# Establishing a Global Digital Format Registry

Stephen L. Abrams

Abstract

Detailed knowledge of the internal properties of digital representation formats is necessary to interpret properly the full information content of otherwise opaque digital objects. These properties form an important component of the representation information needed by repository workflows regardless of local preservation strategy and infrastructure decisions. The Digital Library Federation (DLF) has sponsored preliminary investigations toward establishing a Global Digital Format Registry (GDFR) that will function as a sustainable utility for maintaining the bindings between public identifiers for digital formats and the significant syntactic and semantic properties of those formats. A sustainable GDFR should prove to be of great utility to archives, libraries, digital repositories, and other organizations and individuals interested in the long-term viability of digital assets.

## Digital Formats

It has become commonplace for digital objects to be acceptable and valued assets under the collection development policies of many libraries, archives, museums, and other scientific and cultural heritage repositories with long-term preservation mandates. In general, a digital object can be considered as the encapsulation in digital form of some piece of abstract intellectual content. More specifically, a digital object is the aggregation of one or more formatted content streams representing the primary content of the object as well as associated descriptive, administrative, technical, and structural metadata. Without a thorough understanding of the format of those content streams, the ability to recover the original intellectual content from which those streams were derived is severely compromised,

LIBRARY TRENDS, Vol. 54, No. 1, Summer 2005 ("Digital Preservation: Finding Balance," edited by Deborah Woodyard-Robinson), pp. 125–143

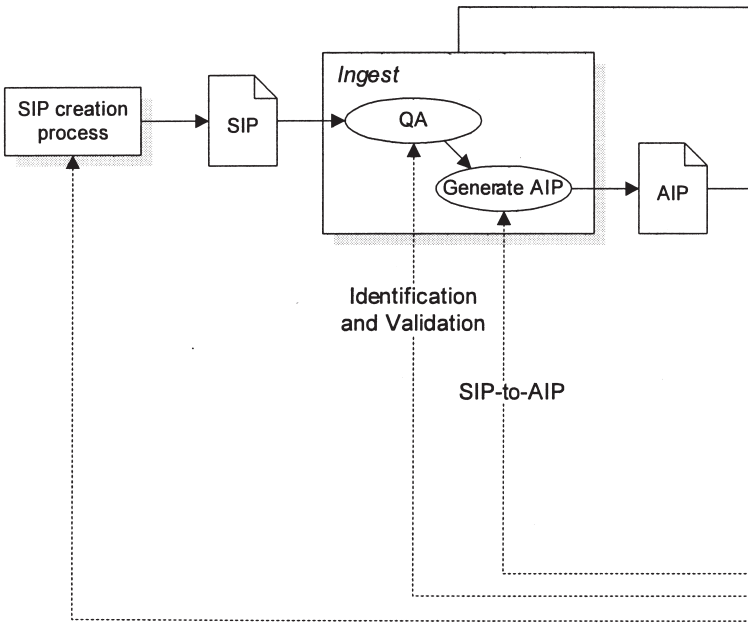
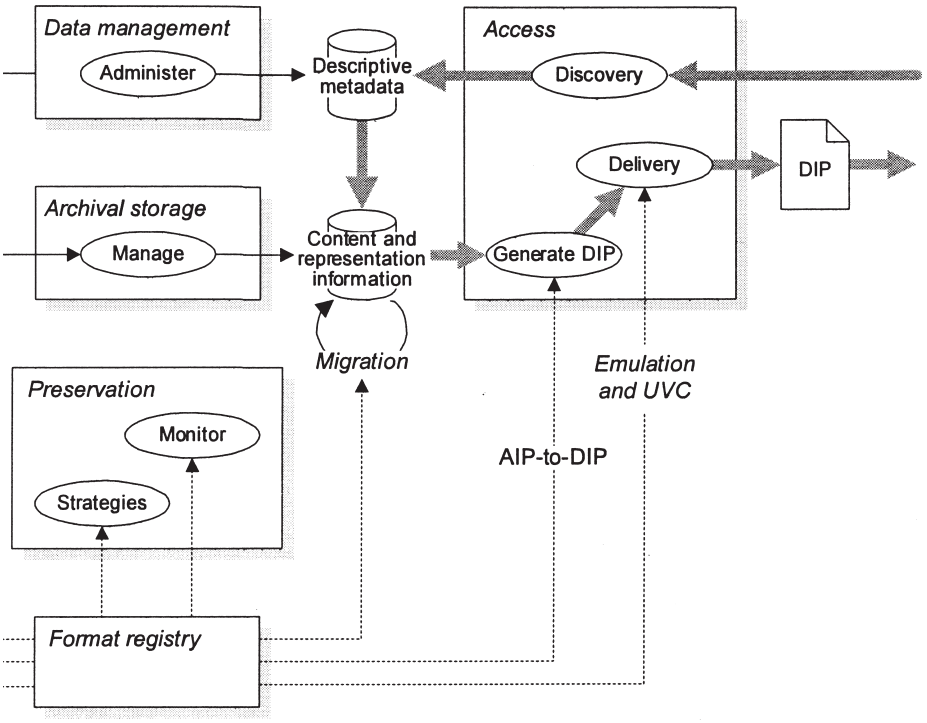



*Figure 1.* Repository Workflow Format Dependencies

if not made impossible. Furthermore, common agreement on the syntax and semantics associated with an object's formatted content streams is necessary for the effective interchange of that object, whether between institutions implementing different technological infrastructures or between the various processing steps applied to the object as it passes through its intra-institutional life cycle. In essence, a format is the property associated with a content stream that provides the typing information necessary for its proper interpretation.

