Automatic Metadata Generation – Use Cases

File Format Metadata (Definitive for Preservation)

Milena Dobreva, Yunhyong Kim

Why are file formats important?

This report discusses the file format of digital objects as an essential part of the metadata required for management of the digital object within the digital preservation lifecycle[1].

The notion of a file format in itself is somewhat fluid and can include a variety of aspects relating to the digital object. Here, we are adopting the definition that a file format is "The internal structure and encoding of a digital object, which allows it to be processed, or to be rendered in human-accessible form." (see Brown, A., 2006b. , p.4). Simply put, a digital object is stored on the computer as a bitstream consisting of 1's and 0's and the file format information enables the computer to convert the stream correctly into information readable by humans or another agent. The role of the file format can be viewed to be akin to the tokenisation, syntax and parsing rules required in understanding natural language text expressed as a string of characters.

The file format affects several elements within the digital preservation lifecycle:

1. As part of the ***ingest process*** to include a digital object in a collection, metadata on file format should form a part of the Submission Information Package (SIP) in the Open Archive Information System (OAIS) reference model[2]. If they are supplied by the producer, they can be validated[3]. If they are not supplied by the producer, they will have to be generated.[4]
2. For the purpose of ***access to content***, it is essential to be able to reproduce the information embodied within a digital object (e.g. by ensuring correct syntax is used for the stored bitstream) and to be able to reproduce the information so that it is as close as possible to its initial instantiation (e.g. formatting of a text document, colour of an image, fidelity of audio clips). The availability of sufficient metadata regarding the file format is crucial for this.
3. In ***using*** and ***re-using*** the objects, the functionalities available at the time of its first instantiation should ideally be preserved, e.g. in the way elements of the object can be viewed, printed, copied and pasted, displayed and contextualised. The file format and its compatibility with hardware and software environments are integral to the availability of these functionalities.
4. ***Transferability*** (e.g. via email and internet download) and **interoperability** (e.g. platform compatibility) are reliant on the method of encoding and compression used for the object, an aspect of file format.
5. Within ***preservation planning*** it is essential to know the formats of the stored digital objects in a digital archive. Evaluating the viability of preservation actions (e.g. migration) and identifying adequate preservation tools (e.g. format conversion tools) depend on the file formats of the digital objects. Best file formats for the storage of objects need to be decided on the basis of preservation purposes (for example, formats using lossless compression such as TIFF are recommended as an alternative to formats using lossy compression such as the older versions of the JPEG format). Sustained risk assessment with respect to new emerging file formats is essential as well (e.g. assessment of how the new

[1] See digital curation lifecycle at http://www.dcc.ac.uk/lifecycle-model/. Please note that this and all susbsequently mentioned web resources were accessed on July 3, 2009.

[2] ISO/IEC 14721, http://www.ccsds.org/documents/650x0b1.pdf

[3] Compare use case in http://www.gdfr.info/docs/use_cases/nyu-1.pdf

[4] Compare use case in http://www.gdfr.info/docs/use_cases/harvard-1.pdf

JPEG 2000 format compares to TIFF).

File format metadata support the assessment of security risks that might threaten digital collections. For example, some media file formats include extra metadata in the file headers which might result in infecting the computer with malicious content. File formats are also used to differentiate executables from non-executables, as a means of detecting or predicting possible viruses.

Metadata included implicitly in the files could lead to inadvertent disclosure of private information and could be a crucial factor in legal disclosure (e.g. in a case of legal disclosure, it is important to understand whether a Excel spreadsheet or an image/printout of the spreadsheet would meet legal requirements). Format information has been also known to play a role in digital forensics (e.g. metadata about the computer that produced the document - information included in the case of some formats - might lead to the identification of a hacker).

The importance of file formats in establishing trustworthiness within a repository is clearly expressed in the Trustworthy Repositories Audit & Certification (TRAC)[5] audit criteria checklist:

> *B2.7 Repository demonstrates that it has access to necessary tools and resources to establish authoritative semantic or technical context of the digital objects it contains (i.e., access to appropriate international Representation Information and format registries).*
>
> *B2.8 Repository records/registers Representation Information (including formats) ingested.*
>
> *B3.2 Repository has mechanisms in place for monitoring and notification when Representation Information (including formats) approaches obsolescence or is no longer viable.*

The examples presented above show clearly that the identification of file formats of digital object in a repository is central to every aspect of managing, curating, and preserving digital information.

