BRIEFING PAPER: FILE FORMAT AND XML SCHEMA REGISTRIES

ALEX BALL

kim12rep002ab10.pdf

ACCESS LEVEL: 1

ISSUE DATE: 31 MAY 2006

APPROVED BY: CHRIS MCMAHON

DATE APPROVED: 4 JUNE 2006

1 INTRODUCTION

Digital files are, fundamentally, strings of binary digits (bits). In order to process them, one must know the format they are in, and further, what software is needed to read that format. Even after the file has been successfully opened, extra information may be needed in order to fully understand the contents. In the terms of the Open Archival Information System (OAIS) Reference Model, the information required to transform a stream of bits into something intelligible is called *representation information* [CCSDS 2002].

Providing sufficient representation information so that digital files remain perpetually intelligible is a significant burden, even if one restricts one's attention to the machine-readable aspects of the problem. The OAIS Reference Model describes a technique for reducing this burden on individual archival information systems: using representation networks. Quite simply, this means referring to external pieces of representation information instead of reproducing them for each file to which they apply.

One obvious use for the representation network idea is to separate out representation information solely concerned with file formats into its own database. That way, whenever a file is added to a repository, the format-related representation information can be expressed by a single reference to the database. Also, should some format-related representation information need updating, the changes need only be made once. Such a database is known as a file format registry; it could be maintained by the repository itself or it may be shared between several repositories.

There are two levels at which a file format registry can run. It can either be maintained as a node in a wider network, referring to externally-held file format specifications and compatible software, or it can attempt to be a terminus, collecting and preserving copies of specifications and software in a local repository. The former type of registry is easier and cheaper to maintain, but may run into trouble if the specification for a format ceases to be easily available. While this is unlikely to happen to formats that have been registered as standards — for example, ASCII [ANSI X3.4-1986], Unicode [Unicode Consortium 2003], Portable Network Graphics [ISO/IEC 15948:2004] or Open Document Format [OASIS 2005] — it is a real danger for proprietary formats, especially those that have been superseded. For example, PKWARE only publishes the specifications for the latest version of its ZIP format;[1] older specifications must be sourced from elsewhere. The problem is even more acute when it comes to software, which can easily become obsolete without a dedicated preservation strategy.

Given the amount of work it takes to keep the information in a registry current, a number of projects are aiming to set up registries that many repositories can use, to reduce unnecessary duplication of effort. This paper looks at a selection of these, as well as some established sources of representation information.

2 FILE FORMAT WEBSITES

A number of privately-maintained websites contain information about file formats. For example, Wotsit's Format is a site containing short descriptions, file extensions and format specifications for a wide variety of formats.[2] A similar level of information is held by the File Format Encyclopedia.[3]

1. URL: ⟨http://www.pkware.com/business_and_developers/developer/appnote/⟩.
2. URL: ⟨http://www.wotsit.org/⟩.
3. URL: ⟨http://pipin.tmd.ns.ac.yu/extra/fileformat/⟩.

At the more detailed end of the spectrum, FILExt, 'the File Extension Source', is a web-accessible database that collects representation information about file formats and their associated extensions.[4] It appears to be part of a personal project that started in 1996. Currently records can only be searched for or browsed by file extension; however, since a given extension can be used by multiple formats, there may be several records for each extension.

Individual records have the following data fields:

- *Extension.* The characters following the last dot in a filename.
- *Program and/or extension function.* In case of formats native to a particular program, the program is named and where possible a link provided to a website from which the program may be obtained. In the case of generic formats, the format is named, and where possible a link provided to a website detailing the format.
- *Company.* The program vendor, or the owner of the proprietary format, with a link to their website.
- *Specific notes.* For example, a description of the nature of a program, or issues to bear in mind when opening/viewing the file (e.g. security implications, version incompatibilities, alternative viewers).
- *MIME type.* All MIME media types (as defined by RFC 2046) commonly associated with the format, whether or not they are the official MIME type registered with the Internet Assigned Numbers Authority (IANA).[5]
- *File classification.* If the format belongs to a recognisable genus, such as graphic, source code, or CAD/CAM, this is recorded here. This classification did not appear to be consistently or thoroughly applied at the time of writing.
- *Associated links.* Additional links to file format descriptions, specifications, or free viewers.
- *Identifying characters.* If files in this format have a characteristic binary header, it is given here in hexadecimal octets. If they have a characteristic text header, it is given here in both hexadecimal octets and ASCII text.
- *Program ID.* This refers to the string used by MS Windows in its registry to identify a given program. Programs that are known to have been associated with this extension (respecting format) in a Windows registry are listed here by ID.
- *General notes.* For example, that too many viewers to mention exist for this format, or that the record's veracity has not been checked.
- *Last modified.* Date of the last modification to the record.
- *Record ID.* Unique numerical identifier for the record.