More formally, a format is a reversible, byte-serialized encoding of an abstract information model, which is itself a formal expression of exchangeable knowledge (International Organization for Standardization, 2003). A format defines the syntactic and semantic rules for the mapping from an information model to a byte stream and the inverse mapping from that byte stream back to the original information model. Historically, discussions of formats have been couched in terms of "file formats." However, as there are many contexts, such as the network transport of formatted content streams or consideration of content streams at a level of granularity finer than that of an entire file, where specific reference to "file" is inappropriate, the more general term "digital formats" will be used in this article.




## Format Dependencies in Repository Operation

Digital repository operations can be distinguished into two broad categories: (1) those that are performed independent of the internal properties of its digital objects; and (2) those that are performed dependent upon the internal characteristics of the objects or, in other words, their format. With regard to the latter category, format dependencies exist in many, if not most, phases of repository operation. Figure 1 presents an idealized repository workflow based on the Open Archival Information System (OAIS) reference model (International Organization for Standardization, 2003). Although originally developed by the space science community, the OAIS model defines a general approach that is broadly applicable to repositories operating in nonscientific domains. It has been widely adopted as the conceptual framework for repository architecture and operation and has become part of the *lingua franca* within the digital preservation community.

### *Ingest Dependencies*

In OAIS terms, digital objects are delivered to an archive or repository in the form of a Submission Information Package (SIP), a conceptual data structure that encapsulates both primary content and representation infor-

mation about that content. Representation information is information that is necessary to map object content into more meaningful constructs relative to some designated community—in other words, metadata (Holdsworth & Sergeant, 2000). The specific format of an object content stream within a SIP is an important technical component of SIP metadata.

The OAIS Ingest function is responsible for Quality Assurance (QA) validation of SIP content. Some repositories may operate under local policies or statutory regimes that mandate an obligation to accept all SIPs regardless of validation status, while others may implement more stringent policies that reject SIPs that are not well formed or well characterized. Regardless, it is a reasonable repository best practice to validate incoming SIP content streams relative to the stated or inferred formats of those streams. Even for repositories that do not use validation status as an acceptance criterion, that status is nevertheless an important preservation metadata property that characterizes the state of a digital object at the point of ingest. Validation is performed with respect to the specific syntactic and semantic rules established by the format to which a content stream purportedly conforms. The Ingest function is the most effective point at which to detect and remediate errors occurring in archival materials (National Archives and Records Administration et al., 1999). Once digital objects are accepted into a repository, where they may not be accessed for significant periods of time, effective channels of communication with the original creators to ascertain their authorial intent with respect to those objects may become difficult, if not impossible.

The Ingest function is also responsible for disaggregating a SIP, passing the descriptive metadata to the archive Data Management function, and transforming the SIP into an Archival Information Package (AIP) encapsulating primary content and administrative and technical metadata. It is not necessary for object content streams within an AIP to have the same formats as the corresponding content streams in the SIP. In the interest of data homogeneity and its concomitant impact on operational efficiencies, many repositories may choose to define a restricted set of canonical AIP formats to which SIP content streams are transformed during the SIP-to-AIP conversion process. Quality assurance checks must be applied subsequent to all content stream transformations in order to ensure that none of the significant properties of the original content have been lost (Hedstrom & Lee, 2002). In addition to knowing the context in which the content will be accessed, the selection of appropriate tools for both the transformation and QA steps requires knowledge of the source and target formats.

*Discovery and Delivery Dependencies*

Object discovery and delivery are handled by the OAIS Access function. Object content and associated metadata are delivered in the form of a Dissemination Information Package (DIP), which is created from an AIP. As

in the Ingest SIP-to-AIP conversion, there is no requirement for content stream format to remain constant during the AIP-to-DIP conversion. Many repositories may choose to provide external access to archived content in a wider range of formats than are used internally to store that content. As with all format transformations, the selection of appropriate tools requires knowledge of the source and target formats.

*Migration Dependencies*

Additional format dependencies are introduced to repository operation by the choice of specific preservation strategies. A migration strategy entails the periodic transformation of object content streams from formats that are in danger of becoming obsolete to other formats with a longer period of viability (Wheatley, 2001). (See figure 2(a). The notation $C_F$ refers to a content stream of format *F;* $D_0$ represents a delivery service for $C_F$ extant at time $t_0$ that executes in the context of a contemporaneous computing platform $P_0$. Similarly, $C_G$ is a content stream of format *G* delivered by $D_1$ at time $t_1$ in the context of platform $P_1$, and so on.) As with the SIP-to-AIP and AIP-to-DIP transformations, a preservation migration requires an understanding of the source and target formats as well as appropriate tools that can perform the mapping. Since any transformation introduces the potential for irretrievable information loss, such tools and processes must be carefully selected and configured to mitigate against any possible loss.

Note that a required transformation path may be indirect. Based upon the specific formats supported as inputs and outputs of available tools, a migration from format *F* to *H* may involve multiple intermediate steps. (See figure 3. The notation $T_n$ represents the process or service implementing transformation step *n.*) In such cases, potential processing paths must be evaluated carefully, as even seemingly insignificant data loss can multiply rapidly.