The best practices, however, are not consistently presented and identified within the preservation community. This is illustrated by the fact that the German repository checklist (The nestor Catalogue of Criteria for Trusted Digital Repository Evaluation and Certification)[6], in contrast to TRAC does not impose any explicit criteria related to the file formats.

Even with best practices adequately identified, it is not straightforward to collect file format metadata. Automated generation of file format relies on a combination of elements such as file extensions, magic numbers, internal metadata, file headers, codes within the file system, identifiers, and external metadata (See section Tools). However, none of these have been applied consistently and reliably, and there is no guaranteed method of extracting the format information from the object itself.

This report briefly describes sample scenarios that highlight the importance of file format metadata (see section Scenarios) and summarise the initiatives within the digital preservation and curation community that have been proposed to support the development of methods to identify file formats, and to encourage registering format specification at *format registries* (see sections Tools and Standards below). Based on this, we attempt to raise a few areas that future projects might address in defining their objectives (see Key Issues and Recommendations) in view of what has been discussed in this report.

[5] Trustworthy Repositories Audit & Certification: Criteria and Checklist, An RLG-OCLC Report, 2002. http://www.crl.edu/PDF/trac.pdf

[6] http://www.ils.unc.edu/tibbo/JCDL2006/Dobratz-JCDLWorkshop2006.pdf

Some Scenarios

Scenario 1.

Title	**Resource discovery and use in portals with multiple content contributors**
Author	Milena Dobreva and Yunhyong Kim
Narrative	Leslie is a student and she is working on a course assignment. She has to prepare a presentation on historical bells. Leslie looks at the resources in Europeana to select audio samples for her presentation. Her search for "bells" returns about 3000 resources on bells and 16 of those are grouped under "sounds". When she examines these resources she finds two types of thumbnails which are visually different. The first thumbnail can be "played" by clicking on it, but, the objects represented by the second type of thumbnail can not be played. Leslie wonders why this happens and looks at the descriptions of several objects represented by the thumbnail. She finds a term "Format" in the descriptions of the objects. She sees that the objects she can not play come with "**Format: audio/x-mpeg3**" (which Leslie recognises as one particular audio-format). Some of the digital objects she can play have values like "**Format: 00:07:15**" (which seems to be the length of the recording) or "**Format: 00:26:32; electronic**" (which combines a length and genre but is not the technical file format information that would help Leslie identify the application able to play the accompanying audio file). Leslie is not familiar how resources are ingested in Europeana and just expects that the objects she finds there can be rendered using a software application on her machine. This example illustrates one effect of misleading format metadata in the case where objects are being ingested from multiple sources: the metadata were recorded without ensuring that a consistent interpretation and vocabulary is established. Automated control of the values of this metadata element would help to identify which records need to be updated in order to achieve uniformity in the descriptions. The control of the values being supplied for the element format could be easily checked using a controlled vocabulary containing lists of possible formats. In the cases incorrect values are assigned, the format type could be supplied by applying a tool like DROID. Even if the value of the format element seems to be correct, the file format still needs to be checked and validated. If this portal wants to meet the requirements of TRAC, it should also be able to provide evidence that it meets the TRAC criteria B2.7, B2.8 and B3.2.

Scenario 2.

Title	**Finding a tool which could make old device-specific files usable**
Author	Milena Dobreva and Yunhyong Kim
Narrative	Steve is a technical editor and he wants to print and extract text from newsletters he had been publishing over a decade ago. The original desktop publishing files have not survived but he has an archive of PRN files. Unfortunately, the files can not be printed directly from these files because the PRN files had been produced for a specific printer type. Steve does not remember what printer he used back at the time the files were created, and he wonders how he can recover this information, and, further, how he

	might print the files using his current printer. In an attempt to find guidance regarding what to do, Steve checks the File-Extensions website (http://www.file-extensions.org/) and discovers that there are 11 file types which could have the same extension: Calcomp Raster bitmap graphics, DataCAD Windows printer file, Generally printer output file, HP Printer Control Language file, PostScript file, PostScript file, Printer driver (Signature), Text file (Lotus 1-2-3/Symphony), XYWrite printer driver, Plan de Negocio file, DataFlex graphic device driver. The file extension is not sufficient to identify the file format. The file format used for Steve's files was bound to a particular hardware device. The device itself may be already obsolete and difficult to find. The device may require an obsolete driver. A record with the full specifications of the file format including any dependencies on hardware, operating systems, and software would have aided Steve in generating the appropriate metadata to read and print the files. In the current state Steve has the option to try to identify whether the file is binary or text and try to parse the PRN file.

