

File format typing and format registries

Document details

Project: SHERPA DP
Work Package: 6.3
Author: Gareth Knight
Version: Version 1.0
Document date: 02/03/2007
Change history:

<i>Date</i>	<i>Version</i>	<i>Author</i>
02/05/2006	First draft	Gareth Knight
02/03/2007	First version	Gareth Knight

Contents

Introduction.....	2
Functions of a Format Registry.....	2
Identification of Representation Information.....	5
Repository-level tools	6
Summary	9
Recommendations for the AHDS	9
References	10

Introduction

Repository policies and procedures regarding the ingest, storage, management, and distribution of digital content is frequently influenced by the file format – the method a software application encodes information. Institutional repositories often establish restrictions on the file formats that they are willing to accept in their deposit guidance and staff are issued with instructions on effective methods of converting the data into a suitable distribution format. To manage and preserve digital content in the long-term, an understanding of encoding methods is essential. However, the large number and complexity of file formats in existence make this a difficult task. Digital archive staff may gain familiarity with a small number of file formats. However, they are unlikely to have extensive knowledge of a large number of formats, particularly if they operate in many different subject fields. Two types of format identification and validation services may be identified – remote format registries that provide centralized access to representation information and repository-level software that operate on local hardware. This paper describes functionality currently provided by format registries and repository-level tools to identify, describe and manage file formats.

Functions of a Format Registry

The JISC Information Environment (Bruce, et al, 2003) is developed on the premise that it is unnecessary to develop and implement all required services in a single institution. The multi-level structure of the digital environment, it indicates, allow different organisations to provide different levels of service, according to their funding level and infrastructure. These services may be offered to other institutions in the information environment, avoiding potential duplication of funding and effort. Format registries are promoted as a centralised service that is able to provide detailed information on a number of file formats, offering significant benefits for repositories responsible for the management of different data types. Specifically, they may offer several services identified by the OAIS reference model as important for the management process and inform preservation practices. This may include the validation of the Submission Information Package (SIPs) and the automatic extraction of representation information necessary to understand the resource. Brown (2005) identifies six functions that a format registry may perform:

1. Identify the format of a digital object
2. Validate that a digital object is the format it claims to be and it complies with a published specification
3. Provide advice on transforming the digital object from its source format to a destination format.
4. Identify the significant properties of a digital object and creating Representation Information (RI).
5. Assess the risk of obsolescence to a digital object
6. Identify the method to render an object.

Format registries are currently in the early stages of development by institutions such as the UK National Archives (PRONOM) and Harvard University Library (Global Digital Format Registry). The first goal of these organisations is to establish a basic service to store technical and descriptive information regarding file formats and developing a method to express them through a valid URI (Uniform Resource Identifier). However, they do not perform further operations of validating the assets stored in a digital repository and assessing the obsolescence risk.

Three services may be identified at the time of writing that operate as format registries:

1. PRONOM

PRONOM is an online system developed by The National Archive (UK) for managing information about the file formats used to store intellectual content created by government departments. It stores information about file formats, software products and software vendors responsible for their development. A recent estimate (Soper,

undated) indicates the system currently holds details of 550 file formats, 250 software products, and 100 software vendors. The system stores signature information that may be used to perform automatic format identification of a digital object. This signature may be internal (e.g. magic numbers) or external (a file extension) to the bitstream. A unique PRONOM identifier is assigned to each format (e.g. info:pronom/fmt/14) that allows the repository service to locate further information on the resources that they store.

The National Archive has not as yet, published details of the PRONOM API. However, it is available to project partners, such as the PRESERV project and may be accessed through the DROID client software.