Such websites are geared more towards the casual user than the curator, and their longevity may be considered doubtful; indeed, three of the seven sites mentioned by Leeds University [2003] are no longer accessible. They would therefore not form part of any long-standing representation information network, but may be a useful resource for those setting up their own registry.

4. URL: ⟨http://filext.com/⟩.
5. URL: ⟨http://www.iana.org/assignments/media-types/⟩.

3 LIBRARY OF CONGRESS DIGITAL FORMATS

Sustainability of Digital Formats: Planning for Library of Congress Collections is a web-based resource aimed primarily at providing advice on the suitability of various digital file formats for long-term preservation.[6]

In order provide a balanced view of the advantages and disadvantages of each format, the site assesses each format according to seven sustainability factors and content-type specific quality and functionality factors [Arms & Fleischhauer 2005]. The sustainability factors are:

- *Disclosure.* How fully is the format documented, and is this documentation freely or publicly available?
- *Adoption.* How widely used is the format among content producers and consumers?
- *Transparency.* How easy is it to analyse the format using basic tools? For example, SGML, XML, postscript, RTF and PDF, along with source code of all varieties, can be written as text files in UTF-8 encoding (using Base64 encoding for embedded non-textual bitstreams), and hence can be analysed very easily with a text editor. Similarly, raster images represented by raw bitmaps, and audio clips represented by pulse code modulation with linear quantisation, are easily analysed with basic tools. On the other hand, compiled, compressed or encrypted file formats are much harder to analyse.
- *Self-documentation.* Is it possible to embed (easily discoverable) metadata, in particular preservation metadata, within the format itself? For example, it is possible to embed XML-formatted metadata within PDF (v1.4 and higher), PNG, TIFF and JPEG format files [Adobe 2005].
- *External dependencies.* To what extent does the format depend on a particular piece of hardware, operating system or other piece of software, and how complex will it be to manage those dependencies in the future? Dynamic content typically suffers from more complex dependencies than static content.
- *Impact of patents.* Where patents are associated with particular formats, patent-holders can impede preservation activities by imposing high license fees or restricting the development of software tools. On the other hand, once patents have expired, the patent documentation can be useful source of disclosure.
- *Technical protection mechanisms.* Certain formats permit various security measures, such as password protection, restrictions on copying or moving the file, time-limits, or dependence on responses from a vendor's server.

Quality and functionality factors vary between content types, although in all cases there are functionality factors based on normal rendering and those based on beyond normal rendering. For example, static images have the following quality factors:

- *Clarity.* To what extent can the format support high resolution images? Various measures can be taken into account, e.g. pixels per linear unit, (colour) bit depth, and the presence or otherwise of compression-related artefacts.
- *Colour maintenance.* Can the format support the specification of a colour profile, thereby allowing the image to be adjusted for different output devices?
- *Graphic effects and typography.* How versatile is the format with regard to applying shadows, filters, transparency and fill effects (and so on) to vector objects and text? Can fonts and patterns be specified?

6. URL: ⟨http://www.digitalpreservation.gov/formats/⟩.

They also have the following functionality factors:

- *Normal rendering.* Tasks frequently performed on images include: zooming, resizing, resampling, publication-quality printing.
- *Beyond normal rendering.* Examples of specialist functions for images include: using multiple layers within the image to perform offset printing, storing the same image at different resolutions within the one file, interlacing (where the image is displayed with increasing clarity as it is rendered), simple animation/looping, variable resolution, metadata attachments.

