World Library and Information Congress:
69th IFLA General Conference and Council

1-9 August 2003, Berlin

Code Number: 128-E
Meeting: 165. Information Technology and Preservation and Conservation
Workshop
Simultaneous Interpretation: -

Towards a global digital format registry

Stephen L. Abrams
Harvard University
Cambridge, MA 02138, USA

David Seaman
Digital Library Federation
Washington, DC 20036, USA

Abstract

The concept of format permeates all technical areas of digital repository architecture and operation. Proper interpretation of otherwise opaque content streams is dependent upon knowledge of how typed content is represented. The international digital library and archival communities have expressed their need for a sustainable, global registry of digital representation formats for the purpose of fulfilling their mission. The current MIME Media Types registry does not provide sufficient granularity of format typing or sufficient standardized representation information about formats. We present a summary of an ongoing international effort to establish a new global format registry, which will maintain persistent, unambiguous bindings between public identifiers for digital representation formats and the significant syntactic and semantic properties of those formats.

Introduction

The concept of digital representation format permeates all technical areas of digital repository architecture and operation. Policy and processing decisions regarding ingest, storage, access, and preservation are frequently, if not uniformly, conditioned on a format-specific basis. Proper

interpretation of otherwise opaque content streams is dependent upon knowledge of how typed digital content is represented. For purposes of long-term preservation of digital objects, this knowledge of representation formats must be sustainable over archival time-spans. Additionally, effective interchange of digital objects between repositories and other consuming agents requires mutual agreement on format syntax and semantics. In order to facilitate the complementary goals of archival preservation and interoperability, what is needed is a sustainable public registry for the authority control of identifiers of digital representation formats. Such a registry will provide an unambiguous and persistent association between an identifier for a format and a set of important syntactic and semantic information about that format, which can be recovered now or in the future in order to facilitate the operation of digital repositories that make use of that format. A preliminary international effort is underway to investigate the technical, policy, and governance issues surrounding the development and operation of such a global digital format registry.

The only global mechanism for format typing in current widespread use is the MIME (Multipurpose Internet Mail Extensions) Media Types [5] registry operated by IANA (Internet Assigned Numbers Authority) [11]. However, for many digital repository operations MIME typing does not provide sufficient granularity to disambiguate important format distinctions, whether based on versioning or profiling. For example, both a tiled RGB TIFF image using LZW compression and a striped bi-tonal TIFF image with Group 4 compression are typed with the same MIME identifier: image/tiff. Similarly, the entire PDF family – PDF 1.0 through 1.4, PDF/X-1 through 3 (ISO 15930), and the proposed PDF/A standard [1] – are all typed with a single identifier: application/pdf. In both cases, the variant digital objects may undergo different parallel workflows dependent upon the specifics of their internal structure or semantics. To facilitate this, the proposed format registry will allow typing, and unambiguous identification, at arbitrary levels of granularity.

It is important that the registry is able to provide detailed authoritative representation information about formats. The MIME registry has varying requirements regarding the level of disclosure of the specific internal structure of MIME types. In particular, for registrations made outside of the IETF tree such details are “encouraged but not required” [6]. Additionally, publication of this technical information is accomplished through the RFC (Request for Comments) process, which relies on discursive text meant for human consumption. The proposed format registry seeks both to develop an appropriate trust mechanism to encourage the deposit of detailed representation information about proprietary formats, and to make such information available in standardized human and machine-readable forms, using controlled vocabularies to the fullest extent possible.

Ad-Hoc Working Group

During the summer of 2002, discussions between team members of the Harvard Library Digital Initiative (LDI) [9] and MIT DSpace [19] projects led to a realization that the existence of a format registry was a shared concern of the wider digital library community, and indeed, anyone operating a digital repository. With initial funding from the Digital Library Federation (DLF), two invitational workshops were organized to explore the potential for establishing a global digital format registry [8]. The ad-hoc working group was selected with an eye towards international participation and draws members from national libraries and archives, academic research libraries, and other library and archive-related organizations, including:

- Bibliothèque nationale de France
- California Digital Library
- Digital Library Federation (DLF)
- Harvard University
- Internet Engineering Task Force (IETF)
- Joint Information Systems Committee (JISC), UK
- JSTOR
- Library of Congress
- Massachusetts Institute of Technology
- National Archives and Records Administration (NARA)
- National Archives of Canada
- New York University
- National Institute of Standards and Technology (NIST)
- Online Computer Library Center (OCLC)
- Public Records Office (PRO), UK
- Research Libraries Group (RLG)
- Stanford University
- University of Pennsylvania

Participation in the working group was deliberately kept small (although with some difficulty, as interest in participation was widespread) in order to facilitate the exploratory nature of the early meetings. Once initial group consensus is reached on questions regarding data and service models, governance structure, and business issues, the process will be opened to all interested stakeholders for wider community review, comment, and refinement.