Scenario 3.

Title	**Using an executable file over time**
Author	Milena Dobreva and Yunhyong Kim
Narrative	Moira is a researcher working on a digital humanities project and wants to illustrate a sequence of events in time using a dynamic timeline. She discovers a widget which can produce dynamic timelines. It looks simple because an XML file is used to represent the data. To use the widget Moira downloads a component which should help her to construct her timeline. At this stage Moira realizes that the executable file needs to be created as a Java class file which uses the XML data. Moira understands how to present her data in XML but is not familiar with Java. At this point she asks herself, what will happen with her timeline if this particular programming language will change: she thinks that XML data will continue to be usable but she is not sure about Java libraries. The combination of the data in mark-up language and an executable file is an ad-hoc compound structure; to address these two components without making sure that they can work together would not solve the problem.

Scenario 4.

Title	**Open standards in automatic generation file format metadata for validating and repairing audio files**
Author	Yunhyong Kim and Milena Dobreva
Narrative	Anonymous post at Musepack Forum: *"Once upon a time I got sick and tired of broken mp3 files floating around and wrote a small utility called mp3ck<http://mp3ck.sf.net> that just parsed MPEG frames and ID3 tags one by one to see if their stream was continuous. The utility made no attempt to actually decode the stream, but it proved to catch most defects that resulted from broken transfers over the net, buggy FTP and HTTP servers, etc.*

	With more and more files in Musepack format appearing around, I'd like to extend my little utility to verify such files, too, if possible at all. However, I failed to find a document on the format except for the source code. Is there any out there? Of course, I mean a rather cursory description allowing one just to parse the stream." This example illustrates a case where the open standards of a file format and open source of the rendering software supported the generation of metadata regarding the continuity of frames in the audio file to support the user in validating the integrity of a file. The user, however, failed to use the same method for the Musepack format, because this format is not an open standard, even though the software for rendering files in this format is an open source code.

Tools

File format identification is hindered by the abundance of file formats currently in use and continuously being created. Not only are the formats of digital objects numerous, but every format can have different versions. The complexity of formats also differs ranging from flat file formats to wrappers and compound formats. The documentations on the formats are not unified and in some cases – like in the case of proprietary formats - might not be disclosed to the general public. Further, there is no generally accepted recommendation on what information regarding the file format needs to be stored in order to guarantee correct future use of the digital object.

The number of file formats that are now in use has not been determined. A number of bodies and projects are collecting information on existing formats (see e.g. the Alphabetical list of File Formats[7], the File format encyclopedia[8], or The Digital Formats Web site of the Library of Congress[9]). The most widely-spread classification of format types currently is MIME[10]. This classification of file formats is done from the point of view of use within the Internet and is grouped in the following main digital object types: application, audio, example[11], image, message, model, multipart, text, and video. MIME does not express versioning differences and is not considered sufficiently complete for long-term preservation purposes. As a way of easily tracking formats, several digital preservation community projects have developed *format registries*.

Format Registries

A format registry is a collection of records that characterize existing file formats. For example, a file format entry could include name and version number, characterization elements and links presenting dependencies with other formats. There is no consensus on how this information should be structured. Three examples of format registries with different approaches are **PRONOM[12]**, **Global Digital Format Registry (GDFR)[13]** and **IBM Preservation Manager[14]**. Most of the current advanced technologies in automated file format identification rely on some information from an internal or external format registry.

