Cairo tools survey

a survey of tools applicable to the preparation of digital archives for ingest into a preservation repository

21 May 2007

Version 1.0

MANCHESTER 1824
The University of Manchester

WELLCOME LIBRARY

Cairo Project team

Susan Thomas,
Project Manager/Digital Archivist,
Oxford University Library Services

Fran Baker,
Digital Archivist, John Rylands University Library

Renhart Gittens,
Software Engineer,
Oxford University Library Services

Dave Thompson,
Digital Curator, Wellcome Library

Cairo is funded under the 'Tools and innovations' strand of JISC's Repositories and Preservation programme.

JISC

Table of Contents

Introduction to the Cairo tool....3
Introduction to the Cairo tool review....5
Survey of tools relevant to the ingest workflow....5
Establishing tool categories....10
Tool survey results....13
Tool Listing....14
Appendix 1 : Common Open Source Licenses....44

Introduction to the Cairo tool

The aim of the Cairo project is to develop a tool which creates an interface for the ingest workflow, which brings together ingest tools, especially metadata creation tools, into a single coherent, usable and documented tool, which is suitable for use by professional archivists with limited technical competencies. The tool should be capable of processing formats commonly found in personal digital archives and be extensible, so that support for other formats and the metadata they need can be added as necessary. The tool's output should be digital archives that have been subject to ingest processes, together with repository-independent metadata packages in the form of METS files, which document that workflow and record metadata that will provide the basis for long term lifecycle management.

The principal user of the Cairo tool will be an archivist performing the everyday tasks of receiving archival material, preparing it for placement in long-term storage, and running queries or generating reports on work processed by the Cairo tool. The archivist will present an arrangement of digital archives to Cairo whereupon the tool will coordinate an ingest workflow, with a minimum of input from the archivist, resulting in the metadata packages needed for the lifecycle management of the digital archives.

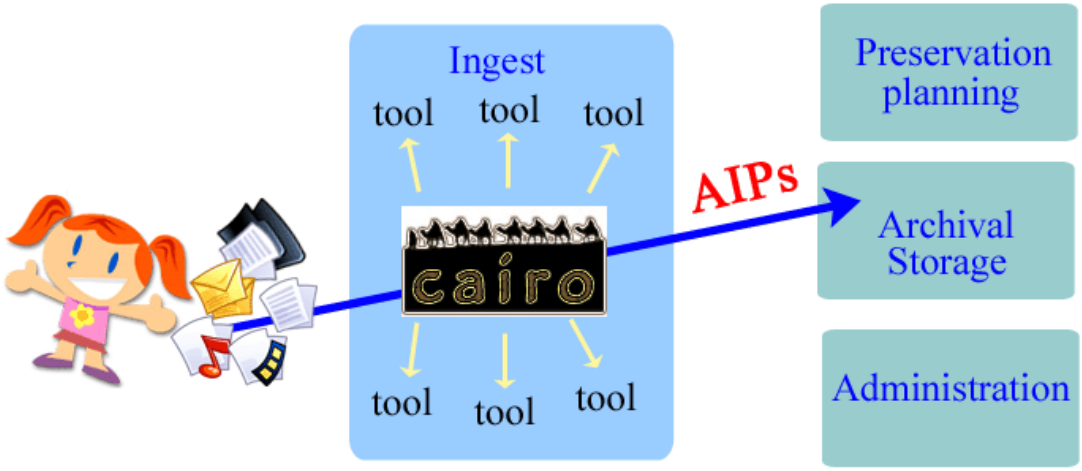

Illustration 1: Archivist presents personal digital archive to Cario, which creates Archival Information Packages (AIPs) for the lifecycle management of the archive and adds them to archival storage.

The digital archives and their metadata may then be presented to some kind of archival storage, perhaps a digital repository system, as Archival Information Packages (AIPs).

Currently, the process of preparing digital archives for ingest requires knowledge of many file formats, ingest-related tools and metadata standards:

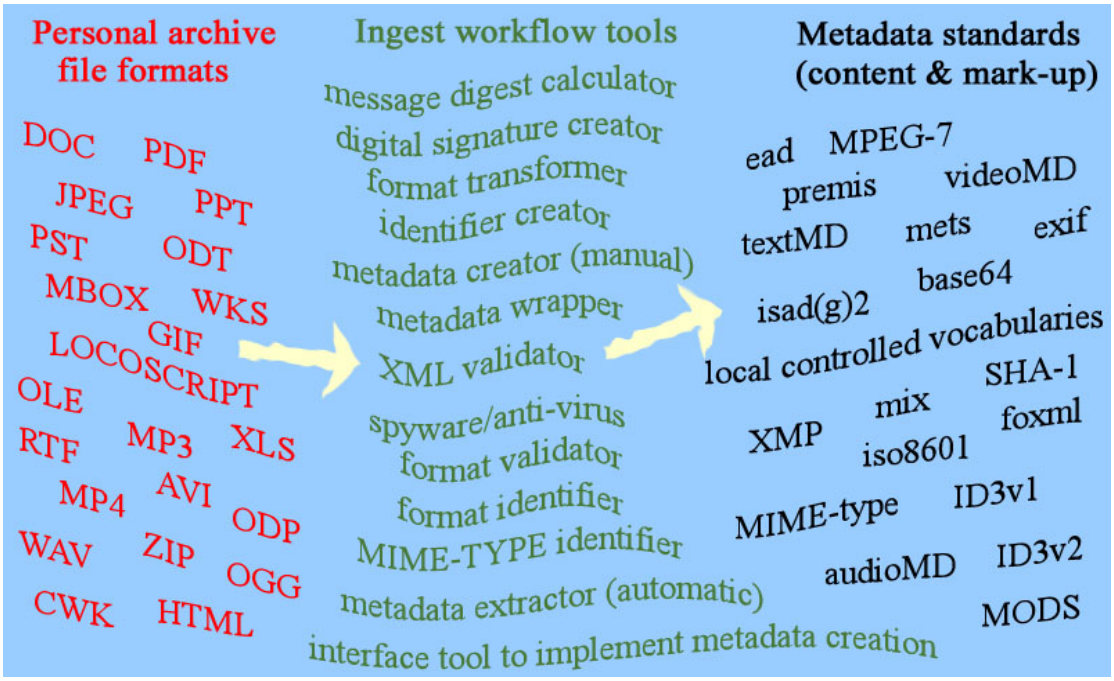

Illustration 2: The range of file formats, ingest tools and metadata standards applicable to the preparation of personal digital archives for preservation.

The quantity and complexity of new knowledge required presents a significant barrier to engaging professional archivists in digital preservation.

The Cairo tool will alleviate this complexity by orchestrating an ingest workflow, composed of several discrete components and standards, from a single user interface. The archivist is therefore not expected to be an expert user of XML, command-line metadata utilities or the means by which the output of individual tool components is aggregated into a standards-based metadata package. Nonetheless, it is advisable for the archivist to learn something about the nature of the preservation metadata that the Cairo tool is creating. It is hoped that Cairo can help in this by providing a gentler introduction to the creation of such metadata than is presently available: users of the Cairo tool will interact with the ingest workflow through selecting options on 'radio buttons', drop-down menus or from pre-determined options taking effect as a consequence of user selections.

The archivist is expected to understand the basic principles of archival practice and especially those principles that relate to authenticity, context and provenance, such as the creation of meaningful hierarchical arrangements of material. These principles are derived from the practice of working with physical materials and must also be applied to their digital equivalents. While the Cairo tool will offer an easy-to-use interface it will not do away with the need for professional archivists to apply specialist skills and understandings.

Further information about the roles of Cairo users is available in the Cairo Use Cases document.

Introduction to the Cairo tool review

The purpose of this tool review is to identify a set of ingest and metadata-related applications, tools or code that could form the components of the overall Cairo tool package. This approach will avoid unnecessary re-invention/replication and leverage a global pool of expertise, saving the Cairo team time and effort. The components selected for use in Cairo must be licensed in such a way that the project is permitted to incorporate them into a larger tool, and modify them as necessary. For this reason, preference will be given to Open Source applications, where license to use the code in this way is granted up front. The project team do not have the resources to seek permissions to utilise software with proprietary or unclear licensing arrangements.

Associated documents

This Tools Survey document should be read in conjunction with other documents designed to inform the development of the Cairo tool:

- Cairo Content Typology Model
- Cairo Use Cases.

Survey of tools relevant to the ingest workflow

Potential Cairo tool components have been identified by a scoping exercise that involved searching open source software repositories such as Sourceforge[1] and Fresh Meat;[2] consulting websites relating to digital repository activity and by conducting broader searches via standard Internet search engines.

Criteria for the identification of potential Cairo components

Criteria for the identification of potential Cairo components were designed to be relatively inclusive at this stage of the project. Hence the tool listing below includes tools that cover many of the same roles or functions, and tools which cannot be used by the Cairo tool (owing to platform or licensing restrictions), but which may prove useful in the ingest context to organisations wishing to use stand-alone tools.

General criteria for the *survey* include applications and tools that:

- are available now;
- are available for public re-use;
- may be open source, 'free to use' or commercially available;
- extract or generate some form of metadata from or about a file object or objects;
- operate on a variety of computing platforms;
- have some basic information about them available.

[1] Sourceforge; http://sourceforge.net
[2] Freshmeat; http://freshmeat.net/

General criteria for the *Cairo tool* have not been fully established at this time, but preference will be given to applications and tools that:

- are available now;
- are clearly licensed under an open source license which permits incorporation into other software and modification;
- extract or generate some form of metadata from or about a file in a form which can be manipulated (such as CSV or XML);
- operate on a variety of computing platforms;
- are well documented;
- scale well;
- are well constructed.

Some of the criteria for the Cairo tool cannot be established within the scope of this survey, as they require developer input. This survey is designed to inform the specification phase of the project, which will see a shortlist of components developed and may include a more thorough examination of the technology base and documentation of the tool components by the project's developers. A more detailed examination of the compatibility of shortlisted components' licenses will also be required.

Problems with recycling software code in a modular tool

The approach of developing a modular tool in which individual components are brought together under a single ‘wrapper’ is sound, but requires care and attention to detail.

One of the greatest problems we anticipate is identifying a suitable combination of modules, which provide comprehensive coverage for formats, ingest processes and metadata types, whilst ensuring that the tool does not become too unwieldy to be practical. Cairo has found that there are few tools developed with a sufficiently general approach to cater for a wide range of file formats, metadata types and sources; and many catering for very specific, often only current, formats.

Another major issue is licensing. The use of inappropriate or unclear licensing has prohibited the inclusion of some potential components. Websites like Sourceforge and FreshMeat, from which many of the potential components were drawn, offer a framework within which open source projects can provide detailed information, including licensing Information, about their work. Even here sufficient detail can be lacking or incomplete.

In addition to functional and legal problems, Cairo has identified some very specific technical issues relating to the ‘recycling’ of potential components. These relate to the design of the Cairo tool and to its distribution and maintenance in the future.

The degree to which issues like these can be resolved will have a great impact on the usability of the tool and its sustainability.