Use Cases for the Registry

The working group has collected a series of potential use cases for the digital format registry, which fall into the following broad categories:

- Identification – "I have a digital object; what format is it?"
- Validation – "I have an object purportedly of format *F*; is it?"
- Transformation – "I have an object of format *F*, but need *G*; how can I produce it?"
- Characterization – "I have an object of format *F*; what are its significant properties?"
- Risk assessment – "I have an object of format *F*; is it at risk of obsolescence?"
- Delivery – "I have an object of format *F*; how can I render it?"

With respect to the Open Archival Information System (OAIS) reference model, now formalized as ISO 14721:2002 [13], these format dependencies exist in the Ingest, Access, and Preservation Planning functions. (See Figure 1.) Relevant Ingest function sub-tasks include Submission Information Package (SIP) validation, and SIP-to-AIP (Archival Information Package) transformation. Access sub-tasks include AIP-to-DIP (Dissemination Information Package) transformation, and the extraction of technical metadata from the AIP for inclusion in the DIP. Preservation Planning sub-tasks include format monitoring for incipient obsolescence, and

defining and carrying out preservation strategies. Format representation information is necessary for preservation regardless of strategy, whether migration (AIP-to-SIP transformation) [4], emulation (deploying new delivery mechanisms for an existing DIP) [7], or the Universal Virtual Computer (UVC) approach [18].

Mission Statement

The working group has provisionally endorsed the following mission statement for the registry: "The registry will maintain persistent, unambiguous bindings between public identifiers for digital formats and representation information for those formats." A format is defined expansively as a fixed, byte-serialized encoding of an information model, which in OAIS terms is a formal expression of exchangeable knowledge. Representation information is also an OAIS concept and refers in this case to the mapping of typed formats into more meaningful concepts by capturing the significant syntactic and semantic properties of those formats. Significant properties are defined as those aspects of a format that are the primary carriers of the format's intellectual value.

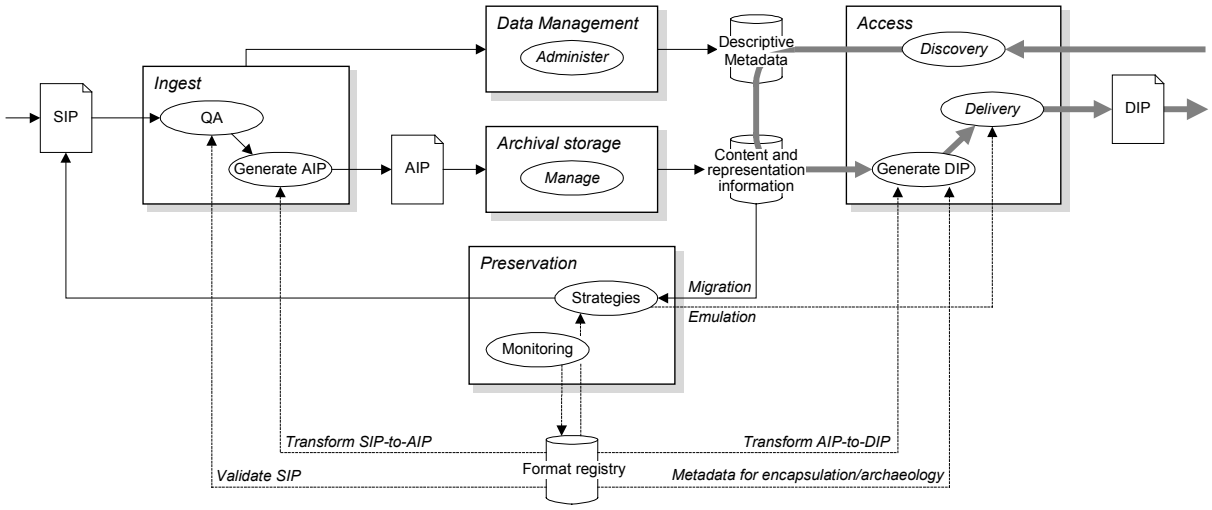

Figure 1. Format dependencies in an OAIS-compliant repository

Data Model

Development of the data model for the format registry is informed by similar investigative work previously performed with regard to format-related preservation metadata. ISO 14721:2002 defines the concept of representation information containing both structural (or syntactic) and semantic levels. The OCLC/RLG whitepaper on preservation metadata [23] suggests concrete elements of information necessary to properly interpret digital objects, drawn from a review of preservation projects undertaken by CEDARS (CURL Exemplars in Digital Archives), NEDLIB (Networked European Deposit Library), NLA (National Library of Australia), OCLC, and RLG. The European Commission's Information Society Technologies (IST) Programme has funded the development of the Diffuse project, which created a web site [12] providing reference and guidance information on format standards and specifications. (Unfortunately, maintenance of the web site ended in January 2003.) JISC has recently started a format representation project