*Emulation Dependencies*

Whereas a migration-based preservation strategy manipulates a content stream as necessary to allow it to interoperate with a current delivery system, an emulation-based approach maintains the data integrity, or fixity, of the content stream as originally deposited. Emulation then requires a delivery system that both supports the original format and executes in the context of the computing platform current at the time of access (Digital Preservation Testbed, 2003). This system is provided either by implementing a new system that mimics the behavior of the original delivery system or by developing an interface layer that sits between a copy of the original delivery system and the current computing platform. (See figures 2(b) and (c). The notation $D_1$ represents a delivery service created to mimic the behavior of $D_0$ but execute in the context of platform $P_1$. $E_1$ represents the emulation interface between the delivery system $D_0$ extant at time $t_0$ and the

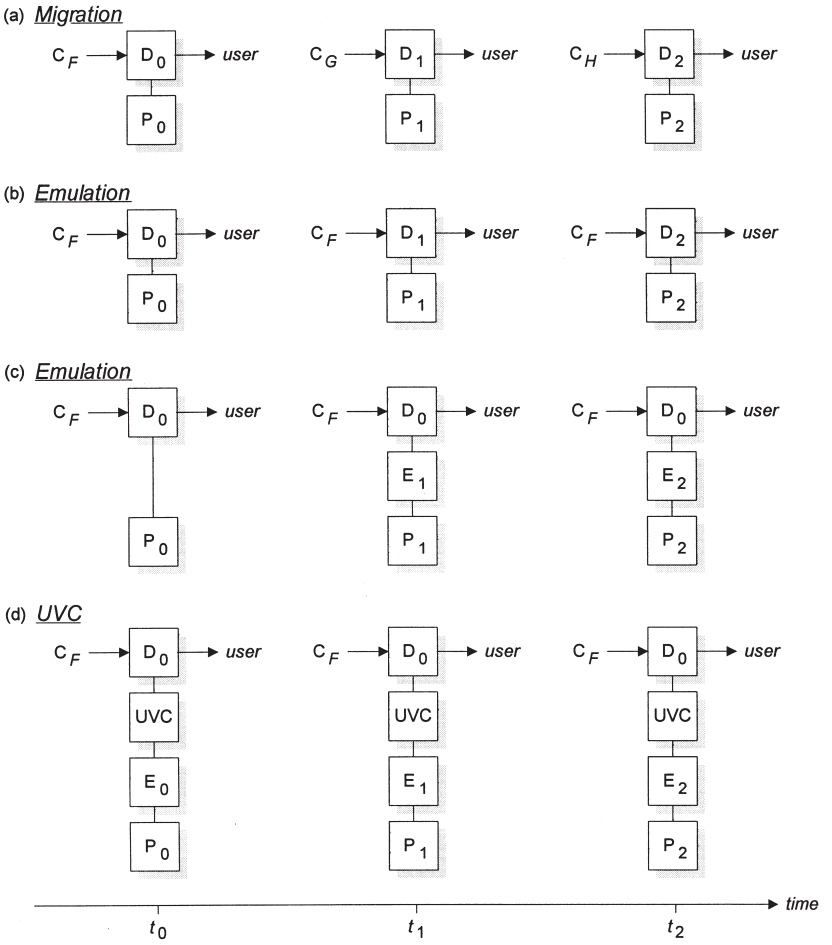



*Figure 2.* Preservation Strategies

computing platform $P_1$ extant at time $t_1$. This interface layer provides the appearance of the context of $P_0$ to $D_0$.) Implementation of a new delivery system requires knowledge of the content stream format; implementation of an emulation interface requires knowledge of the delivery system that supports that format.

*Universal Virtual Computer Dependencies*

The Universal Virtual Computer (UVC) approach is a variant of emulation (Lorie, 2002). Under this approach a delivery system for a given format that executes in the context of a UVC is implemented once. The UVC is a software construct rather than a physical processor. Like traditional emu-

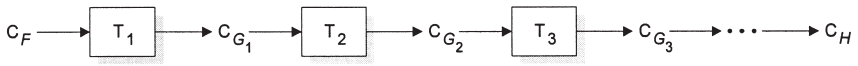



*Figure 3.* Multistep Migration

lation, the UVC itself requires an emulation interface to the underlying computing platform at the time of content stream access. Unlike traditional emulation, however, the emulation interface does not have to be concerned with the specific requirements and behaviors of the delivery system but rather only with the general capabilities of the UVC. [See figure 2(d). The notation $E_0$ represents the UVC interface to the underlying computing platform $P_0$ extant at time $t_0$, $E_1$ is the interface to $P_1$ at time $t_1$, and so forth.] However, the implementation of the format-specific delivery system does require knowledge of the internal syntax and semantics of that format.

## Format Registries

The collection of comprehensive and authoritative representation information for digital formats requires extensive and specialized knowledge. While most digital repositories will need the same types of information, it is unlikely that they will all have the technical resources available to acquire that information locally. The existence of a public registry responsible for the centrally organized maintenance and distribution of format-specific representation information provides an effective mechanism to share scarce technical expertise within the wider digital preservation community.