2. GDFR

The Global Digital Format Registry (GDFR) is a service currently being developed by the Harvard University to collect and store format representation information. The most significant architectural aspect of the GDFR is the distributed, redundant approach to the storage, discovery, and delivery of representation information. The GDFR architecture is envisaged to consist of several format registries that communicate with each other and synchronise their information. Each format registry will serve as a network node that will have the ability to submit new information and propagate it through the registry network. The registries operating in the GDFR programme (OCLC, University of Leeds, JSTOR, MIT, PRO) will provide services necessary to manage digital objects in the user community:

1. **Management services:** The management service will perform functions necessary to identify new formats and subsequent versions, as well as store format obsolescence information. It will also review new and modified format information, and notify clients of new/updated formats.
2. **Access services:** The access services will provide representation information to other format registries or preservation repositories on request. This may be information for a single or multiple formats stored by the format registry.
3. **Representation services:** Representation services will identify and validate the digital object by comparing the attributes of a digital object to the profiles stored by the format registry.
4. **Brokerage services:** The brokerage service will provide functions necessary to identify the rendering requirements of a digital object (e.g. specific software), convert the digital object to a target format, and extract relevant metadata.

Each registry will store format information necessary to identify and manage the digital object. This will include the format name, MIME type, external signatures (file extension), internal signatures (e.g. magic number), authoritative specification document(s), ontological classification, relationships to other formats (e.g., subtype-of, new-version-of, can-be-encapsulated-by), references to systems, services, and tools that support the format as an input or output and details of the format author, IPR holder, and (if different) maintenance agency. Format information may be disseminated through the network using two methods:

1. **Vetted:** Vetted representation information must be reviewed and authenticated as technically valid prior to distribution.
2. **Non-vetted:** Non-vetted information is immediately propagated through the network without further technical review. The credibility of the information is based on the reputation of the submitting agent.

The intra-registry interaction will be maintained through a management database that records the date of the most recent metadata harvest, and additional information that may be useful (GDFR Data Model, 2004).

3. FRED: A format registry demonstration

FRED (Format REgistry Demonstrator) is a proof-of-concept prototype format registry demonstrator developed from the Mellon-funded TOM (Typed Object Model) project. Its purpose is to provide some idea of the requirements that developing a global registry would entail and will not form the basis of a digital format registry. The FRED demonstrator complies with the XML specification defined for the GDFR prototype specification (<http://tom.library.upenn.edu/fred/schemas/format.xsd>) and is designed as a practical implementation of the Typed Object Model (TOM) for the description of data types and formats. The model consists of two components:

1. An underlying data structure intended to describe the representation and, in some cases, the behaviour of an "information source" (a file format, classification system, or registry).
2. Software that interprets the data structure provides information to the client system, and contacts third party services to interpret and convert data stored in these formats.

The FRED demonstrator provides limited functions of the TOM 'Type Broker' – it maintains a database of information on data types and associated services that may be queried by a client. However, it does not contact other services – which did not exist when the registry was implemented – to perform requested functions (e.g. migration) (Ockerbloom, 2004).

The format registry provides limited information on 19 file formats (AIFF, Microsoft Wave, PNG, TIFF, HTML, XHTML, XML, Microsoft Word, RTF, MARC records, Microsoft Ole2 Compound Document Format, ASCII, UTF-8, Adobe PDF, Adobe Postscript, Microsoft Rich Text Format, JPEG, Microsoft Excel & Microsoft PowerPoint). The information provided for each format differs according to the data type. Common components include: [the persistent identifier \(info:gdfred/f/png\)](mailto:info:gdfred/f/png); the format title, format type (e.g. ISO standard) and value (e.g. ISO/IEC 15948), mime type, registered organization (e.g. Microsoft Corporation, World Wide Web Consortium), organization type (non-profit entity), web site location, and additional information regarding specific characteristics.