[16] with many of the same goals as the proposed format registry. The TOM (Typed Object Model) project [22] provides a formal syntax for capturing the underlying grammar of a format, as well as a brokerage system to discovery services for manipulating formatted objects. The PRONOM system [24] developed by PRO captures information about tools used for generating, manipulating, and rendering objects on a per-format basis. Suggestions for administrative properties useful in any registry are provided by ISO/IEC 11179 [14] and OASIS/ebXML [21].

A number of other projects have concentrated on capturing various technical characteristics of formatted instance objects, rather than those of the formats themselves. Regardless, the information modeling of these projects may still suggest useful data elements relevant to the format registry. The NIST National Software Reference Library (NSRL) Reference Data Set (RDS) [20] provides file-level profiling of the distribution packages for popular commercial and non-commercial software, including vendor and product information. Media feature tags [10] can be used to define format-specific characteristics of content streams for client/server content negotiation. The Bitstream Syntax Description Language (BSDL), an XML-based schema under development as part of the MPEG-21 content adaptation mechanism [2], defines a formal syntax that may be useful for capturing the underlying grammar of a format.

The provisional data model for the registry includes elements for the mainly administrative properties of the registry itself as well as the various properties of the individual registered formats. These include data elements in the following four categories:

1. General descriptive properties, including canonical and alias identifiers for formats
2. Characterization properties, detailing the syntactic and semantic properties for formats
3. Processing properties, describing systems and services for which registered formats are inputs or outputs
4. Administrative properties, capturing important events in a registration's provenance

Table 1 lists the high-level format properties included in the current working data model ('*' indicates a cardinality of 0 or more; '+' indicates 1 or more):

Format		
Name	*Type*	*Function*
Identifier	URI	Primary, or canonical identifier
Alias *	URI	Variant identifier
Author *	Agent	Author
Owner +	Authority	Owner
Maintenance *	Authority	Maintenance agency
Classification *	Class	Ontological classification
Relationship *	FormatRelation	Arbitrary typed relationship
Specification *	Document	Specification document
Signature *	Signature	Internal or external signature
Tool *	System	Process or service
Status	Enumeration	"Active", "Withdrawn", "Unknown"
Provenance +	Event	Provenance event
Note *	UTF-8	Informative note

Table 1. High-level format properties

A format can have multiple URI-based identifiers (the specific syntax of which has yet to be determined); however, one must be unique and declared the canonical identifier for the format. A format may have one or more authors, each of which can be either a personal or corporate agent. Format owners and maintenance agencies are agents associated with a specific, though possibly unbounded time-span.

All formats in the registry are given an ontological classification. The two top-level categories are Content Stream, for formats that can be considered usefully as content streams independent of the physical medium underlying their manifestations, and Physical Media, for content streams manifest in tangible form on some physical memory structure. The Content Stream category subdivides on the basis of gross media type: Logical, Numeric, Text, Image, Audio, and Application (i.e., arbitrary binary data), while Physical Media subdivides on the basis of storage technology: Magnetic, Optical, and Paper. The definition of the more granular levels of the ontology remains an ongoing process.

Arbitrary typed relationships can be established between formats in the registry, including previous and subsequent version, dependency (e.g., a spreadsheet macro format might have an operational dependency on the worksheet format), and sub-typing with inheritance and a strict requirement of functional substitutability of the sub-type for its parent. The specification information for a sub-type needs only to document the deviation of the sub-type from its parent. Relationships can be established to formats in external registries, enabling a distributed architecture where a central registry could maintain formats of broad global applicability, while more obscure formats or local format profiles can be stored in local institutional, regional, or consortial registries.

Multiple specification documents can be associated with a format. These are qualified by author, title, publisher, date, public or standard identifier (e.g., DOI, ISBN, RFC, URI), canonicity (e.g., authoritative vs. informative), and accessibility. It is the intent of the registry to include actionable links to external documents, as well as maintaining soft and hard copies of the documents within the registry itself. Various levels of access will be provided to these materials according to deposit-time agreements with the copyright holders, including public access, on-site only, licensed access, and escrow. All restricted access regimes will be tied to specific trigger events (e.g., moving wall, corporate dissolution) that will make the specification information publicly available when appropriate.

