.ppt [Accessed: 26 March 2007].

Hunter, J., and S. Choudhury, June 2004, "A Semi-Automated Digital Preservation System Based on Semantic Web Services", *Joint Conference on Digital Libraries*, Tucson, http://www.itee.uq.edu.au/~eresearch/papers/2004/jcdl2004.pdf [Accessed: 21 April 2007].

Kunze, J., November 2005, "Web Archiving Service (WAS)", *DLF Forum*, http://www.diglib.org/forums/fall2005/presentations/kunze-2005-11.pdf [Accessed: 21 April 2007].

Mellor, P., P. Wheatley, and D. Sergeant, September 2002, "Migration on Request, a Practical Technique for Preservation", *6th European Conference on Digital Libraries*, Rome, http://www.si.umich.edu/CAMILEON/reports/migreq.pdf [Accessed: 21 April 2007].

Payette, S., May 2006, "Formalizing Content Models", *Fedora Content Model Workshop*, Karlsruhe, http://www.fedora.info/presentations/cmodel-intro.ppt [Accessed: 4 March 2007]

Phelps, T. and P. Watry, September 2005, "A No-Compromises Architecture for Digital Document Preservation", *9th European Conference on Digital Libraries*, Vienna, http://multivalent.sourceforce.net/Research/Live.pdf [Accessed: 21 April 2007].

Terminology

Format
A class of digitally-encoded assets defined by a set of semantic, syntactic, and serialization encoding rules for converting from abstract information to tangible byte streams.

Information
The fundamental unit of exchangeable knowledge [ISO 14721].

Long term
A period of time long enough for concern about the impacts of changing technologies, including support for new media and data formats, and changing user communities, on the information being held in a repository, extending into the indefinite future. [ISO 14271]

Magic number
An internal signature occurring in the first few bytes (historically, the first 2 bytes of Unix-derived operating systems) of a formatted file.

MIME type
(Multipurpose Internet Mail Extension) A widely used bipartite format typing taxonomy originally developed in the context of rich-content-enabled Internet mail.

OAIS
Open Archival Information System, an organization of people and systems accepting responsibility to preserve information and make it available [ISO 14721].

Representation information
Information, i.e., descriptive, administrative, technical, and structural metadata, that helps to map content into more meaningful concepts [ISO 14721]. In many curation contexts, format is an important component of content representation information.

Representation network
The set of representation information that fully describes the meaning of a content object, recognizing the digitally-encoding representation information often needs its own representation information to be interpreted and rendered properly.

Signature, external
An external characteristic of a formatted file, typically its filename extension, that presumptively identifies its format.

Signature, internal
A sequence of internal byte values that unambiguously identifies the format of a file.

Appendices

Figure 1 illustrates the three-step transformation (abstract model to semantic properties, semantic properties to syntactic data units, and data units to serialized bytes) underlying raster still image formats.

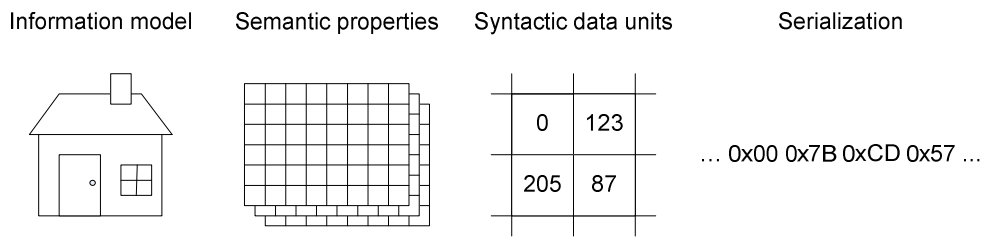

Figure 1. Raster image format in terms of three-stage format transformation

Figure 2 illustrates an example ingestion workflow incorporating automated validation and SIP packaging/unpackaging on both the producer or client-side and the archive or server-side.

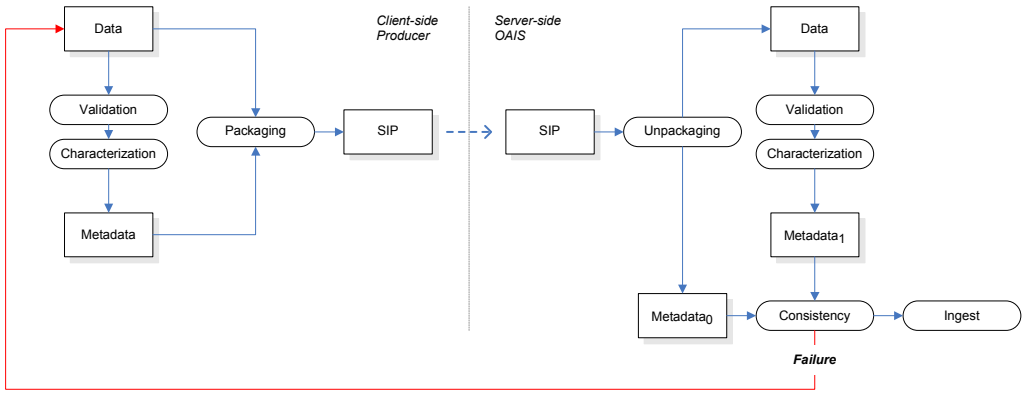

Figure 2. Ingest workflow

An annotated list of key external resources

Diffuse, http://www.dcc.ack.uk/diffuse/. EU-funded portal, now available under the auspices of the Digital Curation Centre.

Preserving Access to Digital Information (PADI), http://www.nla.gov.au/padi/. National Library of Australia (NLA) portal to international digital preservation resources, including information on formats, format standards, and format registries.

Sustainability of Digital Formats: Planning for Library of Congress Collections, http://www.digitalpreservation.org/formats/. Format portal for the Library of Congress's National Information Infrastructure Preservation Program (NDIIPP) initiative.