A format registry is a repository for format representation information or, in other words, descriptive, administrative, and technical metadata about digital formats, including the definition of the syntactic and semantic characteristics of the registered formats. This metadata defines the significant properties of digital formats with regard to the long-term preservation of digital objects. A format registry should provide sufficient information to respond to the following use cases common to digital preservation repositories:

- Identification: "I have a content stream; what format is it?"
- Validation: "I have a content stream that purports to be of format *F;* is it?"
- Characterization: "I have a formatted content stream of format *F;* what are its significant properties?"
- Processing: "I have a formatted content stream; how can I transform (or edit, sample, compress, de-skew, render, etc.) it?"
- Risk assessment: "I have a formatted content stream; is it at risk of obsolescence?"

Descriptions of many digital formats are currently available, at varying degrees of detail and accuracy, through a variety of channels including Web

sites, informal reference books, and formal specification documents. Many of these sources, however, are of a transitory nature. For example, the European Commission's Information Society Technologies (IST) Programme funded the Diffuse project, which operated a high-quality Web site providing extensive information on digital formats and pointers to specification documents (Diffuse Project, 2003). Unfortunately, project funding ended in 2003 and the Web site is no longer available at its previous address. (A snapshot of the Web site can be retrieved from the Internet Archive's Way-Back Machine.) Long-term digital preservation requires that authoritative information concerning digital formats be available indefinitely.

Perhaps the most well-known example of a format registry is the Internet Assigned Names Authority (IANA) MIME type registry (Freed, Klensin, & Postel, 1996). However, MIME registrations are maintained and provided as text documents intended for human consumption, precluding the effective use of automated interactions between the registry and repositories. Furthermore, the MIME registry does not prescribe any specific set of format attributes that must be disclosed, and under some circumstances no technical disclosure of any kind is required. MIME types are also defined at a fairly coarse granularity that makes no provision for families of related formats existing under a common rubric. For example, TIFF/IT (ISO 12639, used for pre-press data exchange), TIFF/EP (ISO 12234-2, output by many digital cameras), and GeoTIFF (used for geo-referenced images) are all variants of the Tagged Image File Format but may require very different preservation processing workflows. Yet all three are identified by the same MIME type, "image/tiff." These conditions render the MIME registry an insufficient resource by itself for digital preservation activities.

A more recent example of a format registry that resolves many of the problems raised by the IANA MIME registry is the UK National Archive's PRONOM system (National Archives of England, Wales, and the United Kingdom, 2005). In its current version, PRONOM stores detailed technical information about various software applications that can be retrieved on the basis of application name, vendor, and supported format. A number of enhancements are planned for PRONOM, including a substantial increase in the amount of information stored about formats themselves, automatic generation of migration paths, and a technology watch service that monitors product support life cycles. Within PRONOM variant formats are identifiable by version and specific profile. Given the nature of the National Archive's mandate, continued support for PRONOM can be assumed with high confidence.

It appears likely that many similar format registries may be developed or at least deployed at institutions around the world. This could result in an undesirable fragmentation of important format representation information that would unnecessarily complicate the process of discovery of relevant

data. To mitigate against this situation, some form of centralized coordination is needed. This coordinating role is a major component of ongoing work toward establishing a Global Digital Format Registry (GDFR).

## Global Digital Format Registry

In recognition of the importance of a format registry as a resource for digital preservation, the Digital Library Federation (DLF) organized a pair of invitational workshops in 2002 to investigate the issues surrounding the development and deployment of a GDFR. The participants in these workshops included representatives from major national, research, and academic libraries and archives; standards organizations; and other institutions involved in digital preservation activities (see table 1). Harvard University is now seeking funding from the Library of Congress under its National Digital Information Infrastructure Preservation Policy (NDIIPP) (Library of Congress, 2002) initiative for a multiyear, two-track project to continue the DLF-sponsored work. The parallel tracks will focus on technical and governance/business model issues respectively. The project makes explicit provision for continued international outreach and consultation in order to reach the widest possible consensus on the GDFR from interested stakeholders in the digital library, archive, and preservation communities. Project deliverables include well-documented data and services models, a complete specification for the inter-nodal communication protocol, and a reference implementation of a GDFR cache. The project plan also envisions a significant period of production operation during which the network protocol will be exercised and integration of the GDFR with repository work flows will be tested.

*Table 1.* Participants in the DLF-Sponsored GDFR Workshops

| |
|---|
| 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 (formerly Public Records Office), UK |
| National Archives and Records Administration (NARA), U.S. |
| National Archives of Canada |
| New York University |
| National Institute of Standards and Technology (NIST), U.S. |
| Online Computer Library Center (OCLC) |
| Research Libraries Group (RLG) |
| Stanford University |
| University of Pennsylvania |

Initially, the GDFR was conceived of as a single centralized repository of format representation information. However, in view of recent developments such as PRONOM and forthcoming work in the area of format registries by the Joint Information Systems Committee (JISC) funded Digital Curation Centre (DCC) in the UK, it has become clear the some form of distributed network of cooperating registries is necessary. This architecture also provides the potential benefit of data redundancy, an important provision with regard to the preservation of the information collected in the various registries.