The task of **automated generation of file format data** can be considered as an outcome of the **file identification** and **validation** tasks. The simplest mechanism for file

[7] http://www.digitalpreservation.gov/formats/fdd/browse_list.shtml

[8] http://pipin.tmd.ns.ac.yu/extra/fileformat/

[9] http://www.digitalpreservation.gov/formats/intro/intro.shtml

[10] Multipurpose Internet Mail Extensions (MIME); the information on MIME types is provided by the Internet Assigned Numbers Authority (IANA), http://www.iana.org/assignments/media-types/

[11] According to IANA, "The 'example' media type is used for examples. Any subtype following the media type syntax may be used in those examples."

[12] http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

[13] http://www.gdfr.info/

[14] http://www-05.ibm.com/nl/dias/preservatiomanager.html

identification is to analyse the file extension and consult a **registry of file extensions** (see e.g. **File Extensions**[15]). One problem with the use of file extensions is that they are not unique - e.g. the search for the popular extension DOC returns 14 possible formats with the same extension. In addition, the users have the freedom to create their own extensions and change the existing ones; thus any judgment based on the file extension can not be trusted.

Metadata Extraction tool[16]

The **Metadata Extraction tool** developed in the National Library of New Zealand does not extract information on the file type but extracts other technical preservation metadata with respect to the file type. The formats currently supported are:

- Images: BMP, GIF, JPEG and TIFF.
- Office documents: MS Word (version 2, 6), Word Perfect, Open Office (version 1), MS Works, MS Excel, MS PowerPoint, and PDF.
- Audio and Video: WAV and MP3.
- Markup languages: HTML and XML.

This approach results in extracting various types of elements which are used within the digital preservation metadata schema, but covers a very narrow set of popular formats.

DROID[17]

DROID was developed by the National Archives. It performs file identification using the PRONOM file format registry. This registry represents each format (assigned a PRONOM Unique Identifier) using signatures that are internal and external to the actual bitstream as a "collections of characteristics which may be used to indicate the format of a digital object" (see Brown A., 2006 a, p. 6). One drawback of the use of format signatures is the level of granularity in detail: if it is not sufficient, files in different formats can be wrongly construed to be the same format.

DROID identifies the file format but does not perform format validation (i.e. to determine whether the file is of the type that it purports to be), nor does it perform any characterization of file formats.

JHOVE[18]

JHOVE (JSTOR Harvard Object Validation Environment) was developed by JSTOR and Harvard University Library. The tool can be used validate and characterise identified formats. The initial JHOVE distribution includes the following standard modules. AIFF, ASCII, BYTESTREAM, GIF, HTML, JPEG, JPEG 2000, PDF, TIFF, UTF-8, WAVE, XML.

Standards

Conforming to recognised standards improves consistency in format registries and format characterisation as well as format validation. There are three types of standards which are relevant:

- file format specification standards;
- preservation metadata standards;
- registry practice standards.

Below we have listed examples of format specifications that have officially accepted by standardisation bodies such as ISO, IEC, ANSI, IEEE as well as those developed and recognised by professional communities. These lists include examples, they are not meant

[15] http://www.file-extensions.org/
[16] http://meta-extractor.sourceforge.net/
[17] http://droid.sourceforge.net/wiki/index.php/Introduction
[18] http://hul.harvard.edu/jhove/index.html

to be exhaustive.

1. Standards and specifications of file formats

- ISO 32000-1*, Document management – Portable document format – Part 1: PDF 1.7*
- ISO/IEC 15948:2003 (E)[19], Portable Network Graphics (PNG) Specification (Second Edition). Information technology — Computer graphics and image processing — Portable Network Graphics (PNG): Functional specification.
- MPEG-7[20] (ISO/IEC JTC1/SC29/WG11) is a standard for describing general multimedia content data.
- MPEG-21[21] (Multimedia Framework (ISO/IEC 21000) was developed to address the need for an overarching framework to ensure interoperability of digital multimedia objects.
- GZIP file format specification v 4.3.[22] This is an example of a specification which defines a lossless compressed data format compatible with a particular utility, GZIP.