Issues arising from the Cairo tools survey

Issues identified in the scoping exercise can be summarised as follows:

Licensing:

- absence of, or poor information about, licenses;
- variety of licenses;
- incompatibility between licenses, which might prevent the combination of or linking of tools released under different licenses;
- even some open source licenses include requirements which might cause problems for the distribution of software, e.g. clauses which stipulate that the distributor must try to obtain explicit consent to the terms of the license from all recipients;
- poor documentation;
- tool base is international and includes developers representing a diverse cross-section of sectors, organisations or personal interests with different priorities and inconsistent terminologies;
- lack of basic information about potential components, no existing web site or little accompanying documentation;
- many open source developers aim the documentation of their software at developers and documentation which would enable managers and users to assess the value of the software is often lacking.

Metadata output:

- failure of potential components to employ metadata standards where appropriate standards exists, meaning that metadata mapping will be required to determine exactly which tools may be useful;
- inability of otherwise useful tools to output metadata for many files in some neutral form, such as XML or CSV;
- inconsistency, and potential incompatibility, of the output produced by potential components which may lead to difficulties in incorporating output into a single coherent METS file;
- difficulty of assessing the quantity and quality of metadata generated.

Sustainability:

- sustainability of many components is questionable - creators often abandon components and do not extend tool functionality to cater for other formats or format versions;
- lack of explicit commitment to on-going long-term tool/module maintenance in the face of changing formats and platforms;
- some tools have not been maintained beyond initial development and may rely on older versions of technological bases (such as older versions of Java), which may create interoperability problems.

Scalability:

- inability of potentially useful tools to process many files concurrently;
- some tools have not been designed to be integrated into a larger workflow (this is especially true of programs developed solely for MS Windows);

- tendency of potential components to focus on single format or type of format.

Interoperability:

- many tools/applications are written only for a Microsoft Windows environment and are available only as a Windows executable binary;
- inconsistency of code use and inconsistency in coding practice;
- incompatibility of code platforms, or versions thereof, and interoperability issues with other tools written using different code base;
- uncertainty as to whether potential components can be easily distributed, and installed, as part of the Cairo package for local deployment.

Potential impact on source data:

- uncertainty over whether or not any individual component modifies or affects the object upon which it acts.

Formats supported:

- most potential components are current and support modern formats; there is concern that when the formats they support move on, then support for the tool will cease causing problems for archivists who may often be working with legacy digital formats in future.

Potential future maintenance problems:

- determining responsibility for deciding whether to update Cairo when a Cairo component is updated and implementing such updates;
- Cairo may need to undertake a 'technology watch' of its component tools;
- if the Cairo tool is widely adopted by the community, governance arrangements for developing the tool further may be needed;

Security:

- some tools with useful functionality use online updates, or are available over a network; personal archives cannot be sent to third party services such as these, so some means of developing local implementations of these services would be useful.

Mitigating risks

Despite the limitations identified in the scoping exercise the project team believes that sufficient candidate components have been identified that fit the criteria for inclusion. To mitigate some of the risks identified in the course of the survey, Cairo will be rigorous in selecting components that comply with three basic criteria:

1. Tools must have clear and appropriate licensing

Components selected for the Cairo tool must have clear and appropriate licensing which allows the component to be used and modified by a third party, and re-distributed as part of the Cairo package without the risk of adverse follow up. More information about common open source licenses is available in Appendix 1 of this document.

2. Tools must have clear and sufficient documentation

Any potential component must be accompanied by clear and sufficient documentation that allows it to be fully understood, its code based explained and its functions explored. The Cairo project team does not have the resources to develop its own documentation for potential components for which it may not have a full understanding.

3. **Components should enable Cairo to support a minimum level of metadata for most objects and more detailed metadata for a limited number**

The Cairo tool should initially aim to support:

- a basic metadata profile, which might supply generally applicable metadata (including PREMIS metadata,[3] such as file format and hash value) for any kind of file format;
- a limited range of sub-profiles for content types, such as 'image';
- a limited range of sub-sub-profiles with support for more detailed metadata associated with their specific format.

The structure of Cairo's metadata application profile for digital files may therefore resemble the following diagram:

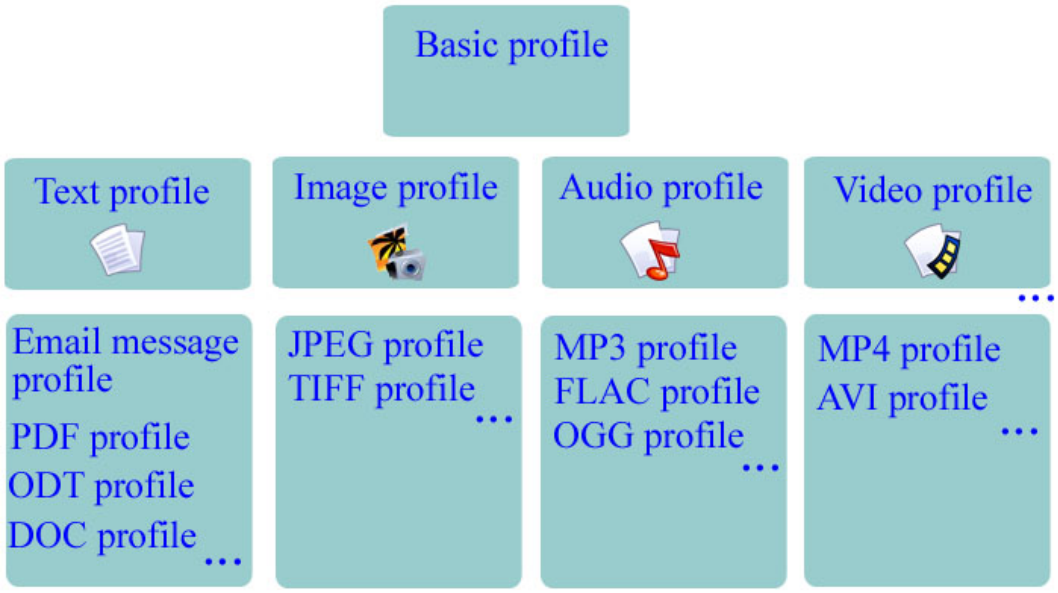

Illustration 3: hierarchy of metadata profiles that might be applied to different file formats.

To remain manageable, the Cairo tool will fix an initial set of file formats and related metadata profiles that the tool will process. An ideal first set would include some or all of the following:

- office suite formats (Appleworks, MS Office, Open Document Format ISO 26300);
- file types common to the Internet (HTML, GIF, PDF etc);
- still and/or moving image formats such as JPEG, TIFF or MPEG;

[3] PREMIS http://www.loc.gov/standards/premis/

- popular legacy formats, though these are not always covered in existing tools which tend to favour current formats.

The final decision as to which formats will be given the most support will be partly reliant on a more technical investigation of the tool components and cannot be determined as yet. Later instances of the Cairo tool may add functionality to cover additional file types, formats or metadata types.

Establishing tool categories

As a starting point, tools have been categorised according to the events they facilitate. To do this, we have used the Event entity of the *PREMIS Data Dictionary 1.0* (PREMIS) which provides a framework for repositories to aggregate metadata about actions taken in respect of objects in the repository's care.[4] PREMIS suggests that: 'Each repository should define its own controlled vocabulary of eventType values'.[5] The standard provides an initial list of suggested event names:

The PREMIS starter list for eventType

PREMIS eventType	PREMIS definition
capture	the process whereby a repository actively obtains an object.
compression	the process of coding data to save storage space or transmission time.
deaccession	the process of removing an object from the inventory of a repository.
decompression	the process of reversing the effects of compression.
decryption	the process of converting encrypted data to plain text.
deletion	the process of removing an object from repository storage.
digital signature validation	the process of determining that a decrypted digital signature matches an expected value.
dissemination	the process of retrieving an object from repository storage and making it available to users.
fixity check	the process of verifying that an object has not been changed in a given period.
ingestion	the process of adding objects to a preservation repository.
message digest calculation	the process by which a message digest ("hash") is created.
migration	a transformation of an object creating a version in a more contemporary format.
normalisation	a transformation of an object creating a version more conducive to preservation.
replication	the process of creating a copy of an object that is, bit- wise, identical to the original.

[4] PREMIS Data Dictionary Version 1.The PREMIS Data Model

[5] PREMIS Data Dictionary Version 1.Event Entity

validation	the process of comparing an object with a standard and noting compliance or exceptions.
virus check	the process of scanning a file for malicious programs.

The PREMIS starter list for an eventType controlled vocabulary refers only to events taking place against digital objects. It should be a basis for an eventType controlled vocabulary for recoding metadata about the events which occur in a digital object's lifecycle, but the granularity available for recording ingest processing activities is not adequate for Cairo's purposes. Cairo aims to record its ingest process in a more granular fashion than the eventType 'Capture' permits. Some (classes of) objects may be subject to different processing tools than others; metadata about these processes and about the manual creation and automatic extraction of metadata is important for authenticity and troubleshooting. The Cairo tool must therefore refine this vocabulary for its own processes, though it can build on the PREMIS starter list.

Proposed CAIRO eventType and tool categorisation

Cairo eventTypes	**Cairo definition**	**Cairo tool categorisation**
assignIdentifier	Assigning a persistent identifier to the material, e.g. Handle System and/or local system identifier	Identifier creator
metadataExtractGeneral	Extracting general metadata that is embedded in an object and common to all formats, e.g. file size. This will be the metadata conforming to Cairo's default base metadata profile.	metadata extractor (automatic)
metadataExtractTypeSpecific	Extracting metadata that is specific to an object type and is embedded in the object, e.g. text or image. This will be metadata conforming to Cairo's sub-type metadata profiles.	metadata extractor (automatic)
metadataExtractFormatSpecific	Extracting metadata that is specific to a format type and is embedded in an object, e.g. TIFF header metadata or photoshop metadata in a .psd file. This will be metadata conforming to Cairo's sub-sub-type metadata profiles.	metadata extractor (automatic)
metadataCreate	Creating metadata by reading file/system information or incorporating metadata added by a human actor.	Metadata creator (manual) Metadata extractor (automatic)
digitalSignatureCreation	Creating a digital signature for addition to an object by the actor carrying out the the ingest process or	Digital signature creator

	using the Cairo tool	
metadataTransform	Take one form of metadata and convert it to another form, e.g. Cairo may normalise dates to the ISO Standard, or transform raw tool output to PREMIS-conformant metadata using XSLTs.	Metadata transformer
metadataValidate	Validate that XML is well-formed; validate that metadata conforms to appropriate XML schemas; validate that metadata values conform to metadata application profile.	Metadata validator
fixityKeyCreation	The creation and assignment of a fixity (Hash) value to an object, e.g. SHA-1	Message digest calculator
formatIdentification	The confirmation, by JHOVE ID, DROID ID or other tools, of a file format	Format identifier
formatValidation	The confirmation that a given file is of a particular format and the recording of that format	Format validator
mimeIdentifcation	The confirmation that a given file is of a particular MIME type and the recording of that type, even if the type is unknown.	MIME-type identifier
malwareCheck	Checking that the object does not contain any malicious spyware or malware that could threaten the integrity of the archive[6]	Spyware/anti-virus
wrapperCreation	The creation of a wrapper or container into which metadata can be placed, e.g. the creation of the METS AIP file to contain disparate XML metadata.	Metadata wrapper
formatTransformation	The migration of an object to a different file format	Format transformer
decompression	The process of reversing the effects of compression. Applied to materials submitted by a donor in compressed formats.	decompressor
decryption	The process of reversing the effects of encryption.	decryptor
n/a	Not associated with an eventType as such, but included as a tool category to denote applications which provide user interfaces/facilitate data entry/navigation – although this doesn't constitute an eventType.	Interface tool to implement metadata creator

[6] For a discussion of Spyware see http://en.wikipedia.org/wiki/Spyware for a discussion of Malware see http://en.wikipedia.org/wiki/Malware

In respect of each of these events, Cairo will record details of the agent (i.e. the tool module being used in the Cairo framework) responsible using the PREMIS agent entity, and METS files representing these agent and event entities will be ingested into our repository with the METS files representing objects and conceptual items such as directories, accessions and collections. Metadata for events should, at a minimum, include those semantic units deemed mandatory for the PREMIS Event entity: eventIdentifier, eventType and eventDateTime.