4. Library of Congress: Digital Formats

A fourth source of information on digital file formats is the Library of Congress' Digital Formats web site. The web site was initially developed to assist library staff in the management of digital collections deposited with the Library of Congress and was later made available for use by others. The public resource provides human-readable guidelines on how to identify and manage specific digital formats. File format information is separated into seven categories:

1. **Identification and description** – Basic information on the format name, a brief characterisation of the format, indication of how the format is generally used during the content life cycle (e.g. during creation, management or distribution), and information on how it relates to other file formats.
2. **Local use** – An indication of local use in the Library of Congress, including experience of existing holdings (i.e. those previously handled by staff) and preferred format for storage.
3. **Sustainability factors** – The level of sustainability examines seven factors necessary to manage the file format: 1) the amount of technical information available about the format (*disclosure*); 2) public citation of a specification or other documentation (*standardization* and *other documentation*); 3) indication of any patents or licensing issues (Licensing and patent claims); 4) a statement regarding the ease of locating or building suitable tools to render

the format (*transparency*); 5) an indication of the ability of the format to incorporate metadata (*self-documentation*); 6) an indication of the need for external software or hardware (*external dependencies*); and 7) support for intellectual property protection, e.g. through digital rights management (*Technical protection considerations*).

4. **Quality and functionality factors** – Specific information related to the characteristics of the file type (image, sound, video)
5. **File type signifiers** – Significant properties that an automated tool may use to identify the file format or encapsulated content. For example, file extension, magic number, etc.
6. **Notes** – Additional information about the file format.
7. **Useful references** – Web links to other resources on the file format

The Library of Congress web site is the most complete, in terms of the amount of information available on file formats. At the time of writing (May 2006), the web site describes 74 file formats and 186 encoding methods. Although the Library of Congress provide much of the same information as The National Archives PRONOM and Harvard University's proposed Global Data Format Registry, the LOC do not describe the resource as a format registry and have not adopted the 'info' resource identifier. The page for each format description is allocated a unique filename (e.g. <http://www.digitalpreservation.gov/formats/fdd/fdd000057.shtml>). The value increases incrementally and appears to be assigned in order of entry (i.e. fdd000058.shtml, fdd000059.shtml, etc.). The Library of Congress do not, at present, provide a machine-readable description of format information to assist with format identification. However, it may be useful to reference the format identifier in the formatRegistryKey of the PREMIS-compliant preservation element set.

Identification of Representation Information

A persistent identifier system is required to link digital objects stored in a digital archive with the correct representation information stored by a format registry. Existing format registries implement the 'Info' URI (Uniform Resource Identifier) – a non-resolvable identifier system developed for use by the library and publishing community to identify information assets (OCLC, undated). The info identifier system may be applied to institutions, such as the Global Data Format Registry (info:gdftr) and the UK National Archives (info:pronom); repository software, such as the Fedora object disseminators (info:fedora); or existing resource identifiers, such as Handle (info:hdl). Although format registries provide similar functionality – to identify and describe file formats stored in a digital archive – there is little consistency between these organizations, at the moment, on refinements to the scheme.

PRONOM Persistent Unique Identifier (PUID)

The PRONOM Persistent Unique Identifier (PUID) is an extensible scheme for the identification of records in the PRONOM database. The PUID Scheme guidelines (Brown, 2005b) indicates a valid PRONOM identifier must be composed of two components: the class of representation information to which the identifier refers, and the unique identifier itself. This is expressed in the following syntax:

`<puid> = <puid type> '/' <identifier>`

The encoding format "fmt" is currently the only type of information asset identified as important in the PRONOM identification system. For example, *info:pronom/fmt/14* to *info:pronom/fmt/20* references representation information for PDF 1.0 – 1.6 (PDF Search Results, 2006). The identifier system does not, at the time of writing, make allowances for ontological classifications, registry information defined by the GDFR namespace, or the identification of embedded formats, such as JPEG images encoded in a PDF document. However, other types of information may be defined in the future.

GDFR identification system

The GDFR identifier system, as used by the Format Registry Demonstrator defines three types of information assets:

1. Ontological classifications (*urn:gdfrc:classid*)
2. File formats (*urn:gdfrf:formatid*)
3. Registries (*urn:gdfrr:registryid*)

The identifier system is currently limited to format identification. For example, *info:gdfrfred/f/html* references representation information for the Hypertext Markup Language (HTML); *info:gdfrfred/f/marc* may be used to identify MARC (Machine-Readable Cataloging) records; and *info:gdfrfred/f/png* references the PNG image format.