These factors are not entirely independent. For example, if a format is is encrypted and compressed (low transparency), this need not be a problem if the methods by which it is encrypted and compressed are fully and openly documented (high disclosure). Similarly, formats may be well adopted despite poor sustainability or quality.

3.1 Record structure

The records used by the Library of Congress Digital Formats website have the following structure:

- *Format description properties.* These include a record identifier, a short name for the format, the categories of content that the format deals with (e.g. sound, still image, text), the categories to which the format itself belongs (e.g. character encoding, file format, wrapper, bundling format), the date the record was last updated, and whether the record is final, fully drafted or partially drafted.
- *Identification and description.* This section includes the full name of the format, a brief (natural language) description, and various statements concerning its relationship to other formats, e.g. has subtype, may contain, has earlier version. It also contains an indication of whether the format is primarily used at point of content creation (initial state), at the point of dissemination (middle state) or at the point of end use (final state).
- *Local use.* This section records how the Library of Congress currently deals with this format, and whether the format is preferred or deprecated in favour of another.
- *Sustainability factors.* This section describes the format in terms of the seven sustainability factors.
- *Quality and functionality factors.* This section describes the format in terms of content-type specific quality and functionality factors.
- *File type signifiers.* This section records various external file type identifiers associated with (but not necessarily uniquely identifying) the format, including: filename extension, MIME media type, magic numbers, Microsoft FOURCC code (codec identifier used in the Windows registry), Microsoft WAVE format indicator (codec identifier used in the Windows registry), ASF GUID, and Apple QuickTime data format code.
- *Notes.* This section provides room for lengthier descriptions of the format and its history.
- *Format specifications.* This section contains URLs for digital versions of the format specification, and references for print versions thereof.
- *Useful references.* Any other online or printed information about the format is referenced here.

While no official global identifiers for the records are advertised, it is possible to reference them using URLs of the form ⟨`http://www.digitalpreservation.gov/formats/fdd/[id].shtml`⟩, where `[id]` is replaced by the record identifier, e.g. `fdd000133`.

Given the advisory nature of the website, the records are designed with human rather than automated readers in mind. Thus the records are unsuitable as a basis for automated tools, but are a useful source of reference for digital curators.

4 PRONOM

PRONOM is a format registry being developed and maintained by The National Archives (TNA).[7] It is primarily intended to support TNA's own preservation work, but since February 2004 it has been available for others to use through TNA's website.

4.1 Architecture and service models

The architecture of PRONOM has so far been monolithic: it has been developed as single database with a web-based front end. In due course, PRONOM will be developed to allow automated data exchange with other databases and to interoperate with various automated services. This interoperability is key for PRONOM to fulfil the role for which it was designed, viz. the support of TNA business activities and processes [Brown 2004]. The specific services that PRONOM will support are as follows.

For electronic documents of unknown format being added to a data archive, PRONOM will support the automatic identification of the file format, and for documents of known format, will also support the automatic validation of the format. Furthermore, PRONOM will support the automatic extraction of metadata from individual files, and of course will provide the format-related representation information one would expect from a file format registry.

To aid in preservation tasks, PRONOM will be able to run analytical operations on the information it holds; for example, scanning for formats whose lack of software or hardware support mark it out as being in danger of obsolescence. PRONOM will also be used to automatically generate optimal migration pathways between formats. This latter function will form the basis for a delivery service so that upon request, electronic documents can be transformed automatically into a format that the end user can render.

Each file format registered with PRONOM is assigned a Persistent Unique Identifier (PUID), consisting of a short lowercase alphanumeric string [Brown 2005*b*]. Only one type of PUID has been defined so far, `fmt` for format, although other types are anticipated for compression methods, character encoding schemes and operating systems. In order to make PUIDs globally unique, and to widen their utility, the PUID scheme has been registered as an info URI namespace, with the form `info:pronom/...`. Hence `info:pronom/fmt/42` uniquely identifies the JPEG File Interchange Format (JFIF) version 1.00. TNA have not yet provided a resolution mechanism by which PUIDs can be used to point directly to the relevant information in the PRONOM database, although there are plans to introduce one in the future.