Tool survey results

The tool survey results are presented in the following format:

Tool	**The name of the tool**
Source URL	A web location where information about the tool can be found or the tool downloaded.
Formats covered	A list of the file formats that a particular tool may be designed to interact with, e.g.: MP3 Microsoft Word JPEG
Technology base	The platform(s) on which the tool was built, e.g.: Java Perl XML
Operating system	The operating system(s) on which the tool can run.
Dependencies	Any technical dependencies the tool may have, e.g. the requirement of a database such as MySQL.
Tool category	The broad function of the tool, e.g.: Metadata creator (manual) Metadata extractor (automatic) Format transformer Tools that perform more than one function are assigned multiple categories.
License	The type of license the tool is available under. Some of the standard open source licenses are outlined in Appendix A and referred to in abbreviated form within the relevant tool entries. Where commercial or specific license arrangements apply, these are outlined briefly in the relevant entry with links to websites containing more information. In some cases licensing information is not readily discernible, and license is given as Unknown/unclear.
Description	A free text description of the tool, based on information taken from project websites.
Output method	For tools that generate metadata, where possible a note of how they are able to output the metadata that has been created.
Maturity	Statement of software maturity, taken from Sourceforge website for relevant software.

Notes	General comments not included elsewhere.

Tool Listing

Tool	**7Train**
Source URL	http://seventrain.sourceforge.net/
Formats covered	Unknown/unclear
Technology base	Java/Saxon/XSL
Operating system	All 32-but MS Windows (95/98/NT/2000/XP) All POSIX (Linux/BSD/Unix/Apple Mac OS X)
Dependencies	
Tool category	Wrapper creator Metadata transformer
License	BSD
Description	7train is an XSLT 2.0 tool for generating METS files from standardised XML input. The tool was originally designed to transform standard XML metadata records exported from the CONTENTdm digital collection management software into METS files which conform to a specific METS profile (designed for a project focusing on digitised images and scanned text objects). However, it can be customised to produce METS files from any kind of standardised XML document, e.g. OAI records.
Output method	XML output in the form of a METS file.
Maturity	
Notes	

Tool	**Ad-Aware**
Source URL	http://www.lavasoftusa.com/products/select_your_product.php
Formats covered	n/a
Technology base	Proprietary
Operating system	Windows 98/98(SE)/ME/NT4 Workstation/NT4 Server/2000 Pro/2000 Server/2003 Server/XP Home/XP Pro/XP Home/XP Professional/XP 64-Bit Edition/Terminal Services
Dependencies	
Tool category	Spyware/anti-virus
License	Commercial: license permits use of one copy of the software on one computer per purchased copy; it prohibits purchaser from modifying, adapting, translating, disassembling, decompiling, reverse engineering or otherwise attempting to discover the source code of the software, and from copying the documentation; purchaser is also forbidden from sub-licensing, assigning, transferring, etc their rights under the license without prior written permission from Lavasoft AB. See http://www.lavasoftusa.com/products/license-

	agreements/SoftwareLicenseAgreement.pdf.
Description	Ad-Aware SE Professional provides a high level of protection against malware and spyware. Its components include: Process-Watch, which allows the user to browse, scan and terminate offending processes which are running; a real-time processing element called Ad-Watch, which runs silently in the background and monitors any malware which tries to install on or modify the system; and an on-demand scanner to detect security threats. It makes use of a new Code Sequence Identification technology, which can identify both new and unknown variants of malware. When spyware is identified, Ad-Aware allows the user to decide whether it should be deleted or kept, rather than automatically deleting it as other security products often do.
Output method	Output to screen.
Maturity	
Notes	To be used alongside, and as well as, anti virus software.

Tool	**Adobe XMP**
Source URL	http://www.adobe.com/products/xmp/
Formats covered	Image formats and PDF
Technology base	C++
Operating system	Windows 98/98(SE)/ME/NT4 Workstation/NT4 Server/2000 Pro/2000 Server/2003 Server/XP Home/XP Pro/XP Home/XP Professional/XP 64-Bit Edition/Terminal Services
Dependencies	The XMP toolkit needs an external XML parser. The source from Adobe is written to use Expat, although adapters for other parsers can easily be written. The most recent version of Expat used with XMP is 1.95.8.
Tool category	Metadata creator (manual) Metadata transformer
License	The XMP Software Development Kit is covered by the Adobe Systems Incorporated Open Source License. It grants the user license to use, reproduce, prepare derivative works from, display, perform, distribute and sublicense the software provided a stipulated copyright notice appears within the source code of the distributed software, and the license is distributed in any documentation of the software distributed by the user. Similar rights are granted over the documentation, although this may not be modified. License available at http://partners.adobe.com/public/developer/en/xmp/sdk/XMPLicense.pdf.
Description	Extensible Metadata Platform. This tool will assist the digital preservation workflow in that it allows the user to attach metadata to a digital object. The metadata syntax is a subset of RDF, expressed in XML. XMP is also extensible, allowing users to incorporate their own existing metadata schemas. It is an open-source product and freely available

	to the community. The tool is particularly pertinent for digital imaging but it is not limited to a particular file format. If archive creators make use of XMP, it will provide valuable preservation metadata for digital archivists, and a number of standards groups with a stake in digital preservation are already working on initiatives based on XMP, e.g. DCMI, RLG. The XMP Software Development Kit provides documentation, tools, and sample code.
Output method	Numerous methods.
Maturity	
Notes	

Tool	**Antiword**
Source URL	http://www.winfield.demon.nl/
Formats covered	Microsoft Word versions 2, 6, 7, 97, 2000, 2002 and 2003
Technology base	C
Operating system	Linux, RISC OS, FreeBSD, BeOS, OS/2, Mac OS X, Amiga, VMS, Netware, Plan9, EPOC, Zaurus PDA, MorphOS, Tru64/OSF and DOS.
Dependencies	
Tool category	Format transformer
License	GNU (GPL)
Description	Antiword is a free MS Word reader which converts documents from Word 2, 6, 7, 97, 2000, 2002 and 2003 to plain text, PostScript, PDF and XML/DocBook (a version of XML originally intended for technical documentation. The conversion to XML/DocBook is still experimental). Antiword tries to keep the layout of the document intact.
Output method	Unknown/unclear
Maturity	
Notes	

Tool	**Apache Xerces**
Source URL	http://xerces.apache.org/
Formats covered	XML
Technology base	Provides C++, Java or Perl implementations
Operating system	Linux, Cygwin, Windows, Mac OS X, BSD, Solaris, AIX, Tru64.
Dependencies	
Tool category	Metadata validator Metadata
License	All contributions to the Apache Xerces project adhere to the Apache Software Foundation License, v.2.0 (http://www.apache.org/licenses/LICENSE-2.0).
Description	A collaborative software development project dedicated to

	providing robust, full-featured, commercial-quality, and freely available XML parsers and closely related technologies on a wide variety of platforms supporting several languages. Xerces was a sub project of the Apache XML project, which contains many useful sub projects.
Output method	XML file
Maturity	
Notes	

Tool	**Aperture**
Source URL	http://aperture.sourceforge.net/
Formats covered	Different formats
Technology base	Java
Operating system	OS independent (written in an interpreted language)
Dependencies	Java 1.4.2 or above
Tool category	Metadata extractor (automatic) MIME-type identifier
License	The core project APIs and architecture are licensed under the AFL v. 3.0. The implementations of these APIs (e.g. extractors for specific file formats) are licensed under the OSL v. 3.0.
Description	Aperture is a Java framework for extracting and querying full-text content and metadata from various information systems (e.g. file systems, websites, mail boxes) and many of the common file formats (e.g. for documents and images) created by these systems. It also offers a facility for MIME type identification and is developing a facility for querying and indexing extracted information. Aperture tutorials for developers are available at http://aperture.sourceforge.net/tutorial/index.html.
Output method	XML file
Maturity	3 – Alpha
Notes	

Tool	**Archivists Toolkit**
Source URL	http://archiviststoolkit.org
Formats covered	N/A
Technology base	Java
Operating system	OS independent (written in an interpreted language)
Dependencies	MySQL 5.0 Community Server
Tool category	Metadata creator (manual) Metadata transformer Metadata wrapper
License	ECL

Description	The Archivists' Toolkit (AT) is an open source archival collection management system, designed to deal primarily with 'traditional' (hard copy) archives. It supports a number of core archival functions: accessioning; location tracking; donor tracking; archival description of items, collections and surrogates; and creating name and subject authority terms. It is modular, which allows users to select the relevant functional areas required for their purposes. Legacy data in multiple formats (including EAD 2002 and MARC XML) can be ingested into the system.
Output method	Exports to EAD 2002, MARC XML, METS, MODS and Dublin Core.
Maturity	Version 1.0 released with Phase 2 development funded from February 2007 for 24 months.
Notes	Metadata must be supplied to the Archivist's Toolkit.

Tool	**BRSoftware - EXIFextractor**
Source URL	www.br-software.com/extracter.html
Formats covered	JPEG
Technology base	Unknown/unclear
Operating system	Windows 95/98/NT4/2000/XP
Dependencies	
Tool category	Metadata extractor (automatic)
License	Freeware; no access to source code and no modifications permitted.
Description	Free program that extracts Exchangeable Image File Format (EXIF) metadata (i.e. the format used by most digital cameras to store metadata about camera settings etc) from JPEG image files. The extracted metadata is saved in a CSV (Comma Separated Values) file, which can be read by any program capable of reading CSV-files (e.g. Microsoft Excel, Microsoft Access and most other databases). The program is still at Beta test stage although a final release is planned, as is a Pro version which will also be able to extract International Press Telecommunications Council (IPTC) metadata.
Output method	CSV-file
Maturity	
Notes	Binary download; executable provided for Windows environment.

Tool	**Caliph and Emir**
Source URL	http://sourceforge.net/projects/caliph-emir/. Also http://www.semanticmetadata.net/
Formats covered	'Digital Photos'; JPEG only?
Technology base	Java / MPEG-7

Operating system	OS independent (written in an interpreted language)
Dependencies	Java 1.5 or greater
Tool category	Metadata extractor (automatic) Metadata transformer Metadata creator
License	GNU (GPL)
Description	Caliph & Emir are MPEG-7 based Java prototypes for both digital image annotation (carried out by Caliph – the Common And Lightweight Interactive PHoto annotation) and retrieval (using Emir – Experimental Metadata based Image Retrieval). Features of Caliph include: the extraction of existing EXIF and IPTC metadata, and its conversion to MPEG-7 metadata; the extraction of some content-based metadata (e.g. colour distribution, dominant colours); the creation of new metadata – both free text or structured; and addition of semantic annotations. Annotations are loaded and saved as an MPEG-7 XML file. Emir allows the search/retrieval of images by various means, e.g. retrieval of textual metadata by keyword searches, and content-based image retrieval by means of three MPEG-7 descriptors. There is a developer page for those who wish to modify or extend Caliph & Emir, at http://www.semanticmetadata.net/wiki/doku.php?id=caliphemir:developerdocs.
Output method	XML metadata compliant with MPEG-7; urn:mpeg:mpeg7:schema:2001.
Maturity	4 - Beta
Notes	Source available; cross-platform.