The scope of the GDFR is to "maintain persistent, unambiguous bindings between identifiers for digital formats and representation for those formats" (Abrams & Seaman, 2003). In other words, so long as a digital object content stream is correctly typed with a format known to the GDFR, the specific syntactic and semantic rules governing that format will be retrievable. As mentioned previously, the GDFR is conceived of as a distributed network of cooperating nodes or caches. Thus, the main work of the GDFR project is to define an abstract data model for format representation information that is used as the basis for communication between network nodes via the GDFR inter-nodal protocol. The specific implementation details of any particular node in this network are left undefined by the GDFR. Compliance with GDFR standards occurs at the level of the network protocol (see figure 4).

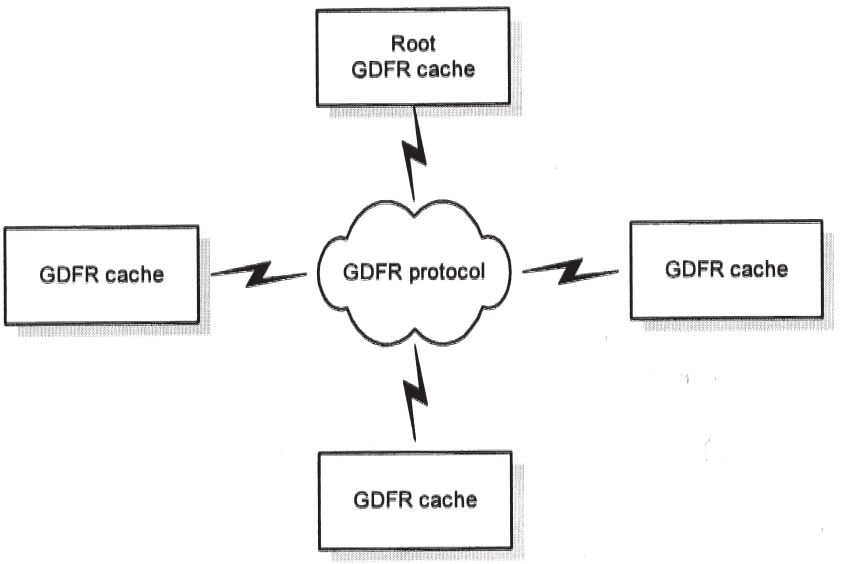



*Figure 4.* GDFR Distributed Architecture

*Data Model*

Development of the GDFR data model has been informed by earlier projects investigating issues regarding format-related preservation metadata. The OAIS reference model defines the concept of representation information containing structural, syntactic, and semantic levels. The Online Computer Library Center/Research Libraries Group (OCLC/RLG) white paper on preservation metadata (2002) suggests specific information elements necessary to interpret digital objects drawn from a review of preservation projects undertaken by CEDARS (CURL Exemplars in Digital Archives), NEDLIB (Networked European Deposit Library), National Library of Australia (NLA), OCLC, and RLG. The UK JISC File Format Representation project investigated many of the issues concerning the collection and maintenance of format representation information (JISC, 2002). Suggestions for administrative properties useful in any registry are provided by the ISO/IEC 11179 standard (International Organization for Standardization, 2004) and the OASIS/ebXML information model (OASIS, 2003).

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 GDFR project. The National Institute of Standards and Technology (NIST) National Software Reference Library (NSRL) Reference Data Set (RDS) provides file-level profiling of the distribution packages for popular commercial and noncommercial software, including vendor and product information (National Institute of Standards and Technology, 2002). Media feature tags can be used to define format-specific characteristics of content streams for client/server content negotiation (Holtman, Mutz, & Hardie, 1999). The Bitstream Syntax Description Language (BSDL), an XML-based schema under development as part of the MPEG-21 content adaptation mechanism, defines a formal syntax that may be useful for capturing the underlying grammar of a format (Amielh & Devillers, 2002).

The provisional data model for the GDFR includes elements for the administrative properties of the registry itself as well as the various properties of the individual registered formats, which fall into four main 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 2.* GDFR High-Level Properties

| Property Name | Type | Description |
|---|---|---|
| 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 | Relation | Arbitrary typed relationship |
| Specification | Document | Specification document |
| Disclosure | Enumeration | Level of disclosure |
| Signature | Signature | Internal or external signature |
| System | Product | Tool, system, or service |
| Status | Enumeration | Format status |
| Provenance | Event | Registration provenance event |
| Review | Enumeration | Level of technical review |
| Note | UTF-8 | Informative note |

Table 2 lists some of the high-level format properties included in the current provisional data model.

A format can have multiple identifiers, which may be based on entirely separate naming schemes; however, one must be unique within the GDFR and declared as 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 specific, though possibly unbounded, time spans. All formats in the registry are given an ontological classification. The two top-level ontological 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 (see tables 3 and 4). The Content Stream category subdivides on the basis of gross media type—Logical, Numeric, Text, Image, Audio, and Application (that is, 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 (for example, a spreadsheet macro format might have an operational dependency on the worksheet format), and subtyping with inheritance and a strict requirement of functional substitutability of the subtype for its parent (Liskov & Wing, 1994). Substitutability requires that a subtype be usable without loss of functionality in any context in which its parent type can be used. (For example, a PDF/X file can be used in any context that a generic PDF can be used but not vice versa. In other words, all PDF/X objects are PDF objects, but not all PDF objects are PDF/X objects; thus, PDF/X is a