Summary

The PRONOM, Global Digital Format Registry and FRED demonstrator offer similar information that may be queried by a digital archive that wishes to identify the file formats stored in its repository. Although The National Archives and Harvard University Library indicate they intend to offer functionality to perform obsolescence checks and extract representation information directly from the digital object, they do not indicate when such functionality will be available. It is therefore impossible to choose a format registry at the current time. However, the decision to standardise the info: identifier and publication of a consistent approach to representation information identification should simplify the process of referencing a particular format registry in the preservation metadata at a later date. The AHDS should create an empty formatRegistryKey element in the preservation metadata that will be used as a placeholder to reference a particular format registry at a later date.

Repository-level tools

Repository-level software tools operate in the local environment of the digital repository. These tools are intended to capture technical information on digital objects at a greater level of granularity through the use of signature information. Two types of information may be identified:

1. External information – An indicator external to the object that may indicate the file format. For example, a DOS/Windows file extension or a Macintosh data fork.
2. Internal information – Information embedded in the object structure that identify the digital object and the bitstream(s) it contains. Many formats contain a “magic number” – an ASCII descriptor in the file header, or a byte sequence that is common to the format.

Embedded information is considered to be more reliable, due to it being less likely to change. A Macintosh data fork may be lost when transferring objects to a different platform and file extensions may be shortened by the operating system (e.g. a 4+ letter extension may be changed to three letters), or the file extension may be changed by the user (e.g. default .doc extension may be changed to .let to indicate it is a letter). The internal identifier may indicate general information, such as the file format, or specific information on the file format, version number, character encoding, number of objects contained in a document, and the number and type of embedded fonts. However, the programmer must have a greater understanding of the file format to develop format recognition modules and, as a result, the number of recognised file formats is much smaller, typically less than 10. Furthermore, the Java-based tools require some manual configuration to accept software updates, such as new format recognition algorithms, thus increasing the amount of work that a system administrator must perform.

Four software tools may be identified that provide some method of format identification and metadata extraction:

1. JHOVE

Jhove is a Java-based application designed to perform file identification and validation functions. The JHOVE architecture operates around a plug-in architecture, allowing modules for new file formats to be added at a later date. In a conference paper, Abrams (2004) indicates the format module outlines the types of representation information that would be provided by the GDFR. The software is able to recognise 12 distinct file formats and 45 encoding methods at the time of writing (April 2006) and various sub-categories. Unrecognised file formats are classified as a bytestream. Table 1 indicates the file formats listed on the Jhove web site.

Format type	Formats
<i>Image</i>	TIFF
	JPEG
	JPEG2000
	GIF
<i>Text</i>	HTML
	XHTML
	ASCII
	UTF-8
	XML
	PDF
<i>Audio</i>	Wave
	AIFF
<i>Other</i>	Unknown bytestream

Table 1: File formats recognised by JHOVE

JHOVE fulfills three functions required by the preservation service. It identifies the file format in which the intellectual content is stored, it determines the level of conformance in comparison to a published format specification and it extracts specific information regarding the significant properties of the digital object. The latter two features are particularly useful when identifying particular problems that will affect the long-term preservation of the digital object.

The JHOVE Representation Information may be output as ASCII text or XML, determined by the *-h* qualifier. The XML output is defined by the JHOVE schema, located at <http://hul.harvard.edu/ois/xml/xsd/jhove/jhove.xsd> in the basic installation. The JHOVE metadata schema is organised into a series of RDF-like property containers that provide specific information – the property name, an arity (the number of values), a type, and a value. Where possible, XML output complies to a published scheme for technical metadata, e.g. the use of MIX for still raster image or the draft AES-X098B Core metadata for audio objects (JHOVE Tutorial, 2005). For the purpose of the Sherpa DP project, XSLT scripts must be developed to transform output to the PREMIS-based XML scheme. In total, eight elements are produced by the JHOVE tool that is considered essential for the preservation of e-prints, outlined in Knight (2005). These are:

1. Size
2. Format name
3. Format version
4. Significant properties
5. Inhibitor type
6. Inhibitor key
7. Message digest algorithm
8. Message digest.