Tool	**CatDoc**
Source URL	http://www.45.free.net/~vitus/software/catdoc/
Formats covered	Microsoft Word, PowerPoint, Excel
Technology base	C
Operating system	UNIX, MS DOS
Dependencies	C Compiler (MS DOS version – binary provided)
Tool category	Format transformer
License	GNU (GPL)
Description	CatDoc is a program developed in Russia which reads one or more Microsoft word files and outputs their textual content in plain text (ASCII or TeX). CatDoc is accompanied by xls2csv, a program which converts Excel spreadsheets into CSV files; and catppt, which extracts textual information from Powerpoint files.
Output method	'Standard output'. Documentation suggests catdoc comes with two output formats: ASCII and TeX, although users can add their own if they wish.
Maturity	

Notes	Designed for Linux and Solaris, may work on Unix. Does not support Microsoft Windows. Catdoc is distributed with Cyrillic character sets as default, so would need to be reconfigured for non-Russian languages.

Tool	**Chiba**
Source URL	http://chiba.sourceforge.net/; http://sourceforge.net/projects/chiba
Formats covered	N/A
Technology base	Javabeans/XML
Operating system	OS independent (written in an interpreted language)
Dependencies	Java 1.5 or greater
Tool category	Interface tool to implement metadata creator
License	Artistic License [unclear whether this is Artistic License 2.0 or the original Artistic License; queries have been raised as to whether the latter is actually a free software license – unless software using it is dual licensed with the GNU GPL as well, as is the case with Perl].
Description	Chiba is an Open Source Java Implementation of the W3C XForms standard. Forms allow web applications to accept input from a user: form elements in a website allow the user to enter information (e.g. via text fields, drop-down menus, radio buttons, check boxes). XForms is the successor to HTML Forms, being more flexible and platform independent; it uses XML for data definition and HTML or XHTML for data display.
Output method	Unknown/unclear
Maturity	4 – Beta; 5 – Production/Stable
Notes	Chiba could be used to produce forms to generate XML metadata via a web browser. It is one of a number of tools mentioned on the DCC Development web site at http://twiki.dcc.rl.ac.uk/bin/view/Main/PreservationDescriptionInformationTool.

Tool	**DRC Bulk Ingest Tool**
Source URL	https://drc-dev.ohiolink.edu/wiki/BulkIngestToolLocalFiles; http://www.fedora.info/download/2.2/services/diringest/doc/index.html
Formats covered	N/A
Technology base	Java
Operating system	OS independent (written in an interpreted language)
Dependencies	Fedora digital repository (dirIngest)
Tool category	Metadata transformer
License	Unknown/unclear

Description	A tool developed by OhioLINK's Digital Resource Commons (DRC) project which is designed to produce a the METS file for a directory of related folder and files, which can then be submitted to the Fedora digital repository using its dirIngest service, but it requires the user to submit this metadata in tab-delimited text file. The resulting METS file can be presented to Fedora zipped up with the folder and files it describes; Fedora then creates foxml files for each file and folder which contain the manually added metadata as well as RDF relationship metadata describing a file/folder's relationships with parent/child objects (derived from the METS structMap).
Output method	XML (METS)
Maturity	Beta
Notes	Source available from website; has open source feel, but no clear licensing information. Development seems to be dormant.

Tool	**Drew Noakes Metadata Extraction Library/Exif-O-Matic**
Source URL	http://www.drewnoakes.com/code/ http://www.instituteofthefuture.org/exifomatic/
Formats covered	JPEG (Exif, Iptc, JPEG Segment)
Technology base	Java
Operating system	OS independent (written in an interpreted language)
Dependencies	
Tool category	Metadata extractor (automatic)
License	No formal license. Statement on Drew Noakes' wesite indicates that the code is protected by copyright, but users are free to use it as they see fit. He encourages users who make changes to the code to contact him so details can be included on the website. Users are free to sell work based on this library, although are encouraged to make a donation.
Description	A code library which has already been implemented in a number of applications (e.g. Exif-o-Matic, which provides a user-friendly front end application). It is designed to extract image metadata from JPEG files; supported metadata is EXIF, IPTC and JPEG Segment (the JPEG format allows for a variety of customised metadata segments).
Output method	Text
Maturity	
Notes	

Tool	**DROID - National Archives**
Source URL	http://droid.sourceforge.net/wiki/index.php/Introduction; http://sourceforge.net/projects/droid/
Formats covered	As per PRONOM technical registry

Technology base	Java
Operating system	OS independent (written in an interpreted language)
Dependencies	
Tool category	Metadata extractor (automatic) Format identifier
License	BSD
Description	DROID (Digital Record Object Identification) is a software tool developed by The National Archives to perform automated batch identification of file formats. Developed by its Digital Preservation Department as part of its broader digital preservation activities, DROID is designed to meet the fundamental requirement of any digital repository to be able to identify the precise format of all stored digital objects, and to link that identification to a central registry of technical information about that format and its dependencies.
Output method	XML (DROID's own schema) or CSV
Maturity	5 – Production/Stable
Notes	DROID can act s a local or an on-line service. Likely to support a growing number of formats, widespread use in the preservation community.

Tool	**Elated**
Source URL	http://elated.sourceforge.net/; http://sourceforge.net/projects/elated/
Formats covered	N/A
Technology base	Java / J2EE
Operating system	OS independent (written in an interpreted language)
Dependencies	Java web application (APT version 2.3) MVC architecture using Apache Struts MySQL/JDBC Apache Lucene Fedora
Tool category	Metadata creator
License	GNU (GPL)
Description	Elated is a lightweight, general-purpose application for managing digital files. It was developed as a user-friendly web interface for the Fedora Repository System: whilst not competing with Fedora in the area of key services, it offers interfaces to those services wherever possible. All the digital objects being preserved, and their associated metadata (in the form of datastreams) are stored in Fedora as normal, along with an extra Elated datastream designed to store extended metadata beyond the basic Dublin Core stream. Other types of data (e.g. Elated authorisation and authentication information, text-search indexes etc) is stored separately in the Elated local database.
Output method	XML

Maturity	4 - Beta
Notes	

Tool	**ExifTool**
Source URL	http://www.sno.phy.queensu.ca/~phil/exiftool/
Formats covered	EXIF, GPS, IPTC, XMP, JFIF, GeoTIFF, ICC Profile, Photoshop IRB, FlashPix, AFCP and ID3 meta information as well as the maker notes of many digital cameras including Canon, Casio, FujiFilm, JVC/Victor, Kodak, Leaf, Minolta/Konica-Minolta, Nikon, Olympus/Epson, Panasonic/Leica, Pentax/Asahi, Ricoh, Sanyo, Sigma/Foveon and Sony. File formats covered include: Windows Bitmap; Microsoft Word Document (FPX-like); HTML; XHTML; JPEG; JPEG 2000; MPEG 4 Audio; MP3; MP4; PDF; PNG; TIFF; WAV; and numerous others.
Technology base	Perl
Operating system	Windows, MAC OS X and Unix systems
Dependencies	Perl (see installation instructions)
Tool category	Metadata extractor (automatic) Metadata transformer
License	As perl - http://dev.perl.org/licenses/ Either GNU (GPL) v1+, or the Artistic License
Description	ExifTool is a Perl module with an included command-line application for reading, writing and editing metadata in image, audio, PDF and video files
Output method	Writes EXIF, GPS, IPTC, XMP, JFIF, MakerNotes, ICC Profile, Photoshop IRB, AFCP
Maturity	
Notes	Full distribution of source is available for Unix/Linux platforms. Also available as a stand-alone Windows executable and a Macintosh OS X package.

Tool	**Expat**
Source URL	http://expat.sourceforge.net/; http://sourceforge.net/projects/expat/
Formats covered	XML
Technology base	C
Operating system	OS portable – can work with many OS platforms
Dependencies	
Tool category	Metadata validator
License	MIT License
Description	XML parser library. A stream oriented parser that requires setting handlers to deal with the structure that the parser discovers in the document.

Output method	
Maturity	6 - Mature
Notes	

Tool	**Ffident**
Source URL	http://schmidt.devlib.org/ffident/index.html
Formats covered	MS Word, images, audio, video, executables and archive files; unclear whether version information is generated. Additional formats can be added by editing a text file.
Technology base	Java
Operating system	OS independent (written in an interpreted language)
Dependencies	Java 1.4 or higher
Tool category	Metadata extractor (automatic) [currently only acts as format identifier; extraction of format-group-specific metadata from files is given as a plan for the future] Format validation
License	GNU (LGPL)
Description	This is the first version of a Java library to extract information from files and identify their formats. It uses the same approach as the Unix command line utility 'file (1)', i.e. collecting information on the most interesting and common file formats, and checking each file to be examined against a list of known signatures (the 'magic (5)' file).
Output method	Described as 'standard output'; probably text.
Maturity	
Notes	

Tool	**GTRI XML Validation tool**
Source URL	http://justicexml.gtri.gatech.edu/gtri_xml_tools.html
Formats covered	XML
Technology base	Java
Operating system	OS independent (written in an interpreted language)
Dependencies	
Tool category	XML validator
License	The Xerces parser is provided under the Apache Software License (which allows redistribution and use, with or without modification, subject to small number of conditions, e.g. inclusion of disclaimer); see http://www.opensource.org/licenses/apachepl.php. The GTRI code is provided under the Georgia Tech Research Institute proprietary License (allows use of software and access to source code for purpose of making derivative works and modifications, for incorporation with third party source code and/or to reverse engineer, subject to some conditions, e.g. disclaimer).

Description	An application developed by the Georgia Tech Research Institute (GTRI) for validating XML documents against their referenced schemas; it makes use of the Apache Xerces Java XML Parser library and can be used for any instances of any XML schemas.
Output method	
Maturity	
Notes	License is included as part of download package.

Tool	**id3lib**
Source URL	http://www.id3lib.org/
Formats covered	ID3v1/IDV3v2 (descriptive tagging for MP3 files)
Technology base	C / C++ / VB
Operating system	All 32-bit MS Windows (95/98/NT/2000/XP) All POSIX (Linux, BSD, Unix)
Dependencies	
Tool category	Metadata extractor (automatic) Metadata transformer
License	GNU (LGPL)
Description	id3lib is an open-source, cross-platform software development library for reading, writing, and manipulating ID3v1 and ID3v2 tags (i.e. the containers used for storing metadata in MP3 audio files, such as song title, artist, album etc). It is an ongoing project which aims to achieve full compliance with the ID3v2 standard, and provide a powerful API with a highly stable and efficient implementation.
Output method	Unknown/unclear
Maturity	5 – Production/Stable
Notes	Last changed Id3lib 3.8.3 release 2003-03-02.

