

Proposal for a Format Registry for Digital Library Preservation

The impetus for this proposal grew out of mutual concerns raised during conversations between the digital library staffs at Harvard University and the Massachusetts Institute of Technology. Through these discussions we found that our preservation projects shared a common challenge in the area of unambiguous data format identification and description. We were each attempting to solve individually what we now realize is a common problem; one that can be addressed more effectively and efficiently through a common solution. This proposal is being circulated in the hope that its central premise – the widespread usefulness of a global format registry – is relevant to other digital library programs. We hope that these programs and their institutions will be interested in joining a collaborative effort to investigate and implement a standardized registry for data format characterization.

The Format Characterization Problem

Digital library programs for long-term preservation of digital materials require a strong typing mechanism in order to characterize effectively the structural and semantic properties, or *data format*, of those materials for purposes of validation, presentation, monitoring for obsolescence, transformation, and interchange. The difficulties facing the digital preservationist today with regard to format typing are twofold:

- (1) Non-standardized format identification
- (2) Non-standardized, ephemeral format description

The first problem can be overcome by the introduction of authority control over format type identifiers [IFLA, Maxwell]. This would insure, for example, the unambiguous semantic interpretation of the common variant names “TIFF”, “*.tif”, “TIFF 6.0”, and “Tag Image File Format”. Once registered, a format identifier would have a persistent binding to a specific, fixed, publicly discoverable meaning.

While pertinent descriptive information about the syntax and semantics of data formats has been available for some time from a variety of sources, both print and online, it has generally been collected on an ad hoc basis and is of unknown quality and long-term persistence [File, Graphics, Murray]. The best example of a controlled source for format identification and description is the IANA media type registry [IANA].¹ For purposes of digital preservation, however, MIME typing does not provide a sufficiently granular characterization of formats. For example, the MIME type `application/msword` by itself doesn't help to determine whether a file is compatible with Microsoft Word 6.0 or 2000.² Similarly, `image/tiff` is equally applicable to both an unstripped, bi-tonal TIFF 5.0 file with CCITT T.4 compression and a tiled, CMYK color TIFF 6.0 file with LZW compression. The important disambiguation between these variant instances requires a type characterization that encapsulates the specific constraints beyond those generically defined by the Word or TIFF specification. For the remainder of this paper the term “format” will be used in the sense of a specific format profile of arbitrary granularity, rather than in its more commonly understood MIME-like sense.

Proposed Activity

Many institutions and programs involved in digital preservation face the same decisions and tasks when confronting the current inadequate state of format typing. In order to prevent wasteful duplication of effort, we propose a global registry of digital object formats that will provide identifier authority control

¹ IANA requires “[a] precise and openly available specification” for all types defined in the IETF tree. Specifications for types in the vendor and personal trees are “encouraged but not required”; see RFC 2048.

² Technically, the IANA registered `application/msword` media type does allow a version parameter, e.g., `application/msword; version=6`, which would permit some level of disambiguation; see <http://www.iana.org/assignments/media-types/application/msword>. However, this is little supported feature of the MIME type and is seldom, if ever, seen in practice.

and standardized format descriptions. More formally, the format registry will establish a persistent, unambiguous binding between unique format *identifiers* and encoded *descriptions* of those formats. For example, the binding (“xyz”, XYZ) establishes the permanent association between the identifier “xyz” and the description of format XYZ. Since varying levels of granular constraints can lead to families of related types, the registry could incorporate an inheritance mechanism to reduce redundant data storage. For example, MyTIFF ⇒ OurTIFF ⇒ TIFF6.0 ⇒ TIFF, where the arrow symbol (⇒) represents a UML-like *generalization* (“IS-A”) operator. A child type would inherit a baseline specification from its parent *in its entirety*, to which it can add additional, more specific constraints.

The determination of the syntax of format identifiers requires that careful consideration be given to questions of semantic opacity, transcribability, and susceptibility to encoding in other identification schemes such as URIs, cf. [Williams]. The descriptive component of the registered bindings should minimally include an authoritative format specification. Ideally, the registry would maintain a physical copy of all relevant format specifications, in either paper or digital form (which would in turn require strong format typing to insure viability). In practice this may prove to be problematic for a variety of policy, technical, and intellectual property rights reasons. In such cases, links to persistent external sources are necessary. Depending upon the nature of the linked-to resources, such links may be actionable (URI, DOI) or not (ISBN, “third shelf in the cabinet down the hall”). Beyond the core format specification, many other descriptive attributes could be usefully registered in the areas of structural and semantic characterization, tool support, and internal registry administration.