4.2 Information model

PRONOM's information model is based on five core entities [Brown 2005*a*]. The Actor entity covers individuals or organisations known to the registry by virtue of their connection to other entities. The Documentation, Intellectual Property Right and Identifier entities are self

7. URL: ⟨`http://www.nationalarchives.gov.uk/pronom/`⟩.

explanatory. The Technical Component entity is used to group sub-entities representing the various aspects of the technical infrastructure required to support a file in a particular format. The four main sub-entities in this area are: File Format, Software Component, Hardware Component and Storage Medium.

The File Format sub-entity has a number of other sub-entities associated with it, for example, Encoding, Compression Type, Internal Signature (i.e. a format and version identifier, recorded under the Byte Sequence sub-entity, that is embedded in the file itself), External Signature (such as a characteristic file extension or Mac OS data type), Classification (according to some named scheme or ontology) and Family (to which the format belongs conceptually).

In addition to these entities, the information model defines four possible types of relationships between entities. These are:

- *Software process.* This describes what a Software Component can do with a File Format (e.g. render, extract metadata).
- *Software requirement.* This identifies a Software Component that is required by a particular Technical Component, and optionally describes the nature of the requirement.
- *Hardware requirement.* This identifies a Hardware Component that is required by a particular Technical Component, and optionally describes the nature of the requirement (e.g. with a minimum of 50MB hard disk space).
- *Generic relationship.* This is intended to be used between two entities of the same type, e.g. to label one File Format as a subsequent version of another File Format.

5 GLOBAL DIGITAL FORMAT REGISTRY

The Global Digital Format Registry (GDFR)[8] is being developed as part of a two year project by Harvard University Library. The name is intended to convey both that the Registry will collect representation information from centres around the world and that it will be available as a resource for any repository in the world [Abrams & Flecker 2005; GDFR 2004].

5.1 Objectives

As part of the preparatory work for the GDFR project, somewhere in the region of thirty use cases were gathered from institutional participants, detailing the ways they would expect to use a file format registry. These use cases fell into six different categories: *identifying* or *validating* the format of a file; *looking up the characteristics* of a format, for example to identify automatic metadata extraction techniques; *assessing the risks* associated with a format, in particular whether the format is in danger of becoming obsolete; and *determining the optimum migration path*, either between the original format and a display format ('*delivery*'), or between the original format and a similarly functional format ('*transformation*').

From an architectural point of view, the project identified further desiderata for the GDFR. Significantly the project wanted to develop a distributed Registry, in order decrease its reliance on any particular institution or funding stream, and in order to maximize participation.

5.2 Architecture and service models

The GDFR will be based on a hierarchical network structure. Each node in the network will be responsible for:

8. URL: ⟨http://hul.harvard.edu/gdfr/⟩.

- adding, updating and deleting representation information, and then either propagating the changes on to its parent and child nodes as non-vetted information, or passing them to the GDFR editorial review board for vetting;
- propagating vetted representation information from its parent node to its child nodes;
- propagating non-vetted representation information from its parent node to its child nodes, or from one of its child nodes to its parent node and other child nodes.

Both the Library of Congress (see section 3) and the National Archives (see section 4), among others, have indicated a willingness to be GDFR nodes; the Digital Curation Centre (see section 6) is also interested in contributing. One registry will be designated as a root node, with special responsibility for the registration of its immediate child nodes (i.e. top level nodes) and the release of vetted representation information.

The information in the GDFR will be referenced by means of a GDFR namespace, in which each digital file format is given a unique, persistent, public identifier. This identifier will point to the relevant representation information for the file format, and work as either a simple unique identifier (starting `info:gdfr/f/`) or a resolvable address (starting `urn:gdfr:f:`). Similar schemes will enable the identification of classes within the GDFR ontology (`gdfr/c/`) and of GDFR nodes (`gdfr/r/`). The GDFR namespace will be administered by the root node.

The services that the GDFR will offer externally, in addition to providing representation information format-by-format, will include bulk delivery of representation information, and notifications of key format-related developments such as impending obsolescence. Each GDFR node will also publicise, or at very least make discoverable, its capabilities, policies and coverage.