Tool	**ImageInfo**
Source URL	http://schmidt.devlib.org/image-info/index.html; http://freshmeat.net/projects/imageinfo/
Formats covered	JPEG, PNG, GIF, BMP, PCX, IFF, RAS, PBM, PGM, PPM and PSD.
Technology base	Java
Operating system	OS independent (written in an interpreted language)
Dependencies	
Tool category	Metadata extractor (automatic)
License	Listed as a Public Domain license at Freshmeat
Description	A free Java class which can recognise the formats listed above, determine image width, height and color depth (bits per pixel).
Output method	text

Maturity	
Notes	Probably in need of refactoring according to the author.

Tool	**Jacksum**
Source URL	http://sourceforge.net/projects/jacksum/; http://www.jonelo.de/java/jacksum/
Formats covered	N/A
Technology base	Java
Operating system	OS independent (written in an interpreted language)
Dependencies	
Tool category	Message digest calculator
License	GNU (GPL)
Description	Jacksum is a platform independent tool for calculating and verifying checksums, hash algorithms and Cyclical Redundancy Checks (CRCs). It supports 58 algorithms (v. 1.7.0).
Output method	Text or file
Maturity	5 – Production/Stable
Notes	Provides an API.

Tool	**Java Beans MIME Type extractor**
Source URL	http://java.sun.com/products/javabeans/glasgow/javadocs/javax/activation/FileDataSource.html
Formats covered	N/A
Technology base	Java Beans
Operating system	OS independent (written in an interpreted language)
Dependencies	
Tool category	Mime-type identifier
License	Presumably covered by the license cited on the Sun Developer Network Site; text at http://developers.sun.com/license/berkeley_license.html. Allows redistribution and use in source and binary forms, with or without modifications subject to retention of copyright notice and specified disclaimer.
Description	JavaBeans Activation Framework class to identify the MIME type of a given file. JavaBeans are reusable software programs that can be developed and assembled to create applications.
Output method	Unknown/unclear
Maturity	
Notes	This is a software component not a fully formed application.

Tool	**Java Metadata Collection**

Source URL	http://www.buckazoid.com/jmdc/; https://sourceforge.net/projects/jmdc/
Formats covered	FLAC, Ogg Vorbis, PDF
Technology base	Java
Operating system	OS independent (written in an interpreted language)
Dependencies	
Tool category	Metadata extractor (automatic) Metadata transformer
License	BSD
Description	The Java Metadata Collection (JMDC) is a set of Java API's for extracting and editing metadata stored in various file types. It is currently able to extract metadata from FLAC and Ogg Vorbis files (both open-source alternatives to MP3 audio format) and PDF. Future versions of the tool will aim to: read and write ID3 tags (i.e. metadata) in MP3 files; add write capabilities to Ogg and FLAC files; and read EXIF and image metadata in both JPEG and TIFF files.
Output method	Unknown/unclear
Maturity	4 - Beta
Notes	This is a software component not a fully formed application.

Tool	**Java Mime Magic Library**
Source URL	http://sourceforge.net/projects/jmimemagic/; http://jmimemagic.sourceforge.net/
Formats covered	N/A
Technology base	Java
Operating system	OS independent (written in an interpreted language)
Dependencies	
Tool category	MIME-type identifier
License	GNU (LGPL)
Description	A Java library for determining the MIME type of files or streams.
Output method	Unknown/unclear
Maturity	3 – Alpha
Notes	This is a software component not a fully formed application.

Tool	**Jhead**
Source URL	http://www.sentex.net/~mwandel/jhead/
Formats covered	Exif / JPEG
Technology base	C
Operating system	Windows/Linux
Dependencies	Unknown/unclear
Tool category	Metadata extractor (automatic)

License	Public domain; unrestricted by license.
Description	A program which parses and extracts EXIF metadata from JPEG files, enables the user to modify or edit some elements of this metadata, and to manipulate the thumbnails included as part of the EXIF header for a JPEG image.
Output method	Text
Maturity	
Notes	Jhead is a command-line cross platform application.

Tool	**Jhove**
Source URL	http://hul.harvard.edu/jhove/
Formats covered	XML, WAV, TIFF, UTF-8, PDF, JPEG, JPEG2000, HTML, GIF, Bytestream, ASCCI, AILL. Also has default profile for objects not conforming to these types.
Technology base	Java
Operating system	OS independent (written in an interpreted language)
Dependencies	
Tool category	Metadata extractor (automatic) Format identifier Format validator
License	GNU (LGPL)
Description	JHOVE, the JSTOR/Harvard Object Validation Environment, is an extensible software framework for performing format identification, validation, and characterisation of digital objects. Format validation conformance is determined at two levels: well-formedness (consistent with the basic requirements of the format); and validity (both well-formed and meeting certain additional semantic-level requirements). Characterisation involves determining the format-specific significant properties of a digital object and reporting these (to form a digital object's Representation Information).
Output method	XML
Maturity	
Notes	

Tool	**Kaa Media Repository – Kaa Metadata module**
Source URL	http://freevo.sourceforge.net/cgi-bin/freevo-2.0/Kaa; http://sourceforge.net/project/showfiles.php?group_id=46652&package_id=213173
Formats covered	Audio (AC3, DTS, FLAC, MP3, OGG, PCM, M4A, WMA); video (AVI, MKV, MPG, OGM, ASF, WMV, FLV, MOV, DVD ISO, VCD ISO); image (JPEG, BMP, GIF, PNG, TIFF).
Technology base	Python
Operating system	POSIX (Linux/BSD/Unix)
Dependencies	

Tool category	Metadata extractor (automatic)
License	GNU (GPL)
Description	The Kaa Media Repository is a set of modules for the Python programming language designed as an umbrella for several previously disparate media-related Python modules, one element of which is metadata extraction. Kaa modules are based on parts from Freevo (open-source application for Linux and BSD designed to run a personal video recorder) and modules created for MeBox (a project to implement home theatre PC software under Linux). Kaa provides a base module that implements the common features needed for application development, and the other modules provide specific media-related functionality. Most relevant is the kaa-metadata module (previously called MMPython), which can extract metadata (e.g. ID3 tags) from a wide range of file formats, and also returns some additional attributes (e.g. length, resolution). Supported formats include audio, video, image and media (i.e. cd, dvd etc).
Output method	Unknown/unclear
Maturity	
Notes	Software components rather than fully formed applications.

Tool	**Kea**
Source URL	http://www.nzdl.org/Kea/
Formats covered	TXT files
Technology base	Java
Operating system	OS independent (written in an interpreted language)
Dependencies	
Tool category	Metadata extractor (automatic) Metadata transformer
License	GNU (GPL)
Description	Kea is an algorithm for extracting key phrases from the text of a document to provide semantic metadata that might be useful for a variety of purposes, most obviously for creating descriptive metadata about a digital object. It can be used either for free indexing (i.e. extracting keyphrases freely chosen by the author/user) or for indexing with a controlled vocabulary - ensuring documents are indexed consistently regardless of their wording; for controlled indexing, the tool supports any vocabulary in the Simple Knowledge Organisation Systems (SKOS) format. SKOS is developing specification and standards to support the use of knowledge organisation systems, like subject heading systems and thesauri, within the framework of the Semantic Web.
Output method	Unknown/unclear
Maturity	
Notes	This is a software component not a fully formed application.

Tool	**Libexif**
Source URL	https://sourceforge.net/projects/libexif; http://libexif.sourceforge.net/
Formats covered	JPEG/Exif
Technology base	C
Operating system	Windows/Linux
Dependencies	
Tool category	Metadata extractor (automatic) Metadata transformer
License	GNU (LGPL)
Description	EXIF Tag Parsing Library which reads and writes EXIF metadata from and to image files.
Output method	Writes Exif metadata to image files.
Maturity	4 – Beta
Notes	Runs under POSIX systems (e.g. GNU/Linux, xBSD, MacOS X, etc.) and Win32. Win64 untested.

Tool	**libextractor**
Source URL	http://gnunet.org/libextractor/
Formats covered	MP3, Ogg, Real Media, MPEG, RIFF (avi), GIF, JPEG, PNG, TIFF, HTML, PDF, PostScript, Zip, OpenOffice.org, StarOffice, Microsoft Office, tar, DVI, MAN, DEB, elf, RPM, asf, NSF, SID, WAV, EXIV2, REAL, RIFF (AVI), MPEG, QT. Various additional MIME-types are also detected.
Technology base	C
Operating system	Mac OS X, MS Windows, POSIX
Dependencies	
Tool category	Metadata extractor (automatic) MIME-type identifier
License	GNU (GPL)
Description	libextractor is a library with an extract tool which is used to extract metadata from different file formats. It detects the MIME-type, and depending on the file format, additional metadata including the name of the software used to create the file, the author, descriptions, album titles, image dimensions, date of creation, etc. It obtains this information by using specific parser code for many popular formats. New formats can be added using plugins. The tool can sometimes obtain useful information even if the format of a file is unknown and unsupported.
Output method	Text
Maturity	
Notes	Binary packages are available online for many UNIX-like operating systems.

Tool	**libwmf**

Source URL	http://wvware.sourceforge.net/libwmf.html; http://sourceforge.net/projects/wvware/
Formats covered	Windows Metafile Format (WMF)
Technology base	C / Unix Shell
Operating system	Unix?
Dependencies	Unknown/unclear
Tool category	Format transformer
License	GNU Lesser (Library) Public License (LGPL)
Description	libwmf offers the capacity to: read vector images in Microsoft's native Windows Metafile Format (WMF); display them; and convert them to more standard or open file formats, e.g. SVG.
Output method	
Maturity	4 – Beta
Notes	

Tool	**Metadata Assistant**
Source URL	http://www.payneconsulting.com/products/metadataent/
Formats covered	Microsoft Office, GroupWise 6.01 and higher, and Lotus Notes 5 and 6
Technology base	Proprietary
Operating system	MS Windows
Dependencies	Unknown/unclear
Tool category	Metadata extractor (automatic)
License	Commercial
Description	The Metadata Assistant analyses Word/Excel/PowerPoint (97 and higher) files to determine the type and amount of metadata associated with each document. It removes (rather than just extracts) this metadata. It can operate as a standalone utility and is also capable of batch processing multiple files located on a local or network folder. It also removes metadata from files attached to outbound e-mail messages, and can convert 'cleaned' files to pdf format; it supports email integration with Outlook 2000-2003, GroupWise 6.01 and higher, and Lotus Notes 5 and 6.
Output method	RTF and XML format
Maturity	
Notes	This is a Windows application.

Tool	**Metaphile**
Source URL	http://www.miniturismo.co.uk/metaphile/
Formats covered	JFIF, JFXX, IPTC IIM (V3 and V4), EXIF (2.1 and 2.2) and XMP (Dublin Core, Photoshop, Iptc4XMPCore, Rights

	Management).
Technology base	Java
Operating system	OS independent (written in an interpreted language)
Dependencies	
Tool category	Metadata extractor (automatic)
License	GNU (GPL)
Description	Metaphile is a free, open source Java library for reading image metadata. It currently extracts segments of metadata found in JPEG files. Once the JPEG format is fully supported, the project aims to move on to other file types.
Output method	Unknown/unclear
Maturity	
Notes	This is a Java library for reading image metadata.