In most circumstances, it will be a simple task to extract the output stored in a specific element and encapsulate it in a PREMIS compliant scheme. An exception is the

Significant Properties component that will require information that may or may not exist in the schema, dependent upon the file format.

2. DROID (Digital Record Object Identification)

DROID (Digital Record Object Identification) is a Java-based software tool developed by The National Archives to perform automated batch identification of file formats and will report the corresponding PUID (if assigned). Format identification signatures are downloaded from the PRONOM technical registry (2005) and stored locally in an XML signature file. The signature record contains the following information: format name, version, internal identifiers (magic number and other unique properties); and the PRONOM unique identifier (PUID). The DROID software may be configured programmatically or manually by the user to identify the file format of data formats submitted into the digital archive. The software produces a “file collection” XML document that reports the software and signature version used to identify the object, and other relevant information provided in the identification process:

1. The path and filename of the digital object(s) selected for identification;
2. The format name that has been identified, stored as a DROID internal value (e.g. “Format C2”);
3. Format version (e.g. “v1.2”)
4. FormatPUID (a plain text description that repeats the format and version in plain text, e.g. “v1.2 of format C”)

The DROID tool will first compare the file with the internal identifier. If a single match is found that identifies it as a specific format and version it is classified as a specific match (“*positive (specific)*” in the XML markup). If the file matches two or more identifiers, it is classified as a generic match (“*positive (generic)*”. A second analysis examines the file extension (external identifier) and compares it to the identifier set. This may confirm the internal analysis or indicate a mismatch (e.g. the file extension does not match an internal signature). Table 2, developed by Brown (2005a), indicates six responses that may be produced by the DROID tool.

Status	Warning message	Meaning
Positive (Specific)		The file matches a specific internal signature
Positive (Specific)	Possible file extension mismatch	The file matches a specific internal signature but the file extension does not match any associated external signatures
Positive (Generic)		The file matches a generic internal signature
Positive (Generic)	Possible file extension mismatch	The file matches a generic internal signature but the file extension does not match any associated external signatures
Tentative		The file matches an external signature but no internal signatures
Negative		The file does not match any internal or external signatures

Table 2: Responses generated by the DROID software tool

DROID has the advantage of recognizing a large number (100+) file formats and versions. However, it is unable to extract further information, such as significant properties, inhibitor types and targets, or generate checksums. If considered for use in the AHDS preservation repository, it must be implemented with other software that is able to perform such functionality.

3. NLNZ Metadata Extraction Tool

The NLNZ Metadata Extraction tool is a Java-based application developed for use by The National Library of New Zealand. The software tool is capable of extracting detailed technical information for a restricted list of file formats. The NLNZ Metadata

Extraction Tool fulfils one function required by the preservation service: it extracts specific information regarding the technical composition of a digital object. The tool does not currently perform any file format identification or verification. Instead it uses the standard three-letter file extension to identify the correct file format to use. It is therefore impossible to determine if an unknown document is in an incorrect format or may simply have the incorrect file extension. Currently the metadata extraction tool produces representation information for the following file formats:

Format type	Formats
<i>Image</i>	TIFF
	JPEG
	GIF
	BMP
<i>Text</i>	Microsoft Word
	Microsoft Works
	Open Office
	Word Perfect
	Adobe PDF
<i>Audio</i>	Wave
	MP3
<i>Presentation</i>	Microsoft Power Point
<i>Other</i>	Unknown bytestream

Table 3: File formats recognised by the NLNZ Metadata Extraction tool

The number of file formats supported is more extensive than the JHOVE software tool. However, less information is extracted. In total, two elements – file format and size – are extracted for PDF documents by the NLNZ metadata extraction tool. Further information is provided on the file format version for Microsoft Word documents.