5.3 Data model

The data model for GDFR encodes representation information as a network of cross-referencing properties and attributes. The two main properties in the model are Format and GDFR, and both refer to a number of subsidiary properties. The Format property has the following attributes:

- *Identifier.* The unique identifier for the format within the GDFR namespace, represented as a Cognomen of type *GDFRFormat*.
- *Alias.* Any identifiers for the format from other namespaces can be supplied as aliases using the corresponding type of Cognomen.
- *Description.* A string describing the format.
- *Version.* A string identifying the version.
- *Author.* The author of the format, described using the Agent property. If the author is a person, the Person property (an special case of Agent) is used instead.
- *Owner* and *Maintainer.* The legal owner and maintainer of the format are given using the Authority property, which references an Agent.
- *Classification.* An ontological class may be assigned to the format using a Cognomen of type *GDFRClass*.
- *Relationship.* If, for example, the format is a subtype, supertype, previous version or subsequent version of another format, this can be noted using the FormatRelation property, a special case of the Relation property.

- *Specification.* The specification for a format can be referenced using the Document property; the accessibility or otherwise of a Document is documented by the Access property.
- *Signature.* If a format has a characteristic file extension or Mac OS data type, this can be recorded as an ExternalSignature. If a format has a format declaration built into its syntax, this can be recorded as an InternalSignature. Both of these are special cases of the Signature property.
- *Application.* Any application that can read the format should be declared using the Application property. This latter property describes the application's capabilities with respect to the format through the Process property, and declares any hardware dependencies by means of a Platform property. Software dependencies are declared by referencing further applications.
- *Provenance.* The provenance of the format can be recorded using the Event property.
- *Note.* An optional informative string.
- *LastModified.* A string containing the modification date and time.

The GDFR property has the following attributes:

- *Version* and *Date.* Strings identifying the version and build date respectively of the registry code base and data model.
- *Aegis.* The Authority responsible for the registry.
- *ExternalRegistry.* Any external registries known to the GDFR node (e.g. its parent and child nodes in the network) can declared using the Registry property.
- *Ontology.* The ontological classification scheme used by the registry, declared using the Ontology property.
- *Format.* All formats known to the registry are listed using the Format property.

By virtue of being a special case of the Registry property, the GDFR property also has the following attributes:

- *Identifier.* The unique identifier for the registry within the GDFR namespace, represented as a Cognomen of type *GDFRRegistry*.
- *Service* A supported GDFR service (such as approval or synchronisation), declared using the Service property, which in turn includes a declaration of the Interface protocol used by the service.
- *Note* and *LastModified.* Defined as before.

An external Registry can also have two dates associated with it: *LastHarvestedBy*, the date and time at which the external registry last harvested data from the present registry, and *LastHarvest*, the date and time at which the present registry last harvested data from the external registry.

5.4 Timescale

The project team aims to have most of the theoretical aspects of the project — the data model, the architectural model, the network protocol, the editorial process, etc. — finalised by August 2006, and a first-attempt reference implementation in place by February 2007. By January 2008, the root node in the GDFR network should be fully operational, with the rest of the network coming online shortly after.

6 REPRESENTATION INFORMATION REGISTRY/REPOSITORY

The Digital Curation Centre's Representation Information Registry/Repository (RI RegRep)[9] is geared especially towards the needs of the e-Science community, with particular provision for curating experimental data sets. Being a registry/repository, it will not only register various file formats, but also hold copies of format specifications and rendering software. The scope of the RegRep is intended to be broader than just format information: it will also include semantic representation information, such as instrument calibrations, data units and other information necessary to interpret scientific data [Giaretta et al. 2005].

As well as being intelligible to humans, the RegRep will also be available as a basis for automated tools. For example, the RegRep will provide the information necessary to generate migration routes, from one digital medium to another, from one format to another, or from one set of data conventions (such as units) to another.

6.1 Classification scheme

The RI RegRep will be implemented as an ebXML Registry [Fuger et al. 2005*a*;*b*].[10] The model for such a Registry is that digital objects — in this case, files containing representation information and associated resources — are stored as Repository Items (in the repository side) while various forms of metadata are stored as Registry Objects (in the registry side). The metadata concerning Repository Items are primarily recorded as Extrinsic Objects, a subtype of Registry Objects. The RI RegRep's information model extends the ebXML Registry model by defining a subtype of Extrinsic Object called Representation Information.