Tool	**NLNZ Preservation metadata Extraction Tool**
Source URL	http://www.natlib.govt.nz/about-us/current-initiatives/metadata-extraction-tool; http://meta-extractor.sourceforge.net/; http://sourceforge.net/projects/meta-extractor/
Formats covered	Bitmap, Microsoft Excel, GIF, HTML, JPG, MP3, Open Office, PDF, Microsoft PowerPoint, TIFF, Wave Audio, Microsoft Word, WordPerfect, Microsoft Works, XML
Technology base	Java
Operating system	OS independent (written in an interpreted language)
Dependencies	
Tool category	Metadata extractor (automatic)
License	Apache License v. 2.0
Description	A tool developed by Sytec for the National Library of New Zealand, which automatically extracts preservation-related metadata from the headers of a range of file formats, and outputs that metadata in a standard format (XML) for uploading into a preservation metadata repository. It consists of a base generic extract process, with adapters written for different file formats. More adapters are planned, and these can easily be plugged into the application in the future. The tool only extracts metadata from file headers, meaning that the files themselves are not opened; this allows the extraction process to take place in a secure, read-only environment. The tool has recently been redeveloped an released as version 3. The tool supports simple and complex objects.
Output method	XML (native and NLNZ schemas)
Maturity	5 – Production/Stable
Notes	

Tool	**OMAR Representation Information Repository**
Source URL	http://registry.dcc.ac.uk/omar/
Formats covered	None as yet.
Technology base	freeEBXML, but used as an external registry
Operating system	
Dependencies	Unknown/unclear
Tool category	Metadata registry
License	Unknown/unclear
Description	A registry which is being developed by the Digital Curation Centre (DCC) to provide an infrastructure for the preservation of Representation Information (RI). RI is defined by the OAIS Reference Model (ISO 14721:2003), and is essentially the information which is needed to ensure that the bitstream of an archival digital object can be transformed into something meaningful and understandable over time; it includes structural, semantic and other information and encompasses things like file format, operating system, hardware dependencies, character encoding etc. RI can either be held locally within a digital repository, or held externally in a reliable repository and referred to from multiple archives. The DCC Registry/Repository aims to fulfil the latter role: it records RI and assists users to populate relevant technical preservation metadata fields for the digital objects they preserve. Repositories will be able to refer to RI held in the registry by means of a RI label (in the form of an XML Schema) which can be attached to a digital object. The Registry/Repository is intended as a collaborative resource, and members of the digital preservation community are encouraged to submit information to it.
Output method	Unknown/unclear
Maturity	
Notes	

Tool	**OpenExif**
Source URL	http://sourceforge.net/projects/openexif/; http://openexif.sourceforge.net/
Formats covered	JPEG
Technology base	C++
Operating system	All 32-bit MS Windows (95/98/NT/2000/XP) All POSIX (Linus, BSD, Unix, Mac OS X)
Dependencies	
Tool category	Metadata creator
License	Common Public License
Description	An object-oriented library for accessing Exif formated JPEG image files. The toolkit allows creating, reading, and modifying the metadata in the Exif file. It also provides

	means of getting and setting the main image and the thumbnail image.
Output method	
Maturity	5 – Production/Stable
Notes	

Tool	**OpenTIFF**
Source URL	http://sourceforge.net/projects/opentiff/
Formats covered	TIFF
Technology base	C++
Operating system	All 32-bit MS Windows (95/98/NT/2000/XP) All POSIX (Linus, BSD, Unix, Mac OS X)
Dependencies	
Tool category	Metadata creator
License	Other/Proprietary License
Description	A TIFF toolkit which provides an object-oriented interface to TIFF image files. It allows an arbitrary set of tags to be defined and used in a TIFF file.
Output method	Unknown/unclear
Maturity	5 – Production/Stable
Notes	

Tool	**Pedro**
Source URL	http://pedrodownload.man.ac.uk/
Formats covered	N/A
Technology base	Java
Operating system	OS independent (written in an interpreted language)
Dependencies	
Tool category	Interface tool to implement metadata creator Metadata validator
License	AFL
Description	Pedro is a free open-source application developed by the scientific community, but it may be suitable for use in other environments. It can be used to create data entry forms to capture the information required by a specified XML Schema. By means of controlled vocabularies and validation routines, the tool ensures that the resulting data file conforms to the relevant Schema. It therefore provides an interface appropriate for users who are not necessarily familiar with XML Schemas.
Output method	XML
Maturity	
Notes	Pedro is one of a number of tools mentioned on the DCC Development web site at http://twiki.dcc.rl.ac.uk/bin/view/Main/PreservationDescriptio

	nInformationTool.

Tool	**Picture Metadata Toolkit**
Source URL	http://picturemetadata.sourceforge.net
Formats covered	EXIF (JPEG), TIFF
Technology base	C++
Operating system	All 32-bit MS Windows (95/98/NT/2000/XP) All POSIX (Linus, BSD, Unix, Mac OS X)
Dependencies	OpenTiff, OpenExif, the IJG JPEG toolkit and Xerces-C++ v. 2.2.
Tool category	Metadata extractor (automatic) Metadata transformer Metadata creator
License	Common Public License
Description	The Picture Metadata Toolkit (PMT) provides functionality for the extraction, creation and editing of metadata associated with or stored in digital image files. The metadata extracted from various formats are all treated as PmtMetadata objects, which are based on an XML Schema. New file formats can be integrated into PMT, and a new XML file format may be created to hold any kind of data for PMT to work with.
Output method	Unknown/unclear
Maturity	5 – Production/Stable
Notes	This tool has good documentation at http://picturemetadata.sourceforge.net/doc/PmtUserGuide.pdf

Tool	**Query Electronic Storage (QUEST)**
Source URL	https://sourceforge.net/projects/quest-archiv/
Formats covered	Archival Information Packages and Xena
Technology base	Java
Operating system	OS independent (written in an interpreted language)
Dependencies	XENA, DPR
Tool category	Metadata transformer
License	Unknown/unclear
Description	QUEST is a Java based application that creates links between digital objects held in a repository and their associated metadata, and enables the user to retrieve Archival Information Packages from a digital repository using various metadata elements. It is designed to work as an integral part of a software suite developed by the National Archives of Australia for carrying out long-term digital preservation; the other components of the suite are the Digital Preservation Recorder (DPR), which creates an

	audit trail recording the complete life history of each digital object being preserved, and XML Electronic Normalisation of Archives (XENA), which converts digital files into XML formats.
Output method	Unknown/unclear
Maturity	
Notes	This is designed to only work with Xena and DPR. The application is still under development and the project hasn't yet created any file release packages. It is possible that the Quest branch of code has been deprecated and included in DPR.

Tool	**SHA**
Source URL	http://www.saddi.com/software/sha/
Formats covered	N/A
Technology base	C
Operating system	All POSIX (Linux/BSD/Unix)
Dependencies	
Tool category	Message digest calculator
License	BSD
Description	SHA is a simple program that creates and assigns a fixity (or Hash) value to a digital object which can be used in a digital repository for the purpose of integrity checking. It can use SHA-1, SHA-256, SHA-384, or SHA-512 cryptographic hash functions, which generate (respectively) hash values of 160, 256, 384, or 512 bits. SHA can be used in scripts, e.g. to implement file integrity checking.
Output method	?
Maturity	
Notes	The current version of SHA is 1.0.4. It is known to build on FreeBSD, Darwin, OpenBSD, Debian Linux, and Solaris (using gcc) on a variety of architectures.

Tool	**SHAME**
Source URL	http://kmr.nada.kth.se/shame/wiki
Formats covered	RDF
Technology base	Java / Swing
Operating system	OS independent (written in an interpreted language)
Dependencies	The Edutella library – for the graph patterns and to search the Edutella P2P network. The Jena2 library – an API to work with RDF Velocity supplied with the download library.
Tool category	Metadata creator (manual) Tool to implement metadata creator
License	GNU (GPL); GNU (LGPL); MPL 1.1

Description	The Standardized Hyper Adaptible [*sic!*] Metadata Editor (SHAME) is a metadata editing and presentation tool. It aims to provide a general purpose form-based graphical user interface, which can be configured to work with a specific class of RDF-graphs, e.g. RDF that describes resources according to qualified Dublin Core. It works by means of Annotation Profiles based on a specific metadata standard or schema; an Annotation Profile determines how the RDF should be read/modified, the input which is allowed (e.g. specified vocabularies), and presentational issues like grouping, labels, ordering etc. The Annotation Profiles are then used to generate user interfaces for either editing, presentation or searching purposes. The user interface may be realised in a web setting or a stand alone application.
Output method	RDF via the Jena Framework
Maturity	3 - Apha
Notes	

Tool	**Signify**
Source URL	http://signify.sourceforge.net/; http://sourceforge.net/projects/signify/
Formats covered	N/A
Technology base	Perl-5
Operating system	OS independent (written in an interpreted language)
Dependencies	
Tool category	Digital signature creator
License	Public Domain
Description	Signify is an automatic random signature generator; it allows a digital signature to be generated from a set of rules. It allows the user to create a signature in multiple sections; each section may be one of an unlimited number of possibilities, and can be allocated its own weighting, thereby enabling some sections to appear more frequently than others.
Output method	
Maturity	6 - Mature
Notes	Signify is an OS Independent (Written in an interpreted language).

Tool	**SIP creator**
Source URL	http://www.fedora.info/download/2.2/services/sipcreator/doc/index.html; http://www.fedora.info/download/2.2/services/diringest/doc/index.html
Formats covered	N/A
Technology base	Java

Operating system	OS independent (written in an interpreted language)
Dependencies	Fedora digital repository (v. 2.2) Servlet container supporting Servlet API v. 2.3
Tool category	Metadata creator (manual) Metadata creator (automatic) Interface tool to implement metadata creator
License	Educational Community License
Description	SIP creator is a tool designed to produce a METS file for a directory of related folder and files, which can then be submitted to the Fedora digital repository using its dirIngest service. The tool automatically generates a METS structural map describing the hierarchical relationships between the files and additional XML metadata of any kind can be manually added (this could be generated automatically from other tools) about any of the component files and folders. The resulting METS file can be presented to Fedora zipped up with the folder and files it describes; Fedora then creates foxml files for each file and folder which contain the manually added metadata as well as RDF relationship metadata describing a file/folder's relationships with parent/child objects (derived from the METS structMap).
Output method	XML (METS)
Maturity	
Notes	

Tool	**Soft Experience - Metadata Miner Catalogue PRO**
Source URL	http://peccatte.karefil.com/software/Catalogue/MetadataMiner.htm
Formats covered	Multiple 'Office' type formats (Microsoft Office, OpenOffice.org, StarOffice) as well as PDF document information, Adobe XMP metadata, HTML documents and image formats such as JPEG, TIFF and PSD.
Technology base	Unknown/unclear
Operating system	MS Windows 95/98/ME/NT4/2000/XP
Dependencies	Unknown/unclear
Tool category	Metadata extractor (automatic) Metadata transformer
License	Commercial. Details at http://peccatte.karefil.com/software/Catalogue/License_metadataminer.html. PRO version can only be used by one registered user on one computer. Shareware version of the software may be freely distributed. However, license forbids any modification of the program; see http://peccatte.karefil.com/software/Catalogue/License_metadataminer.html.
Description	Software which can extract metadata from directories of files, and allows the management of metadata elements for specific file formats. It exports this metadata and generates reports in various formats for different uses, e.g. produces

	listings of the metadata for one or more file directories, transforms the metadata into formats like XML for editing and data exchange.
Output method	HTML, XML, CSV, RTF, TXT, MS Word.
Maturity	
Notes	

Tool	**Spybot-Search & Destroy**
Source URL	http://www.safer-networking.org/
Formats covered	N/A
Technology base	Proprietary
Operating system	MS Windows (Linux/Unix via WINE emulator v. 1.5)
Dependencies	Unknown/unclear
Tool category	Spyware/anti-virus
License	Public license; license agreement detailed at http://www.safer-networking.org/en/license/index.html. The software can be freely used in its entirety (though use of parts only is not permitted); reverse engineering is forbidden,
Description	Spybot-Search & Destroy is a piece of free software which detects and removes spyware. It offers a range of features, e.g. removing adware and spyware (which track a user's web surfing behaviour to create a marketing profile which is sold on), dialers (which connect to the internet without the user's knowledge for fraudulent purposes), keyloggers (which capture the user's keystrokes), trojans (that install malicious programs on the user's computer), and usage tracks (user's history); and provides detailed information about problems encountered.
Output method	Output to screen
Maturity	
Notes	To be used alongside, and as well as, anti-virus software.