4. Metadata Miner Pro

Metadata Miner Pro is a commercial application developed by Soft Experience. Its primary purpose is to extract descriptive information, such as title, author, subject, keyword, from common document formats (Microsoft Office applications, OpenOffice, HTML, Adobe PDF, JPEG/TIFF/PSD IPTC-NAA fields and Apple Mac file comments). Limited technical metadata is created on the filename and location, file size, file type (indicated by file extension), creation date and last modification date. It does not provide sufficient information to be considered a viable alternative to JHOVE or the NLNZ Metadata Extraction Tool.

Summary

The repository-level software tools offer a practical method of identifying the file format of a digital object and extracting limited technical information in the short-term. Notably, the use of JHOVE offers significant benefits for a preservation service, such as the AHDS, in creating appropriate Representation Information. Format identifications tools, such as DROID, NLNZ Metadata extractor and Metadata Miner are designed to provide basic information on the file format. However, they do not provide sufficient detail to preserve the digital object. In the short-term, it is recommended the AHDS Preservation Service use JHOVE to generate technical information and consider the implementation of a format registry when they provide functionality, such as obsolescence monitoring that is unavailable in a repository-level application.

Recommendations for the AHDS

1. The AHDS should create an empty element in the preservation metadata to record the format registry reference (info:) for the specific file format. The PREMIS data

- dictionary indicates the formatRegistryKey element should be used for this function. A reference to a valid format may be completed at a later date.
2. Review the suitability of format registries every six months and implement support at a future date.
 3. The AHDS should implement JHOVE as a practical method to generate technical information.

References

Abrams, S. (2004). The role of format in digital preservation. Retrieved on May 01, 2006 from: <http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/2870340201.html>

Anon (2004). Global Digital Format Registry (GDFR) Data Model v.4. Retrieved on May 01, 2006 from: <http://hul.harvard.edu/gdfr/documents/DataModel-v4-2004-01-12.doc>

Brown, A. (2005a). Automatic format identification using PRONOM and DROID. Retrieved on May 01, 2006 from: http://www.nationalarchives.gov.uk/aboutapps/fileformat/pdf/automatic_format_identification.pdf

Brown, A. (2005b). The PRONOM PUID Scheme. Retrieved on May 01, 2006 from: http://www.nationalarchives.gov.uk/aboutapps/pronom/pdf/pronom_unique_identifier_scheme.pdf

Bruce, R et al (2003). Investing in the Future: Developing an Online Information Environment. Retrieved on May 23, 2006 from: <http://www.jisc.ac.uk/ie/>

Knight, G. (2006). Proposal for a minimum preservation metadata element set based upon the PREMIS data dictionary. Internal working document.

Jhove (2005). JSTOR/Harvard Object Validation Environment. Retrieved on May 01, 2006 from: <http://hul.harvard.edu/jhove/>

Library of Congress (2006). Sustainability of Digital Formats: Planning for Library of Congress Collections. Retrieved on May 01, 2006 from: <http://www.digitalpreservation.gov/formats/>

Ockerbloom, J. (2004) FRED: Format REgistry Demonstration. Retrieved on May 01, 2006 from: <http://tom.library.upenn.edu/cgi-bin/fred?cmd=ShowDocu&&id=about>

Ockerbloom, J. (2004). What is TOM?. Retrieved on May 01, 2006 from: <http://tom.library.upenn.edu/intro.html>

OCLC (undated). About "Info" URIs: Frequently Asked Questions. Retrieved on May 24, 2005 from: <http://info-uri.info/registry/docs/misc/faq.html>

Soper, F. (undated). The PRONOM File Format Registry. Retrieved on May 01, 2006 from: <http://www.erpanet.org/www/workgroup/documents/soper-pronom.pdf>

The National Archives (2006). PRONOM: The Technical Registry. Retrieved on May 01, 2006 from: <http://www.nationalarchives.gov.uk/aboutapps/pronom/default.htm>

TOM Project (2003). Fred format registry schema for Format. Retrieved on May 01, 2006 from: <http://tom.library.upenn.edu/fred/schemas/format.xsd>

Various (2006) Wikipedia: Arity. Retrieved on May 23, 2006 from: <http://en.wikipedia.org/wiki/Arity>