Representation Information is an aggregation of three further classes: Structure Information, Semantic Information and Other Representation Information. Structure Information aggregates classes concerning the file type and its format specification, while Semantic Information aggregates classes concerning the data dictionary, natural or programming language semantics, and relevant standards. The remaining class, Other Representation Information, aggregates classes concerning relevant software, hardware and storage media, encryption/compression algorithms, audiovisual codecs and printed documentation.

In the interests of flexibility, individual pieces of representation information registered with the RI RegRep may be attached (i.e. present within the repository) or, if this is not possible, referenced externally using one or more persistent, resolvable identifiers. Furthermore, any piece of representation information may be assigned as many classes as are appropriate.

7 XML SCHEMA REGISTRY/REPOSITORIES

XML (eXtensible Markup Language) as a format is a subtype of SGML (Standard Generalised Markup Language) which is itself a subtype of plain text format.[11] In order to maximize the extensibility of the format, there is very little semantic (or presentational) interpretation within the base specification [W3C 2004]; this is left to XML schemata, which define specialist markup languages in terms of XML syntax. This division between pure syntax and application-related syntax is reflected in the distinct terms *well-formed*, implying conformance

9. URL: ⟨http://dev.dcc.rl.ac.uk/twiki/bin/view/Main/DCCRegRepV04⟩.
10. Electronic Business using eXtensible Markup Language. URL: ⟨http://www.ebxml.org/⟩.
11. The standards for XML and SGML do not prescribe a particular character encoding, although XML processors must support the use of UTF-8 and UTF-16 encodings.

with the XML specification, and *valid*, additionally implying conformance to a named XML schema.[12]

Thus, in order to interpret an XML document correctly, it is necessary to have access to the schema used in its creation; and as XML schemata can be expressed purely formally, it is also necessary to have access to some human-readable documentation for the schema, explaining the semantic (or presentational) meaning of the markup. The implication for digital preservationists is that the XML schemata referenced by documents should either be kept with the document, or be made available through a representation information network. Again, the task of keeping such a network stable is eased considerably by the use of XML schema registry/repositories.

As well as helping to preserve the intelligibility of XML documents, XML schema registry/repositories can also assist data creators by giving them access to ready-made schemata. This can both avoid the duplication of effort and increase compatibility between documents; the latter could prove important when writing tools to mine the documents for information at a later date.

In practice, XML schema registry/repositories tend to be rather basic, with each schema represented by a title, description and link to the locally hosted schema document.

7.1 Examples

One of the earliest XML schema registries was BizTalk.org, offered by Microsoft between 1999 and 2002. As the name implies it was focussed on business applications. At one point the site had over 400 schemata in its registry, but they were all declared using its own proprietary language. Microsoft eventually closed it down, claiming that other registries had made the site redundant [Foley 2002; Kennedy 2000]. In 2000, the OASIS consortium set up a more broadly focussed registry as part of its XML.org site.[13] This registry apparently contains around 250 schemata from various fields of industry, business and academia, but has recently become difficult to access [Kennedy 2000].

The US Government is transforming much of its background technology to work around XML, in order to streamline processes, speed up communications and increase accessibility [Sall 2003; 2004; Walker 2005]. Providing XML registries and repositories is part of this programme, with the aim of improving interoperability. While plans for a dedicated, centralised XML registry have been shelved, a number of different registries have been established, including the Department of Defense Metadata Registry and Clearinghouse (with over 3 300 schemata),[14] the XML Registry for the Environmental Information Exchange Network (with over 880 schemata),[15] and the Grant Application XML Registry (with about 160 schemata).[16] Meanwhile, the Component Organization Registry Environment (CORE.gov), intended as a collaboration space for sharing processes and resources between agencies, is taking on the task of collecting XML schemas, although it is unclear whether this environment will have the same utility as a dedicated XML registry [Jackson 2004; 2005].[17]

12. In the terminology of this paper, XML schemata can be written in several languages, the standard two being the Document Type Definition (DTD, defined at ⟨http://www.w3.org/TR/REC-xml/#dt-doctype⟩ and the XML Schema (defined at ⟨http://www.w3.org/XML/Schema#dev⟩).
13. URL: ⟨http://www.xml.org/xml/registry.jsp⟩. Accessed via Wayback Machine: ⟨http://www.archive.org/⟩.
14. URL: ⟨https://metadata.dod.mil/mdrPortal/appmanager/mdr/mdr⟩
15. URL: ⟨http://iaspub.epa.gov/emg/xmlsearch$.startup⟩.
16. URL: ⟨http://apply.grants.gov/system/MetaGrantApplication⟩.
17. URL: ⟨http://collab.core.gov/CommunityBrowser.aspx?id=2166⟩.