*Table 3.* Sample Content Stream Classification

**Content Stream** [*byte-serialized encoding of abstract information model*]
- **Logical**
  - XDR Boolean (RFC 1832)
- **Numeric** [*data representing mathematical cardinality or ordinality*]
  - **Scalar**
    - **Integer**
      - XDR integer (RFC 1832)
      - **Unsigned integer**
        - XDR unsigned (RFC 1832)
    - **Real**
      - Floating point
        - IEEE 754
- **Text** [*directly interpretable character data*]
  - EBCDIC
  - ISO/IEC 646 (ASCII)
  - ISO/IEC 8859-1 (Latin 1)
  - Mac OS Roman
  - UTF-8
  - Windows code page 1252
  - **Structured text** [*text with structural constraints*]
    - CSV
    - Tab delimited
    - **Mark-up language** [*text with semantic tagging*]
      - HTML
      - LaTeX
      - RTF
      - SGML
- **Image**
  - **Still**
    - **Font** [*character glyph data*]
      - **Outline**
        - Adobe Type 1
        - OpenType
        - TrueType
    - **Graphic**
      - **Vector**
        - **2D**
          - SVG
        - **3D**
          - VRML
      - **Raster**
        - GIF
        - ISO/IEC 10918 (JPEG)
        - JFIF
        - TIFF
    - **Page description**
      - PDF
      - PostScript
      - QuarkXpress
  - **Motion**
    - AVI
    - MPEG
    - QuickTime

*Table 4.* Sample Physical Media Classification

**Physical Media** [*encoding to physical memory structure*]
- **Magnetic**
  - **Tape**
    - **Reel**
      - **9 track**
        - ANSI X3.54-1986
      - **Cartridge**
        - **3480 class**
          - ANSI X3.180-1990
        - **DLT**
          - ISO/IEC 15307
          - ISO/IEC 16382
- **Optical**
  - **Disk**
    - **CD-ROM**
      - ISO 9660

subtype of PDF.) The specification information for a subtype needs only to document the deviation of the subtype from its parent. Relationships can be established to formats in external registries, enabling a distributed architecture where a root registry node or cache 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 (for example, DOI, ISBN, RFC, URI), canonicity (for example, authoritative vs. informative), and accessibility. It is the intent of the GDFR to include actionable links to external documents, as well as maintain 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, ranging from public access to document escrow. All restricted access regimes will be tied to specific trigger events (for example, moving wall, corporate dissolution) that will make the specification information publicly available when appropriate.

The level of disclosure indicates the degree to which complete technical information about a format's syntax and semantics are made publicly available. Signatures are identifying characteristics of a format, either external (for example, customary file extension, Mac OS data type) or internal (for example, magic number). Format-specific software products, systems, and services are qualified by function and vendor contact information. Status indicates whether a format is still supported or has been deprecated or withdrawn by its owner. All provenance events, such as initial registration, update, and delete, are qualified by timestamp, agent, and an explanatory note. All information submitted to the GDFR is subject to technical review for accuracy, completeness, and authoritativeness.

In addition to these properties, the GDFR will investigate the use of for-

mat assessment characteristics. A starting point for this investigation is work being done at the Library of Congress that defines assessment categories dealing with objective sustainability factors applicable to formats independent of content genre and more subjective factors relative to genre-specific quality and functionality (Arms & Fleischhauer, 2003). The sustainability factors fall into six subcategories:

1. Disclosure: the degree to which comprehensive and authoritative technical specifications are publicly available
2. Adoption: the degree to which the format is in common use. Software support for a format is evidence of its adoption. Widespread use tends to impede the onset of obsolescence
3. Transparency: the degree to which the digital representation is open to direct analysis—human readability—with basic tools, such as a non-format-aware text editor. For example, compression inhibits transparency; character encodings are more transparent than binary encodings
4. Self-documentation: the degree to which objects encapsulate intellectual, administrative, and technical descriptions of themselves
5. External dependencies: the degree to which formatted objects depend upon hardware and/or software for rendering or use. For example, highly dynamic or interactive content may rely upon input modalities (for example, mouse, trackball, light-pen) assumed today but unavailable in the future
6. Technical protection mechanisms: the degree to which a format enforces restrictions on use to protect intellectual property rights

These assessment factors are useful for the selection of appropriate formats to represent digital content in specific contexts.

*Service Model*

The GDFR defines a set of core registry services in two broad categories: Management Services and Access Services. The Management Services include the following:

- Approval: providing an appropriate level of technical review of registration information
- Maintenance: creation, updating, and deletion of format entries
- Notification: subscription-based notification of significant events regarding specific formats
- Introspection: machine-discoverable publication of local registry policies and practices

The Access Services include the following:

- Description: query mechanism for specific format representation information
- Export: bulk export of registry data

Service gateways will be provided for both human and machine interaction with the registry. Additional administrative services regarding delegation and synchronization between the individual nodes of the distributed registry network will be integrated into the GDFR protocol. The final determination of the inter-nodal synchronization mechanism will be informed by relevant work in this area by the Open Archives Initiative (OAI) (Van de Sompel & Lagoze, 2002) and LOCKSS (Lots of Copies Keeps Stuff Safe) projects (Reich & Rosenthal, 2001).