The necessity for repository-specific format registries has been recognized by many digital library projects, such as CEDARS and DSpace [Holdsworth, Bass]. The IANA character set and media type registries provide examples of global centralized services [Freed]. The distributed broker model introduced in TOM demonstrates a method for building a federated registry, with its attendant potential for reducing duplicative effort [Ockerbloom]. The question of a centralized vs. replicated vs. distributed architecture is one of many pertinent questions requiring additional analysis. Regardless of the specifics of architecture and implementation, however, a standardized format registry will prove of great value to all institutions concerned with the long-term preservation of digital assets. The registry will be deemed successful if it can provide a sustainable registration, storage, and access platform, and if it can usefully distribute the implementation and operational costs of that platform across its institution users.

Desired Benefits

The intention of the registered format descriptions is to allow the characterization of the well-formedness of digital encodings along a continuum from syntactically correct to intellectually meaningful. For operational digital repositories, this information is of great value for planning ongoing preservation activities. By attaching characterization information to formats, rather than to individual objects, repository ingest workflow can be simplified by reducing pre-ingest metadata gathering. Some measure of the relative risk of format obsolescence can be drawn from examination of the number and quality of supporting tools, utilities, and services. Additionally, repositories could choose to maintain migration provenance by reference to registered formats, rather than by retaining all successive generations of an artifact. For the digital archeologist, registry descriptive information will provide valuable assistance during future reconstruction efforts [Ross].

Open Questions

Additional collaborative work is needed to answer the following non-exhaustive list of questions:

- (1) Scope: Digital library stuff only or digital everything?
- (2) Audience: Who can register? Who can access?
- (3) Is the registry human readable or machine actionable?
- (4) Determination of identifier syntax and required/ optional descriptive attributes
- (5) Source of descriptive information: Effect of DCMA on reverse engineering of closed formats?

- (6) Architecture: Centralized/replicated/distributed registration/storage/discovery/delivery?
- (7) Obligation of registrants (responsible for maintaining validity of registered data in perpetuity?)
- (8) Imprimatur: Is this at some point an IETF, ISO, NISO, W3C, etc. activity?
- (9) Appropriate administrative structure for long-term sustainability
- (10) Governance: Who will decide? Who will pay?

Strawman for Descriptive Attributes and Format Coverage

A wide variety of additional descriptive information could also be maintained in the registry. The added utility of any such expansion of the scope of registration must be carefully balanced against the concomitant costs in increased implementation and operational overhead. An initial set of potential descriptive attributes includes:

Core	
Name	Unique format identifier
Alias ^{O,R}	Equivalent identifier
Specification	(Pointer to) typed authoritative specification of the format
Synonym ^{O,R}	Identifier for the identical format in another registry
See also ^{O,R}	Identifier for the related format (e.g., previous version)
Characterization	
Magic number ^O	Unique internal signature of format
File extension ^{O,R}	Customary (non-Macintosh) file extension
File type ^O	Macintosh-specific data fork file type
Orientation ^O	Binary or text orientation
DRM ^{O,R}	Internal DRM mechanism with implications vis-à-vis DCMA
Inheritance	
Super-type ^O	Identifier of parent format, from which this format inherits a baseline description
Utilities	
Tool ^{O,R}	(Pointer to) software for format instance creation, validation, rendering, etc.
Service ^{O,R}	Pointer to fee-based/free services supporting this format
Miscellaneous	
Schema ^{O,R}	(Pointer to) typed schema appropriate for encoding metadata
Migration ^{O,R}	Migration path/process to and from arbitrary formats (including quantification of potential syntactic/semantic loss)
Canonicalization ^O	Canonicalization process for instances of this format (useful for equality comparison)
Administrative	
Registrant	Contact information of original registrant and creation timestamp
Revision ^{O,R}	Contact information of modifier, modification timestamp, and descriptive note

O = optional; R = repeatable

At a minimum, the format registry should include a core set of important formats in widespread usage for encoding of digital library objects. A reasonable first approximation of such a set would include:

Data	Access, CPC, DXF, HDF, EPS, Excel, FileMaker, GIF, HTML, JPEG, LaTeX, MrSID, MySQL, netCDF, PDF, PostScript, RTF, SGML, TeX, TIFF, Word, XML
Encoding	BinHex, UUEncode
Package	JAR, TAR, ZIP

Note that some of these are closed proprietary formats, for which it may prove difficult to obtain sufficiently rich specification information. Nevertheless, end-user expectation draws no such distinction between open and closed formats. Thus, in order to achieve public recognition of success, digital library programs must tackle the difficult support issues surrounding these proprietary formats.

Example

The following is an example of a potential registry entry for RTF in a simple, human-readable encoding:

```
Name: RTF1.6
Specification: title="Rich Text Format (RTF) Specification, Version 1.6",
date=1996-05, scheme=URI, format="HTML4.0"; charset=UTF-8,
link=http://msdn.microsoft.com/library/?url=/library/en-us/
dnrtfsspec/html/rftfsspec.asp
Synonym: text/rtf; registry=MIME
See-also: RTF1.0; relationship="previous version"
Magic-number: "{\rtf1"; offset=0
File-extension: rtf
Orientation: text; charset=USASCII
Tool: Microsoft Word
Registrant: name="John Q. Registrant", affiliation="Digital Archive",
email=jqr@digiarhive.org, date=2002-07-19
```

References

- Bass, Michael J., et al., *DSpace – A Sustainable Solution for Institutional Digital Asset Services – Spanning the Information Asset Value Chain: Ingest, Manage, Preserve, Disseminate*, March 1, 2002, <http://web.mit.edu/dspace/www/implementation/design_documents/architecture.pdf>.
- File Formats at filespecs.com: Your Source for File Formats, Specification, and Definitions* (July 16, 2002) <<http://www.filespecs.com/index.jsp>>.
- 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>>.
- Freed, N., J. Klensin, and J. Postel, *Multipurpose Internet Mail Extensions (MIME) Part Four: Registration Procedures*, RFC 2048, BCP 13, November 1996 <<http://www.ietf.org/rfc/rfc2048.txt>>.
- Freed, N. and J. Postel, *IANA Charset Registration Procedures*, RFC 2978, BCP 19, October 2000 <<http://www.ietf.org/rfc/rfc2978.txt>>.
- The Graphics File Format Page* (January 15, 2002) <<http://www.dcs.ed.ac.uk/home/mxr/gfx/>>.
- Holdsworth, David, and Derek M. Sergeant, "A Blueprint for Representation Information in the OAIS Model," *Eighth NASA Goddard Conference on Mass Storage Systems and Technologies*, University of Maryland, March 27-30, 2000 <<http://esdis-it.gsfc.nasa.gov/MSST/conf2000/PAPERS/D02PA.PDF>>.
- IANA, *MIME Media Types* (January 2, 2002) <<http://www.iana.org/assignments/media-types/>>.
- IFLA Working Group on GARE Revision, *Guidelines for Authority and Reference Records*, UBCIM Publications – New Series, Vol. 23 (2nd ed.; Munich: Saur, 2001) <<http://www.ifla.org/VII/s13/garr/garr.pdf>>.
- Maxwell, Robert L., *Maxwell's Guide to Authority Work* (Chicago: ALA, 2002).
- Murray, James D. and William VanRyper, *Encyclopedia of Graphics Formats* (2nd ed.; Sebastopol: O'Reilly, 1996).
- Ockerbloom, John, *Mediating Among Diverse Data Formats*, CMU-CS-98-102, Ph.D. thesis, Carnegie-Mellon University, January 14, 1998 <<http://www-2.cs.cmu.edu/People/spok/thesis.ps>>.
- Ross, Seamus and Ann Gow, *Digital Archeology: Rescuing Neglected and Damaged Data Resources*, JISC/NPO, February 1999 <<http://www.ukoln.ac.uk/services/elib/papers/supporting/pdf/p2.pdf>>.
- Williams, Stuart, *TAG Finding: Mapping between URIs and Internet Media Types*, W3C, May 27, 2002 <<http://www.w3.org/2001/tag/2002/01-uriMediaType-9>>.