Signature refers to some identifying characteristic of a format, either external (e.g., customary file extension, Mac OS data fork type), or internal (e.g., magic number). Format-specific software products are further qualified by function and vendor contact information. All provenance events, such as initial registration, update, and delete, and further qualified by timestamp, agent, and an explanatory note.

The format properties maintained in the registry are intended to be factual, not evaluative. Including overtly subjective information can raise issues of liability, and may tend to discourage the deposit of proprietary information by format owners. Regardless, the working group would like to explore mechanisms to make available informative case studies and best practice

guidelines insofar as this will not hamper the registry's primary goal, the collection and maintenance of authoritative format representation information.

Service Model

The working group has identified a set of core registry services in two broad categories, Management Services:

- Approval – Providing an appropriate level of technical review of registration information
- Maintenance – Creation, update, and deletion of format entries
- Notification – Subscription-based notification of significant events regarding formats
- Introspection – Machine-discoverable publication of local registry policies and practices

and Access Services:

- Description – Query mechanism for format representation information
- Export – Bulk export of registry data

A further set of ancillary services has been defined, but for the time being their implementation is being left to external value-added service providers:

- Rendering – Format-appropriate delivery of a formatted digital object
- Transformation – Conversion of an object from its target to source format
- Metadata extraction – Metadata characterization of a formatted object

The development of the registry service model is informed by ANSI X3.285 [3], the OASIS / ebXML Registry service model, and the IANA MIME media type and media feature tags registries.

Governance Structure

The digital format registry will be judged a success insofar as it is perceived as being trustworthy and is sustainable. The ultimate governance structure for the registry must facilitate these two goals. Without trust as to the authoritativeness of the representation information maintained within it, the registry will not be used by operators of digital repositories. Without trust as to the handling of proprietary representation information, such information will not be deposited within the repository by format owners.

Sustainability of the registry is essential in order to support the long-term preservation of digital assets. The OAIS reference model introduces the concept of community knowledge, the assumed knowledge base of some designated community. One of the reasons why it is often difficult to convince people of the necessity for a registry is the assumption of current knowledge about digital formats. Since today's operational repositories are gracefully handling a variety of formatted material it is difficult to imagine that the necessary community knowledge can be easily lost with the passage of time. The format registry will function as the persistent memory

of the digital community to ensure that the format knowledge we often take for granted today will be available to the digital community of the future.

It is not clear to the working group whether the governance of the registry necessitates the creation of new organization, or whether the registry can be appropriately administered under the umbrella of some existing organization. In part, this may be determined by the operational nature of the registry. If the registry is merely a public bulletin board, relying upon the larger community to populate it, then its administrative structure can be quite lean. If, on the other hand, the registry will be carefully managed, with a pro-active staff to harvest representation information, then the administrative structure must be concomitantly more complex and appropriate to deal with policy questions of acquisition strategy, public disclosure, and technical review.

Business Issues

The primary task of the registry business model is to generate a predictable yearly revenue stream with which to fund the ongoing operation of the registry. Unlike traditional archiving, digital archives cannot afford service gaps due to insufficient operating budgets; without constant active management digital materials are susceptible to irretrievable loss. The essential difficulty in funding any preservation activity, however, is how to generate income today for a benefit that may not be accrued until tomorrow.

The working group will consider the appropriateness of various business models – subsidy, subscription, submission fee, query fee, value-added services – but no firm decision will be made until further discussions within a larger community-wide consensual process.

Conclusion

The long-term preservation of digital assets requires a sustainable mechanism to maintain detailed authoritative representation information about the formats of those assets. The proposed global digital format registry will allow typing of digital objects at an appropriate level of granularity, and will permit the future recovery of important syntactic and semantic representation information associated with typed digital objects. Thus, the registry should properly be seen as an essential enabling technology that will support effective digital repository operations and archival preservation activities.

The ad-hoc working group recognizes that the development and implementation of a digital format registry will require the expertise and consensus of a much wider range of participants. The group intends to put into place a process that will encourage and welcome participation in the project from all appropriate stakeholders, including national, academic, and institutional libraries and archives; standards bodies; commercial interests such as regulated industries with statutory requirements regarding long-term record retention, software vendors as both developers and consumers of formatted information, and content providers; as well as others with an interest in the archival presentation of digital assets. The group will be exploring avenues for short-term funding to cover the continuing design and implementation phase for the registry. Two potential sources being investigated are the Library of Congress's National Digital Information Infrastructure Preservation Program (NDIIPP) [17] and JISC's tender for a Digital Curation

Centre [15]. The working group is confident that appropriate funding will be found to support continuing efforts to develop and deploy a global digital format registry for the benefit of the digital community now and into the future.