A further set of ancillary services can be envisioned, but for the time being their implementation is being left to external value-added service providers. These include implementation of, or service brokerage for, format-specific rendering, transformation, validation, characterization, and other relative services. The JSTOR/Harvard JHOVE tool for format-specific object identification, validation, and characterization (Chapman & Abrams, 2004) and the National Library of New Zealand (NLNZ) Metadata Extraction Tool (Searle & Thompson, 2003) are two well-known examples of systems whose implementation and maintenance would be facilitated by the existence of the GDFR to provide sufficiently detailed and authoritative format specifications.

*Governance and Business Model*

Two criteria for success of the GDFR project are long-term sustainability and trustworthiness. The GDFR governance structure and business model must facilitate both of these goals. Without trust in the authoritativeness of the representation information contained within it, the registry will not be utilized by digital preservationists. Without trust in the handling of proprietary representation information, such information will not be deposited with the registry, thereby significantly decreasing its potential value. Sustainability of the registry is essential to providing appropriate support for long-term digital preservation activities. Since today's operational repositories are gracefully handling a variety of formatted material, it is often difficult to imagine how easily that community knowledge of contemporary formats can be lost with the passage of time. The GDFR will function as the persistent memory of the digital preservation community to ensure that the format knowledge often taken for granted today will remain accessible to the community in the future.

It remains unclear if the GDFR should operate under the administrative aegis of some existing institution or if an entirely new organization is required. Regardless, it is important that the GDFR can be ensured of a predictable yearly revenue stream with which to fund its operation. Digital preservation requires an aggressively proactive approach with constant monitoring for obsolescence and periodic intervention to ensure the continuing viability of the digital assets under its managed care. Even a momentary disruption of preservation intervention at the point of major technological

change may result in the irretrievable loss of digital content. As with many common good services, the major business difficulty facing the GDFR is to provide income today for a benefit that may not accrue until tomorrow. In many ways, the administrative and business issues surrounding the GDFR will prove much more difficult to solve than the technical issues.

*Testbed*

The initial GDFR data and service models are being tested in a proof-of-concept prototype registry known as Fred (Format Registry Demonstrator) under development at the University of Pennsylvania Library. Fred (n.d.) is based on the Typed Object Model (TOM) format service broker architecture (Ockerbloom, 2004). When completed, this prototype will serve as a testbed for refining the data and service models and suggesting appropriate architectural and implementation decisions for the GDFR reference implementation.

## Conclusion

The concept of digital format permeates all 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. The proper interpretation of otherwise opaque content streams is dependent upon the internal syntactic and semantic details of formats in which digital content is represented. For purposes of long-term preservation of digital objects, this knowledge of format representation information must be sustainable over archival time spans. Additionally, the effective interchange of digital objects between repositories and other consuming agents requires mutual agreement on format syntax and semantics. This format representation information can be best collected, maintained, and disseminated through a distributed network of registries interoperating via standard protocols for delegation and synchronization.

The Digital Library Federation has sponsored an initial investigation into the technical, administrative, and business issues surrounding the establishment of a Global Digital Format Registry. An ad hoc working group with international participation has created provisional data and service models that are being implemented in a proof-of-concept system. Funding is being sought for a multiyear two-track project that will recommend an appropriate governance and business model for an operational registry and will implement, deploy, and populate a production-quality prototype registry. The development and implementation of the registry will require the expertise and consensus of a wider digital repository and preservation community. The GDFR project will encourage and welcome participation in the project from all appropriate stakeholders, including national, academic, and institutional libraries and archives; standards bodies; commercial in-

terests 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 preservation of digital assets. This project will lead to the establishment of a sustainable registry that can function as a key component of a future digital preservation infrastructure.

## References


Abrams, S. L., & Seaman, D. (2003). *Towards a global format registry.* Paper presented at *World Library and Information Congress: 69th IFLA General Conference and Council,* August 1–9, Berlin.

Amielh, M., & Devillers, S. (2002). *Bitstream syntax description language: Application of XML-Schema to multimedia content adaptation.* Paper presented at WWW2002: The Eleventh International World Wide Web Conference, May 7–11, Honolulu.

Arms, C. Y., & Fleischhauer, C. (2003). *Digital formats: Factors for sustainability, functionality, and quality.* Retrieved February 17, 2005, from http://lcweb2.loc.gov/ammem/techdocs/digform/Formats_DLF2003.ppt.

Chapman, S., & Abrams, S. L. (2004). Steering resources to safe-harbor repositories: The need for reliable, accurate, and affordable ingest services (pp. 98–102). In *Proceedings of Imaging Science & Technology 2004 Archiving Conference, San Antonio, Texas.* Springfield, VA: Society for Imaging Science and Technology.

Diffuse Project. (2003). Home page. Retrieved October 1, 2004, from http://web.archive.org/web/20031022140114.

Digital Preservation Testbed. (2003). *Emulation: Context and current status* [White Paper]. Retrieved February 17, 2005, from http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white_paper_emulatie_EN.pdf.

*Fred: A format registry demonstration.* (n.d.). Retrieved February 17, 2005, from http://tom.library.upenn.edu/fred/.

Freed, N., Klensin, J., & Postel, J. (1996). *Multipurpose Internet Mail Extensions (MIME) Part four: Registration procedures* (RFC 2048). Retrieved February 17, 2005, from http://www.ietf.org/rfc/rfc2048.txt.