Tool	**Typed Object Model (Tom)**
Source URL	http://tom.library.upenn.edu/
Formats covered	Common text/spreadsheet formats
Technology base	Perl / Java
Operating system	OS independent (written in a interpreted language)
Dependencies	
Tool category	Format transformer Format identifier
License	'An open source license' which allows free downloading, use and adaptation of the software.
Description	TOM consists of: a data model for identifying and describing a wide variety of data types and formats; and a system of

	networked software that supports the description and use of these data types/formats (e.g. by providing information about specific data types; enabling the creation of TOM objects for data; and extracting information from unfamiliar formats). It also provides an online conversion tool to carry out 'respectful conversion' of submitted files in around 200 formats; this service is based on the idea that all digital objects can be divided into types which are defined by specific values, attributes, methods and semantics for each class of object. Respectful conversion means the significant properties of each object type are retained during conversion. Downloadable conversion software is being developed, so repositories can carry out conversion (e.g. migration or normalisation) locally, without sending sensitive documents outside their institutional firewall.
Output method	Unknown/unclear
Maturity	
Notes	The core TOM library is in Perl and modules support TOM clients, servers, and brokers. The library can be used to create TOM objects for data, request conversions and other operations, get information about types, and more. Users can also run a TOM broker of their own.

Tool	**TrID File Identifier**
Source URL	http://mark0.net/index-e.html
Formats covered	Extensive format recognition: at least 2488 types.
Technology base	Unknown/unclear
Operating system	MS Windows 32-bit Linux binary available
Dependencies	Unknown/unclear
Tool category	Format identifier
License	Unknown/unclear
Description	TrID is a program which identifies file types from their binary signatures. It makes use of an extensible database of definitions which describe recurring patterns for supported file types. Submitted files are read and compared with the definitions included in the database, and results are presented in order of highest probability; information on who authored the definition and their contact details are provided, as is any related URL. It is possible to scan groups of files or entire folders. TrID can also add guessed extensions to filenames which can be useful when working with files recovered by data rescue software.
Output method	Output to screen?
Maturity	
Notes	

Tool	**UIUC OAI Metadata Harvesting Project**
Source URL	http://uilib-oai.sourceforge.net/
Formats covered	JPEG
Technology base	Dependent on component ASP / XML / Java / Visual Basic
Operating system	Dependent on component MS Windows 32-bit (95/98/NT/2000/XP) OS Independent (written in an interpreted language)
Dependencies	Various dependent on which components are used. Potential dependencies include: Apache webserver, Tomcar servlets, MySQL database, NETPBM graphics library
Tool category	Metadata transformer
License	University of Illinois/NCSA Open Source License, which permits users to use, copy, modify, merge, publish, distribute, sublicense, or sell copies of the software, provided the copyright notice, and list of conditions and disclaimers are reproduced.
Description	The Library of the University of Illinois at Urbana-Champaign (UIUC) has developed a suite of Open Archives Initiative (OAI)-based metadata harvesting services, search services, and tools for discovery and retrieval. These are implemented in Visual Basic and Java, and include various stand-alone packages, e.g. a number of data provider services for storing metadata records and transforming XML metadata into Dublin Core records for OAI-PMH metadata harvesting; and a number of OAI metadata harvesters.
Output method	Unknown/unclear
Maturity	4 – beta; 5 – Production/stable.
Notes	

Tool	**upCast**
Source URL	http://www.infinity-loop.de/products/upcast/
Formats covered	MS Word, RTF, RTF-embedded WMF and RTF-embedded images
Technology base	Java/Commercial
Operating system	OS independent (written in an interpreted language)
Dependencies	J2SE v. 1.3+
Tool category	Interface tool to implement metadata creator
License	Commercial: information and pricing at http://www.infinity-loop.de/buy/upcast/index.html
Description	upCast enables the user to create XML documents using any application which can read RTF as an authoring application. upCast can perform batch conversions on these files to transform them into XML format; there is a built-in XML validator to ensure that documents conform to XML, and an XSLT processor for output. It is possible to export in raw XML (rather than XML which conforms to the upCast DTD) which supports the use of any DTD or Schema.
Output method	XML, XHTML, raw XML
Maturity	
Notes	

Tool	**wvWare**
Source URL	http://wvware.sourceforge.net/; http://sourceforge.net/projects/wvware/
Formats covered	Microsoft Office (primarily MS Word)
Technology base	C / Unix Shell
Operating system	Compiles and works on most Oss. Most development is undertaken in Linux, but also works on BSD, Solaris, OS/2, AIX, OSF1 and (with varying levels of success) AmigaOS VMS.
Dependencies	
Tool category	Metadata extractor (automatic) Format transformer
License	GNU (GPL)
Description	wvWare is both a metadata extractor (its wvSummary facility can print out metadata from Microsoft Office documents), and a format transformer; it allows other programs access to Word documents for the purpose of converting them to other formats.
Output method	Metadata from Office documents is ‘printed out’. Word documents can be converted to HTML4.0, LaTeX, DVI, PostScript, PDF, plain text, WML, RTF.
Maturity	5 – Production/Stable; 6 – mature
Notes	wvWare’s website recommends using AbiWord for some conversions; this is a more actively maintained product

	which supports many more output formats than wvWare, and output has a higher degree of fidelity to the original Microsoft Word document.

Tool	**Xena / Xenalite**
Source URL	http://xena.sourceforge.net/index.html; http://sourceforge.net/projects/xena/
Formats covered	MS-Word, Excel, Powerpoint and Project OpenOffice.org Writer, Calc, and Impress RTF PST email format TRIM email format MBOX email format MSG email format Comma Separated Files (CSV) JPG, GIF, TIFF, PNG, BMP, PCX HTML Plaintext (various encodings) PDF documents XML
Technology base	Java
Operating system	OS independent (written in an interpreted language)
Dependencies	
Tool category	Format transformer Metadata extractor
License	GNU (GPL)
Description	XML Electronic Normalising of Archives (Xena) is a tool developed by the National Archives of Australia (NAA) for converting a range of file formats to XML representations, for the purpose of long-term preservation. The NAA has developed a number of XML schemas to represent supported file formats, and the XML output of Xena conforms to this NAA standard XML.
Output method	XML
Maturity	4 - Beta
Notes	Need to discover how to access the raw XML; ability to process .pst files very useful.

Tool	**XML Batch Validator**
Source URL	http://sunsite3.berkeley.edu/ead/tools/schema_validate
Formats covered	XML/METS
Technology base	Perl/ Xerces
Operating system	32-bit MS Windows (NT/2000/XP)

Dependencies	Unknown/unclear
Tool category	Metadata validator
License	Unknown/unclear
Description	A program developed at the University of California, Berkeley, to batch validate large numbers of METS documents against the official METS schema and related schemata. When batch validation is complete, error messages are displayed in a browser window, showing a list of the files with errors, the location of each error, and an error message. Unlike most XML parsers, Xerces is configured not to abort the validation process at the first error, meaning the user can view all of the errors in an XML document, rather than just the first.
Output method	Output to screen?
Maturity	
Notes	The program has been used internally at Berkeley, and has not yet been extensively tested. It is unlikely to work with DTDs.

Appendix 1 : Common Open Source Licenses

Some of the more common standard licenses under which tools are made available are listed here. All of these have been approved by the Open Source Initiative.[7] The list includes a brief overview of each license (focusing on the elements which are likely to be most relevant for Cairo purposes), a link to its full text, and a note on its compatibility with other licenses. Compatibility is important in the context of Cairo, which envisages combining some existing tools as components of the final Cairo product. Where licenses for existing software packages are incompatible, we may be prevented from using tools in combination. The most widely used open-source license is the GNU General Public License; this license states that any modified piece of software which contains elements derived from an original GPL-licensed program should be licensed under the terms of the GPL. Compatibility with the GPL is therefore likely to be an important factor in the selection of tools for incorporation into Cairo. Of the licenses covered in this document, it may be advisable to avoid, or think carefully about how we use in combination, software issued under: the Academic Free License; the Original Artistic License; the Common Public License; and the Open Software License. The Free Software Foundation (FSF), which publishes the GNU GPL, has produced lists of licenses it considers to be compatible and incompatible with the GPL; this can be found at http://www.fsf.org/licensing/licenses/ and provides the source of most of the compatibility information given below.

All of the licenses include warranty disclaimers, which make clear that the work covered by the license is made available without warranty, and that the risk as to the quality of the original lies entirely with the user. Most of them also stress that if any modification is made to the work covered by the license, this should be made clear in the modified version. The majority of the licenses (with the exception of the GPL, LGPL and MPL) also expressly prohibit using the name of the original creator to endorse or promote products derived from the original software covered by the license.

Academic Free License (AFL)

This license (currently at version 3.0) uses the term ‘Original Work’ to denote the tool/piece of software covered by the license.

Three principal grants are made under the license:

- Grant of copyright license: this permits the user to reproduce the original work, and adapt or otherwise transform it to create derivative works; and to distribute copies of the original or derivative works to the public under any license that does not contradict the AFL .
- Grant of patent license: granting the non-exclusive right to make, use and sell the original or derivative works worldwide.
- Grant of source code license: the licensor must provide a machine-readable copy of the source code of the original work along with the work; it may also fulfil this obligation by placing a copy of the source code in an easily accessible information repository.

[7]Open Source Initiative http://www.opensource.org/

Conditions of the license include:

- The stipulation to include an ‘Attribution Notice’ in the source code of any derivative work, which reproduces all IPR notices from the original work, and makes clear that the original work has been modified.
- Anyone distributing copies of the original work or derivative works must make a ‘reasonable effort under the circumstances’ to obtain the express assent of recipients to the terms and conditions of the AFL.

Compatibility with GNU GPL:

The FSF states that versions 1.2 and 2.1 of the AFL are incompatible with the GNU GPL; while it does not comment on version 3.0, the requirement to obtain the express assent of recipients is given by the FSF as a reason why the Open Software License is incompatible with the GNU GPL. It therefore seems likely that version 3.0 of the AFL would also be considered incompatible.

Text of the AFL:

- Academic Free License v3.0

Artistic License

This license uses the term ‘Package’ to denote the tool/software covered by the license. The ‘Standard Version’ of the package is the original version in its unmodified form, or a version which has been modified in accordance with the wishes of the copyright holder. There are at least three versions of this license – the original license (and some versions of this also differ slightly), version 2.0, and the Clarified Artistic License.