2. Technical Preservation Metadata

These are data dictionaries and metadata standards specifying technical metadata requirements for preservation of digital objects:

- PREMIS Data Dictionary[23], specifying technical and descriptive metadata specifically aimed for the purpose of digital preservation.
- LMER Long-term preservation Metadata for Electronic Resources[24] builds XML schemas for technical metadata that supports digital preservation.
- National Library of New Zealand Metadata Standards Framework - Preservation metadata data model.[25]

3. Metadata registries standards

These include standards similar to the Standards for Information Technology – Metadata registries (MDR), initiated by ISO since 1994 (with various parts released in different subsequent years) to address the existence of multiple metadata standards:

- 11179-1: Framework (this part of ISO/IEC 11179 introduces and discusses fundamental ideas of data elements, value domains, data element concepts, conceptual domains, and classification schemas essential to the understanding of this set of standards).
- 11179-2: Classification (this part of ISO/IEC 11179 provides a conceptual model for managing classification schemas).
- 11179-3: Registry metamodel and basic attributes (specifies a conceptual model for a metadata registry, and a set of basic attributes for metadata for use when a full registry solution is not needed).
- 11179-4: Formulation of data definition (provides guidance on how to develop unambiguous data definitions).
- 11179-5: Naming and identification principles (provides guidance for the identification of administered items.
- 11179-6: Registration (provides instruction on how a registration applicant may register a data item with a central Registration Authority and the allocation of unique identifiers for each data item).

The metadata registries are created for a particular application domain; we were not able to identify an existing registry on preservation metadata. The format registries listed above

[19] http://www.w3.org/TR/PNG/
[20] http://www.chiariglione.org/mpeg
[21] http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm
[22] http://www.isi.edu/in-notes/rfc1952.pdf
[23] http://www.loc.gov/standards/premis/v2/premis-2-0.pdf
[24] LMER description and LMER schema: http://www.d-nb.de/eng/standards/lmer/lmer.htm
[25] http://www.natlib.govt.nz/catalogues/library-documents/downloadpage.2007-02-15.6613783926

are not linked to this standard – they model file format structures, but a metadata registry for them is still not in place.

Key Issues

Obstacles in successful implementation of the automatic generation of file format metadata could include:

Interoperability Issues: Failure to create a common vocabulary, and/or standardised and consistent metadata and file format registries. Note, for example, that metadata could be stored separately or be embedded, i.e., encoded in the digital object. One popular example of a technological solution which allows to embed metadata into the file is Adobe's Extensible Metadata Platform (XMP)[26]. Further, different information levels and granularity associated with file format registries make it difficult to make the most of its role in file format characterization and validation in addition to format identification.

Note that, a basic difficulty in the field of file formats is the big number of formats and the fact that a commonly accepted model on format registries is not yet developed.

Issues related to IPR: The automatic generation of file format metadata is highly dependent on a clear understanding of the syntax being used in the description of the file format as well the way in which the software is rendering the information within the object. If the formats are proprietary, the integration of the tool which can recognize the format becomes more difficult. The difficulty already caused by the lack of documentation on file format structure can be exacerbated by the obscurity of proprietary formats.

Insufficient platforms for experimentation, evaluation, and comparison: the tools that have been developed for format identification, validation and characterization have not been compared on a laboratory controlled environment using a consolidated dataset to meet the requirements of different user scenarios. To bring the tool development to maturity and to evaluate the tools on the basis of preservation quality metadata and to compare tools using competitive evaluation we need to use testbeds such as the environment developed by Preservation and Long-term Access through NETworked Services (PLANETS)[27].

Lack of consensus in professional domain about best practice: for example**,** the metadata schema of the PREsrevation Metadata Implementation Strategies (PREMIS)[28] working group uses an extensive description of format registries as part of their standard; the semantic unit **format** has the components **formatDesignation** and **formatRegistry**. The semantic unit **formatRegistry,** in turn, has the components **formatRegistryName**, **formatRegistryKey** and **formatRegistryRole**. In contrast to PREMIS, The metadata schema at the German national Library LMER[29] includes an element **format** which comprises a single string value pointing to the format identifier used within an external repository.

Recommendations

1. **Develop a set of tools which will improve the quality control of metadata element values during ingest.**
   The file format value either is not entered in advance, or might follow different interpretations. In order to improve the homogeneity in digital repositories w.r.t. the metadata quality, in the cases of ingest of digital objects it would be helpful to control the values related to file formats and to extract file formats where values had not been supplied. To make this information usable it would be also helpful to provide a connection to a specific metadata registry, e.g. PRONOM.