Outside the US, there does not appear to be much activity in government XML registries, although the Hong Kong Government has one.[18] Much of the recent work in this area has been centred on specific topics. For example, the Open Geospatial Consortium hosts around 280 schemata for GIS data in its registries,[19] while the HR-XML Consortium hosts around 30 schemata for Human Resources applications.[20]

8 CONCLUSION

This paper has looked at a number of different implementations of file format registries. Some provide very basic levels of detail (such as Wotsit's Format and the File Format Encyclopedia from section 2), while others give much more comprehensive information. Some are just human-readable, while others are machine readable as well. Some are merely registries, while some have an additional repository element. All of them fulfil a need of a particular community of users.

The subset of registries dealing exclusively with XML schemata are much more uniform in approach, and tend towards a basic filestore implementation. This is reasonable given that XML schemata are written in a formal markup language that can be interpreted by automated tools. Differences between such registry/repositories tend to be on popularity and longevity, with the more successful registry/repositories catering for communities committed to interoperability.

As most of the registries examined are still in their early stages, it is impossible to comment on their long-term impact, although they appear to be promising tools. A similar effort for file formats specific to engineering would appear equally promising, especially if implemented as a shared resource.

9 ACKNOWLEDGEMENTS

This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) and the Economic and Social Research Council (ESRC) under Grant Numbers EP/C534220/1 and RES-331-27-0006.

REFERENCES

Abrams, Stephen & Dale Flecker. 2005. 'A proposal for a global digital format registry.' URL: ⟨http://hul.harvard.edu/gdfr/documents/Proposal-2005-09-29.doc⟩.

Adobe. 2005. 'XMP specification.' Adobe Systems Incorporated. URL: ⟨http://partners.adobe.com/public/developer/en/xmp/sdk/xmpspecification.pdf⟩.

ANSI X3.4-1986. '7-bit American national standard code for information interchange.'

Arms, Caroline R. & Carl Fleischhauer. 2005. 'Digital formats: Factors for sustainability, functionality, and quality.' In: *IS&T Archiving Conference*. Washington, DC: Society for Imaging Science and Technology. 26–29 Apr. URL: ⟨http://memory.loc.gov/ammem/techdocs/digform/Formats_IST05_paper.pdf⟩.

Brown, Adrian. 2004. *PRONOM 4 User Requirements*. Kew: The National Archives. URL: ⟨http://www.nationalarchives.gov.uk/aboutapps/fileformat/pdf/pronom_4_user_reqs.pdf⟩.

———. 2005*a*. *PRONOM 4 Information Model*. Kew: The National Archives. URL: ⟨http://www.nationalarchives.gov.uk/aboutapps/fileformat/pdf/pronom_4_info_model.pdf⟩.

18. URL: ⟨http://www.xml.gov.hk/⟩.
19. URLs: ⟨http://schemas.opengis.net/⟩ and ⟨http://www.opengeospatial.org/specs/?page=specs⟩.
20. URL: ⟨http://hr-xml.org/channels/projects_main.cfm⟩.

———. 2005*b*. 'The PRONOM PUID scheme: A scheme of persistent unique identifiers for representation information.' Digital Preservation Technical Paper 2, The National Archives, Kew. URL: ⟨http://www.nationalarchives.gov.uk/aboutapps/pronom/pdf/pronom_unique_identifier_scheme.pdf⟩.

CCSDS. 2002. 'Reference model for an Open Archival Information System (OAIS).' Blue Book CCSDS 650.0-B-1, Consultative Committee for Space Data Systems. Also published as ISO 14721:2003. URL: ⟨http://www.ccsds.org/documents/650x0b1.pdf⟩.