The principal grants made by all versions are:

- The right to copy and give away the standard version of the package, provided all original copyright notices and disclaimers are duplicated.
- The right to modify the package and distribute the modified version, provided a prominent notice of changes made (with dates) is inserted in each modified file, and provided one or more other stipulated conditions are met. These conditions vary from one version to another, and include: placing the modifications in the public domain or making them freely available by other means; renaming any executables which have been altered from the standard version, and providing information on their modification; and allowing anyone who receives a copy of the modified version to make the source form of the modified version available to others under the original license or a similar open license.
- The right to distribute the programs of the package in object code or executable form, provided one or more of certain conditions are met, the most relevant of which is likely to be: accompanying the distribution with the machine-readable source of the package with any modifications which have been made.
- The right to distribute the package in aggregate (‘possibly commercial’) with other programs as part of a larger (‘possibly commercial’) software

distribution, as long as the package is not advertised as a product of the user's own.

Compatibility with GNU GPL:

The FSF criticizes the Original Artistic License for vagueness, and suggests it does not qualify as a free software license ('free' in this case meaning freedom to distribute copies, access to source code, and the right to modify or use pieces of the software in other free programs). Version 2.0 of the license was produced by the Perl Foundation; this is a free software license, and is compatible with the GNU GPL. However, FSF advises against its use, except as part of the disjunctive license of Perl. The Clarified Artistic License is a free software license, and is compatible with the GNU GPL.

Text of the Artistic License:

- Original Artistic License as given by the Open Source Initiative
- Artistic License v2.0
- Clarified Artistic License

The BSD License

The BSD License (named for Berkeley Software Distribution, the operating system for which it was originally created) has few restrictions compared with some of the other free software licenses. It grants the user the right to redistribute and use the software covered by the license, in source and binary forms, with or without modification, provided that redistributions retain the copyright notice, list of conditions, and warranty disclaimer given in the license template.

The current version of the license is a revision of the original version which omits an advertising clause that caused widespread criticism. It is sometimes referred to by other names, including 'Modified BSD License', '3-clause BSD license', 'BSD-new' and 'revised BSD'.

Compatibility with GNU GPL:

The BSD License allows commercial use, and it also allows for a product to be distributed with both a BSD and a different license attached. The Modified BSD license is compatible with GNU GPL, although the FSF highlights the problem of referring to it as 'the BSD license' which could lead to confusion with the original, deprecated, version. It recommends using the X11 license in preference, which is more or less equivalent to the modified BSD license.

Text of the BSD License:

- BSD License

Common Public License

Currently at version 1.0, this license outlines a number of grants made by the 'contributors' to a software program (i.e. the initial creator and any subsequent contributors who have made changes to the program).

The principal grants made under this license are:

- The right to reproduce, prepare derivative works of, distribute and sublicense each contribution to the program in source and object code form.
- The right to a worldwide, royalty-free patent license to make, use, sell, import and otherwise transfer each contribution to the program in source and object code form (with a liability disclaimer for contributors).

The conditions include:

- If a contributor distributes the program in object code form under their own license agreement, the license must: comply with the terms and conditions of the Common Public License; include warranty and liability disclaimers; state that any provisions differing from the Common Public License are offered by that contributor alone; and make clear how the source code can be obtained.
- If a contributor distributes the program in source code form, it must be made available under the Common Public License.
- Copyright notices must not be removed or altered, and each contributor must identify themselves as the originator of their contribution.

Compatibility with the GNU GPL:

The Common Public License is incompatible with GNU GPL because it has various requirements that are not included in the GPL.

Text of the Common Public License:

- Common Public License

Educational Community License v. 1.0

Currently at version 1.0, this license has few restrictions. It grants the user permission to use, copy, modify, merge, publish, distribute and sublicense the original work covered by the license, and its documentation (with or without modification) for any purpose, and without fee or royalty, provided that all copies of the original or modified work include: the full text of the license; any pre-existing IPR disclaimers or other notices; and a notice of any modifications to the original work (including dates).

Compatibility with GNU GPL:

The FSF does not include the Educational Community License in its GPL compatibility lists. However, the OpenSouceLegal.org website, which focuses

on legal issues in relation to open source software, suggests that it is compatible.[8]

Text of the Educational Community License:

- Educational Community License 1.0

GNU General Public License (GPL)

The GNU GPL is currently at version 2, but the versions are largely backwards and forwards compatible, so when a version number is not specified the user can choose from any version. The GPL is a widely-used license, under which many of the tools listed in this document are released. The license refers to the tool/software covered by the license as a 'Program'.

The principal grants made under this license are:

- The right to copy and distribute verbatim copies of the original program's source code, provided certain conditions are met (including the reproduction of notices/disclaimers, and providing a copy of the license).
- The right to modify all or part of the original program, and to copy and distribute the modified version under the same terms as verbatim copies, provided certain other conditions are met, including: prominent notices of changes made to modified files (with dates); any program which is in whole or in part derived from the original program should be licensed as a whole at no charge under the terms of the GNU GPL; and if the modified program reads commands interactively, it must at starting display copyright notice and warranty disclaimer. This means that any modified program containing elements derived from the original GPL-licensed program must, when distributed as a whole, be licensed under the terms of the GPL. This does not apply where independent sections of the modified version, which were not derived from the original program, are distributed as separate works.
- The right to copy and distribute the program (unmodified or modified) in object code or executable form, under the same terms as above, provided one of three other conditions are met, principally to make available the corresponding machine-readable source code.

The license contains numerous other clauses aimed to guarantee that the program covered by the license meets all the conditions for free software.

Compatibility:

The Free Software Foundation provides a list of licenses which are and are not considered compatible with the GNU GPL.[9]

Text of the GNU General Public License:

- GNU General Public License v. 2.0

[8] OpenSourceLegal.org http://www.opensourcelegal.org/licensedb/detail.php?lid=54&SEARCH=Educational%20Community%20License

[9] Free Software Foundation – Licenses http://www.fsf.org/licensing/licenses/

GNU Lesser (or Library) Public License (LGPL)

The GNU Lesser General Public License 2.1 is the successor to the GNU Library Public License, version 2. The LGPL is slightly more flexible than the GPL with regard to combining other programs with the program covered by the license. The LGPL is primarily intended for software libraries, and hence the work covered by the license is defined as a 'Library' (meaning a collection of software functions and/or data which can be linked with application programs to form executables); however it may also be used for standalone applications.

The license also defines a 'work that uses the library', which is a program that is not derived from the library but is a standalone executable that dynamically links to it. Taken alone a 'work that uses the library' is therefore not covered by the LGPL. However, linking a 'work that uses the library' with the library creates an executable that is a derivative of the library and is therefore covered by the license, although there is an exception to this (see the fourth point below).

The principal grants made by the LGPL are:

- The right to copy and distribute verbatim copies of the library, provided all notices and disclaimers are published, and a copy of the license distributed.
- The right to modify the library and distribute the modified version under the same terms as verbatim copies provided that certain other conditions are met, including: files should carry prominent notices of any changes made (with dates); and the whole work should be licensed at no charge under the terms of the LGPL (except where independent sections which were not derived from the original are distributed as separate works).
- The right to copy and distribute part or whole of the library (or derivative of it) in object code or executable form, under the above terms, provided it is accompanied by the corresponding machine-readable source code, or the source code is made accessible from a designated location.
- The right to combine or link a 'work that uses the library' with the library to produce a work containing portions of the Library, and distribute this under terms of the user's choice, provided: the terms permit modification of the work by the customer and reverse engineering for debugging such modifications; a notice indicates that the library is used in the work, and is covered by the LGPL; a copy of the LGPL is supplied; and copyright notices are included. One of five further conditions must also be met, the most relevant being: to accompany the work with (or place in an accessible location) the corresponding machine-readable source code for the library, including changes made; and if the work is an executable linked with the library, to accompany it with the complete machine-readable 'work that uses the library' as object and/or source code.
- The right to place library facilities derived from the library side-by-side in a single library together with other library facilities not covered by the LGPL, and distribute this combined library, provided: the combined Library is accompanied by a copy of the uncombined work based on the

Library; and prominent notice is given of the fact that part of it is based on the original Library, giving its location.

Compatibility with GNU GPL and other licenses:

The license is compatible with the GPL, and will also be compatible with the other licenses which are compatible with the GPL. [10]

Text of the GNU Lesser (Library) Public License:

- GNU Lesser (Library) Public License v. 2.1

Mozilla Public License 1.1 (MPL 1.1)

The MPL 1.1 is somewhat more complex in structure than the other licenses.

The grants made by the license are divided into grants made by the Initial Developer and grants made by each contributor who has modified the original version. The grants made are essentially the same in each case, and are:

- The right to use, reproduce, modify, sublicense and distribute the original code or modifications to it (or parts thereof), with or without further modifications and/or as part of a larger work.
- The right to make, use, sell, and otherwise dispose of the original code and/or modifications (in whole or in part).

The conditions include:

- Modifications: are governed by the terms of the MPL; must be made available in source code form under the MPL; and must include a statement documenting changes and identifying the original developer.
- Relevant notices and the license must be reproduced in modified versions.
- If the code covered by the license is distributed in executable form (i.e. any form other than the source code): a prominent notice should state that the source code version is available under the terms of the MPL, and outline how the distributor has fulfilled the obligations of the license.
- If the code is distributed in executable form, it may be distributed under a license other than the MPL, provided the user is in compliance with the MPL and the license does not limit/alter the recipient's rights in the source code version from the MPL rights.
- Where the user creates a larger work by combining the original code with other code not governed by the terms of the MPL and distributes this as a single product, the terms of the MPL only apply to the original code.
- There is also a clause relating to termination of the license, in particular termination for patent action (as with the Open Software License).

Compatibility with the GNU GPL:

The MPL is not considered compatible with the GNU GPL. However, the provision which allows for combining the original with other code under a

[10]Free Software Foundation – Licenses http://www.fsf.org/licensing/licenses/

different license may mean that the GNU GPL (or a GPL compatible license) could be used for that part of the program.

Text of the Mozilla Public License:

- Mozilla Public License v. 1.1

Open Software License (OSL) v. 3.0

The tool/software covered by the license is referred to as the Original Work. This license, as with the AFL, contains three principal grants:

- Grant of copyright license: permitting the user to reproduce the original work alone or as part of a collective work; to modify or otherwise transform the original to create derivative works; and to distribute copies of both original and derivative works to the public, provided they are licensed under the Open Software License.
- Grant of patent license: granting the non-exclusive right to make use, sell, import the original and any derivative works.
- Grant of source code license: the licensor must provide a machine-readable copy of the source code of the original work along with each copy it distributes, or in an accessible information repository.

Conditions of the license include:

- The stipulation to include an 'Attribution Notice' in the source code of any derivative work, which reproduces all copyright, patent or trademark notices from the original work, and makes clear that the original work has been modified.
- When distributing the original or derivative works, the user must make a 'reasonable effort under the circumstances' to obtain the express assent of recipients to the terms of the Open Software License.
- Termination for patent action: the license includes a clause stating explicitly that the license terminates automatically if the user commences a legal action against the licensor or any licensee alleging that the original work infringes a patent.

Compatibility with the GNU GPL:

The Open Software License, version 1.0 is considered incompatible with GNU GPL in several ways, notably the requirement for distributors to try to obtain explicit assent to the terms of the license.

Text of the Open Software License:

- Open Software License v. 3.0