[26] http://www.adobe.com/products/xmp/

[27] http://www.planets-project.eu/

[28] PREservation Metadata Implementation Strategies working group documents: see *PREMIS Data Dictionary for Preservation Metadata, v. 2* (March 2008) http://www.oclc.org/research/projects/pmwg/

[29] LMER - Long-term preservation Metadata for Electronic Resources, German National Library, http://www.d-nb.de/eng/standards/lmer/lmer.htm

2. **Encourage the use of open standards and open source and build action plans for formats that are stored within the repository.**
   Open standards and source codes refer to standards and codes (respectively) for which the technical specifications are made publicly available. The transparency of standards and source codes used in creating and reading file formats enables future users to re-create the environment or software, if necessary, to access, copy, migrate, display and re-use previously created files. Open standards and source also make it easier to build preservation action plans[30] by making it easier to evaluate the archival quality of the file format, to identify a range of automated methods for normalising and migrating the file, to isolate the characteristics of the file essential for validating and repairing its integrity.

3. **Provide a best practice set of examples in the field of file format metadata.**
   The existence of pragmatic and clear guidance on the use of registries and tools for file format identification and validation would ne of help to the users.

4. **Use testbeds for experimentation, evaluation and comparison of tools being developed** (see section Key Issues for more detail). This would contribute to improve the trustworthyness of the ongoing research and implementation work in digital preservation.

Further reading

Resources on File Formats

- **Alphabetical list of File Formats** (96 formats on 30 June 2009, with listed versions), From: Sustainability of Digital Formats Planning for Library of Congress Collections, http://www.digitalpreservation.gov/formats/fdd/browse_list.shtml
- **File Extensions,** http://www.file-extensions.org/
- **Florida Digital Archives,** http://www.fcla.edu/digitalArchive/formatInfo.htm
- **National Software Reference Library (NSRL)**, http://www.nsrl.nist.gov/index.html
  repository of known software, file profiles, and file signatures for use by law enforcement and other organizations involved with computer forensic investigations. Currently contains over 10,000 software products of various types: benign, malicious, corporate, electronic voting; over 75,000,000 files
- **Universal Preservation Format,** http://info.wgbh.org/upf/
- **Wotsit.org**, http://www.wotsit.org/
  Set of resources on file and data types for programmers.
- **File format encyclopedia**, http://pipin.tmd.ns.ac.yu/extra/fileformat/
- **The Digital Formats Web site, Library of Congress**, http://www.digitalpreservation.gov/formats/intro/intro.shtml

Format Registries

- **Global Digital Format Registry (GDFR),** http://www.gdfr.info/
- **FRED - Format REgistry Demo,** http://tom.library.upenn.edu/fred/
- **IBM Preservation Manager,** http://www-05.ibm.com/nl/dias/preservatiomanager.html
- **PRONOM**, http://www.nationalarchives.gov.uk/PRONOM/Default.aspx
- **TOM - Typed Object Model,** http://tom.library.upenn.edu/

File Format Identifiers & Validators

- **File Format Identification wiki**, http://www.forensicswiki.org/wiki/File_Format_Identification
- **DROID** - Digital Record Object Identification, http://droid.sourceforge.net/wiki/index.php/Introduction
- **JHOVE** - JSTOR/Harvard Object Validation Environment, http://hul.harvard.edu/jhove/index.html
- **File Investigation tools**, http://www.forensicswiki.org/wiki/File_Format_Identification

[30] http://www.fcla.edu/digitalArchive/formatInfo.htm

Publications

Abrams, Seaman: Towards a global digital format registry. In: *69th Congress IFLA 2003*. http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf

Arms, C. & Fleischhauer, C., 2004. *Digital Formats: Factors for Sustainability, Functionality, and Quality*, Library of Congress.

Brown, A., 2006a. *Automatic Format Identification Using PRONOM and DROID*, The National Archives.

Brown, A., 2006b. *The PRONOM PUID Scheme: A scheme of persistent unique identifiers for representation information*, The National Archives.

Calhoun, W. & Coles, D., 2008. Predicting the types of file fragments. *Digital Investigation*, 5, S14-S20.