References

[1] AIIM, *PDF-Archive* (May 20, 2003) <http://www.aiim.org/standards.asp?ID=25013>.

[2] Amielh, Myriam and Sylvain Devillers, "Bitstream Syntax Description Language: Application of XML-Schema to Multimedia Content Adaptation," *WWW2002: The Eleventh International World Wide Web Conference*, May 7-11, 2002, Honolulu, Hawaii, USA <http://www2002.org/CDROM/alternate/334/>.

[3] ANSI X3.285, *Metamodel for the Management of Shareable Data* <http://metadata-stds.org/Document-library/Draft-standards/X3-285-Mgmt-of-Sharable-Data/X3-285.PDF>.

[4] Digital Preservation Testbed, *Migration: Context and Current Status*, December 2001 <http://www.digitaleduurzaamheid.nl/bibliotheek/docs/Migration.pdf>.

[5] Freed, N. and N. Borenstein, *Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types*, RFC 2046, November 1996 <http://www.ietf.org/rfc/rfc2046.txt>.

[6] Freed, N., J. Klensin, and J. Postel, *Multipurpose Internet Mail Extensions (MIME) Part Four: Registration Proceduress*, RFC 2048, BCP 13, November 1996 <http://www.ietf.org/frc/rfc2048.txt>.

[7] Granger, Stewart, "Emulation as a Digital Preservation Strategy," *D-Lib Magazine* 6:10 (October 2000) <http://www.dlib.org/dlib/october00/granger/10granger.html>.

[8] Harvard University Library, *Global Registry for Digital Format Representation Information* (April 9, 2003) <http://hul.harvard.edu/formatregistry/>.

[9] Harvard University Library, *Library Digital Initiative* (January 2003) <http://hul.harvard.edu/ldi/>.

[10] Holtman, K., A. Mutz, and T. Hardie., *Media Feature Tag Registration Procedure*, RFC 2506, BCP 31, March 1999 <http://www.ietf.org/rfc/rfc2506.txt>. See also RFCs 2533 and 2534.

[11] IANA, *MIME Media Types* (January 2, 2002) <http://www.iana.org/assignments/media-types/>.

[12] Information Society Technologies, *Diffuse Standards and Specifications List* (December 2002) <http://www.diffuse.org/standards.html>.

[13] ISO 14721:2003, *Space data and information transfer systems -- Open archival information system -- Reference model* <http://wwwclassic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf>.

[14] ISO/IEC 11179, *Information technology – Specification and standardization of data elements*.

[15] JISC, *Discussion Paper – Draft ITT for a Digital Curation Centre*, Version 2.2 <http://www.jisc.ac.uk/uploaded_documents/digitalcurationcentrev3.pdf>.

[16] JISC, *The File Format Representation and Rendering Project* (December 12, 2002) <http://www.jisc.ac.uk/dner/preservation/fileformatting.html>.

[17] Library of Congress, *Welcome to the NDIIPP* (February 26, 2003) <http://www.digitalpreservation.gov/ndiipp/>.

[18] Lorie, Raymond A., "A Methodology and System for Preserving Digital Data," *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries*, July 13-17, 2002, Portland, Oregon, USA (New York: ACM Press, 2002), pp. 312-319.
[19] MIT Libraries, *DSpace: Durable Digital Depository* (May 20, 2003) <http://www.dspace.org/>.
[20] NIST, *Data Formats of the NSRL Reference Data Set (RDS) Distribution*, December 30, 2002 <http://www.nsrl.nist.gov/documents/Data-Formats-of-the-NSRL-Reference-Data-Set-12.pdf>.
[21] OASIS, *ebXML Registry TC* (May 20, 2002) < http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=regrep>
[22] Ockerbloom, John Mark, *Introduction to the Types Object Model (TOM)* (May 20, 2003) <http://tom.library.upenn.edu/intro.html>.
[23] OCLC/RLG Working Group on Preservation Metadata, *A Metadata Framework to Support the Preservation of Digital Objects*, June 2002 <http://www.oclc.org/research/pmwg/pm_framework.pdf>.
[24] Public Records Office, *PRONOM System User Guide*, November 27, 2002 <http://www.pro.gov.uk/about/preservation/digital/pronom.htm>.