Foley, Mary Jo. 2002. 'Microsoft shutters BizTalk.org.' *eWeek*. Online news article. URL: ⟨http://www.eweek.com/article2/0,3959,494981,00.asp⟩.

Fuger, Sally; Farrukh Najmi; & Nikola Stojanovic. 2005*a*. 'ebXML registry information model version 3.0.' Approved draft specification, OASIS. URL: ⟨http://docs.oasis-open.org/regrep/v3.0/regrep-3.0-os.zip⟩.

———. 2005*b*. 'ebXML registry services and protocols version 3.0.' Approved draft specification, OASIS. URL: ⟨http://docs.oasis-open.org/regrep/v3.0/regrep-3.0-os.zip⟩.

GDFR. 2004. 'Global digital format registry data model v4.' URL: ⟨http://hul.harvard.edu/gdfr/documents/DataModel-v4-2004-01-12.doc⟩.

Giaretta, David; Manjula Patel; Adam Rusbridge; Stephen Rankin; & Brian McIllwrath. 2005. 'Supporting e-Research using representation information.' In: *Proceedings of the UK e-Science All Hands Meeting*. Nottingham. 19–22 Sep. URL: ⟨http://www.allhands.org.uk/2005/proceedings/papers/447.pdf⟩.

ISO/IEC 15948:2004. 'Information technology — Computer graphics and image processing — Portable Network Graphics (PNG): Functional specification.' URL: ⟨http://www.w3.org/TR/PNG/⟩.

Jackson, Joab. 2004. 'Would a governmentwide XML schema registry cut duplication?' *Government Computer News* 23(16). ISSN 0738-4300. URL: ⟨http://www.gcn.com/print/23_16/26367-1.html⟩.

———. 2005. 'An XML registry is key to sharing data.' *Government Computer News* 24(3). ISSN 0738-4300. URL: ⟨http://www.gcn.com/print/24_3/35006-1.html⟩.

Kennedy, Dianne. 2000. 'XML.org and BizTalk.org make peace?' *XML Files* (22). URL: ⟨http://www.gca.org/whats_xml/xml_files/issue22/edit.htm⟩.

Leeds University. 2003. 'Survey and assessment of sources of information on file formats and software documentation.' Final report, Representation and Rendering Project, University of Leeds. URL: ⟨http://www.jisc.ac.uk/uploaded_documents/FileFormatsreport.pdf⟩.

OASIS. 2005. 'OpenDocument format for office applications (OpenDocument) v1.0.' Under development as ISO/IEC DIS 26300. URL: ⟨http://www.oasis-open.org/committees/download.php/12572/OpenDocument-v1.0-os.pdf⟩.

RFC 2046. 'Multipurpose Internet Mail Extensions (MIME) part two: Media types.' URL: ⟨http://www.isi.edu/in-notes/rfc2046.txt⟩.

Sall, Kenneth. 2003. 'How the US Federal Government is using XML.' In: *XML Conference and Exposition*. Philadelphia, PA. 7–12 Dec. URL: ⟨http://www.idealliance.org/papers/dx_xml03/papers/05-01-04/05-01-04.html⟩.

———. 2004. 'How the US Federal Government is using XML: One year later.' In: *XML Conference and Exhibition*. Washington, DC. 15–19 Nov. URL: ⟨http://www.idealliance.org/proceedings/xml04/papers/150/How-US-Govt-Using-XML-1YL.html⟩.

Unicode Consortium. 2003. *The Unicode Standard, Version 4.0*. Boston: Addison-Wesley. ISBN 0-321-18578-1. Also published as ISO/IEC 10646:2003. URL: ⟨http://www.unicode.org/versions/Unicode4.0.0/⟩.

W3C. 2004. 'Extensible markup language (xml) 1.1.' URL: ⟨http://www.w3.org/TR/xml11/⟩.

Walker, Richard W. 2005. 'XML excises the Army's ancient forms system.' *Government Computer News* 24(3). ISSN 0738-4300. URL: ⟨http://www.gcn.com/print/24_3/35006-1.html⟩.

All links were correct on 31 March 2006.