Chou, C. 2007, Format Identification, Validation, Characterization and Transformation in DAITSS, In *Archiving 2007*, pp. 33-36.

Data Dictionary for Preservation Metadata, 2005. OCLC and RLG.

Dobratz, S., Schoger, A. & Strathmann, S., 2006. The nestor Catalogue of Criteria for Trusted Digital Repository Evaluation and Certification. In *JCDL Workshop*. pp. 6.

Dunckley, M. et al., The Use of File Description Languages for File Format Identification and Validation. In *PV 2007 (October 9-11 2007, Munich)*. pp. 9.

Guercio, M. & Cappiello, C., 2006. *File formats typology and registries for digital preservation*, Urbino: DELOS D6.3.1.

Guttenbrunner, M. et al., 2008. Evaluating Strategies for the Preservation of Console Video Games. In *iPRES 2008: The Fifth International Conference on Preservation of Digital Objects*. London: The British Library.

Lawrence, G.W. et al., 2000. *Risk Management of Digital Information: A File Format Investigation*, Washington, D.C.: Council on Library and Information Resources.

Perkins, R., 1995, File Formats on the Internet, Computer and Geosciences, vol 21, no 6, 775-777 Elsevier

Steinke, T., 2008. Harvester results in a digital preservation system. In *iPRES 2008: The Fifth International Conference on Preservation of Digital Objects*. London: The British Library.

Stevenson, J. & Team, F.T., 2005. *Jorum Preservation Watch Report*, *Survey and assessment of sources of information on file formats and software documentation (Final Report)*, University of Leeds.

Sweetkind-Singer, J., Larsgaard, M.L. & Erwin, T. 2006. Digital Preservation of Geospatial Data. *Library Trends*, 55(2), 304-314.

Authors:
Milena Dobreva and Yunhyong Kim

Dr. Milena Dobreva is a Senior Researcher in the Centre for Digital Library Research of the University of Strathclyde. She is coordinating the work of the Work Package setting the foundations of the reference architecture of the SHAMAN integrated project of FP7. In 2006-2007 she contributed to the activities of the Preservation cluster of the DELOS project in the area of automated metadata extraction the work on the DELOS DLRM (Digital Library Reference Model). She created and served as the founding head of the first specialised Digitisation Centre in Bulgaria (2004-2007), currently part of the Humanities Informatics Department of the Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences. She holds an Academic Award for young researchers for original achievements in the computer representation of mediæval Slavonic texts (Bulgarian Academy of Sciences, 1998). Milena coordinated the project KT-DigiCult-Bg project (Knowledge Transfer in Digitisation of Cultural and Scientific Heritage to Bulgaria), FP6 project MTKD-509754, Marie Curie programme, May 2004-2007, and served as a member of the Executive Board of the National Commission of Bulgaria for UNESCO (2006-2007); the Expert Committee Information and Documentation of the Bulgarian Committee for Standardisation (1998-2006). She is a member of the Europeana v1.0 working group on Standards and Interoperability since 2007. Milena is a Honorary Research Fellow of HATII at the University of Glasgow since 2008.

Dr. Yunhyong Kim is a Research Fellow at the School of Computing, Robert Gordon University. She is involved in the EPSRC funded AutoAdapt Project, developing domain models to enhance information search in the intranet and local collection environment. She has also worked as the Digital Curation Centre (DCC) Resources Researcher at the Humanities Advanced Technology and Information Institute (HATII), University of Glasgow, at which time she engaged in investigating methodologies for testing digital preservation strategies. Her research experiences have focused on methods to automate the extraction of semantic metadata from digital material, as part of the ingest and appraisal processes related to digital repositories. In particular she has presented papers at several international conferences on the role of automated genre classification in semantic metadata extraction.

Contact Details:
Dr. Milena Dobreva
Centre for Digital Library Research (CDLR)
Information Resources Directorate (IRD)
Livingstone Tower 12.12
26 Richmond Street, Glasgow, G1 1XH
Phone: +44 (0) 141 548 4753
Email: milena.dobreva@strath.ac.uk

Dr. Yunhyong Kim
St. Andrew Street
School of Computing
Robert Gordon University
Aberdeen
AB25 1HG
Email: y.kim1@rgu.ac.uk