Hedstrom, M., & Lee, C. E. (2002). Significant properties of digital objects: Definitions, applications, implications. In *Proceedings of the DLM-Forum 2002, 6–8 May, 2002, Barcelona: @ccess and preservation of electronic information: Best practices and solutions* (pp. 218–27). Lanham, MD: Bernan Associates.

Holdsworth, D., & Sergeant, D. M. (2000). A blueprint for representation information in the OAIS model. In B. Kobler & P. C. Harihan (Eds.), *Eighth Goddard Conference on Mass Storage Systems and Technologies: In cooperation with the 17th IEEE Symposium on Mass Storage Systems* (pp. 413–28). Greenbelt, MD: National Aeronautics and Space Administration.

Holtman, K., Mutz, A., & Hardie, T. (1999). *Media feature tag registration procedure* (RFC 2506). Retrieved February 17, 2005, from http://www.ietf.org/rfc/rfc2506.txt.

International Organization for Standardization. (2003). *ISO 14721: Space data and information transfer systems—Open archival information system—Reference model.* Geneva: Internationl Organization for Standardization.

———. (2004). *ISO/IEC 11179-1: Information technology—Metadata registries (MDR)—Part 1: Framework.* Geneva: International Organization for Standardization.

Joint Information Systems Committee (JISC). (2002). *The File Format Representation and Rendering Project.* Retrieved February 17, 2005, from http://www.jisc.ac.uk/index.cfm?name=project_fileformat.

Library of Congress. (2002). *Digital preservation: Welcome to the National Digital Information Infrastructure and Preservation Program.* Retrieved February 17, 2005, from http://www.digitalpreservation.gov/.

Liskov, B., & Wing, J. M. (1994). A behavioral notion of subtyping. *ACM Transactions on Programming Languages and Systems, 16*(6), 1811–41.

Lorie, R. (2002). A methodology and system for preserving digital data. In *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries, Portland, Oregon* (pp. 312–19). New York: ACM Press.

National Archives and Records Administration et al. (1999). *Archival Workshop on Ingest, Identification, and Certification Standards, October 13–14, 1999, College Park, MD* (ISO Archiving Workshop Series). Retrieved August 15, 2005, from http://ssdoo.gsfc.nasa.gov/nost/isoas/awiics/.

National Archives of England, Wales, and the United Kingdom. (2005). *PRONOM: The File Format Registry.* Retrieved September 27, 2004, from http://www.nationalarchives.gov.uk/pronom/.

National Institute of Standards and Technology. (2002). *Data Formats of the NSRL Reference Data Set (RDS) Distribution.* Retrieved February 17, 2005, from http://www.nsrl.nist.gov/documents/Data-Formats-of-the-NSRL-Reference-Data-Set-12.pdf.

OASIS. (2003). *OASIS/ebXML Registry Information Model v2.5.* Retrieved February 17, 2005, from http://www.oasis-open.org/committees/regrep/documents/2.5/specs/ebrim-2.5.pdf.

Ockerbloom, J. M. (2004). *The Typed Object Model: Support for diverse formats.* Paper presented at the ERPANET seminar, File Formats for Preservation, May 10–11, Vienna. Retrieved February 17, 2005, from http://tom.library.upenn.edu/pubs/TOM-ERPANET.ppt.

OCLC/RLG Working Group on Preservation Metadata. (2002). *Preservation metadata and the OAIS information model: A metadata framework to support the preservation of digital objects.* Retrieved February 17, 2005, from http://www.oclc.org/research/pmwg/pm_framework.pdf.

Reich, V., & Rosenthal, D. S. H. (2001). LOCKSS: A permanent Web publishing and access system. *D-Lib Magazine, 7*(6). Retrieved February 17, 2005, from http://www.dlib.org/dlib/june01/reich/06reich.html.

Searle, S., & Thompson, D. (2003). Preservation metadata: Pragmatic first steps at the National Library of New Zealand. *D-Lib Magazine, 9*(4). Retrieved February 17, 2005, from http://www.dlib.org/dlib/april03/thompson/04thompson.html.

Van de Sompel, H., & Lagoze, C. (2002). Notes from the interoperability front: A progress report on the Open Archives Initiative. In M. Agosti & C. Thanos (Eds.), *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries, September 16–18, 2002, Rome* (Lecture Notes in Computer Science, 2458, pp. 144–57). London: Springer-Verlag. Retrieved February 17, 2005, from http://www.openarchives.org/documents/ecdl2002-oai.pdf.

Wheatley, P. (2001). Migration—A CAMiLEON discussion paper. *Ariadne, 29.* Retrieved February 17, 2005, from http://www.ariadne.ac.uk/issue29/camileon/.

---

Stephen L. Abrams, Digital Library Program Manager, Harvard University Library, 1280 Massachusetts Avenue, Suite 404, Cambridge, MA 02138, stephen_abrams@harvard.edu. Stephen Abrams is the Digital Library Program Manager at the Harvard University Library, where he provides technical leadership for strategic planning, design, and coordination of the Library's digital systems, projects, and assets. He is currently engaged in research and implementation of effective methods for archival preservation of digital objects. Mr. Abrams was the project manager for the joint JSTOR/Harvard JHOVE project and is the ISO project leader and document editor for ISO/TC 171/SC 2/WG 5, the joint working group developing the PDF/A standard. He is a member of ACM, ALA, ASIS&T, and the IEEE Computer Society.