

Automatic Metadata Generation: Use Cases and Tools/Priorities

Guidance on different automated metadata generation approaches for service providers in HE

This report was produced for JISC by Intrallect Ltd (Charles Duncan and Peter Douglas)

24 August 2009

Contents

Background and Context	3
Approach	4
Business Cases	5
Cost.....	5
Benefits	5
Tools and Services	6
Automatic recognition/extraction tools and services	6
Authoritative source tools and services.....	9
Translation tools and services	11
Metadata quality validation services.....	12
Activity aggregation services.....	12
Relationship services.....	12
Examples of services using these tools.....	13
Implementing Services	13



Background and Context

This report brings together guidelines derived from a short study (March – July 2009) on automatic metadata generation. The purpose of this study was to identify a set of use cases which illustrate the role and importance of automatic metadata generation and, as a result, to identify tools and services – both available and required – to provide guidelines for service providers and to recommend priorities for further work.

This report is one of several outputs of the project. All the project reports and their target audiences are listed below:

Guidelines Report (this report): This is intended for anyone considering implementing automatic metadata generation. It offers some cost-benefit analysis which can support building a business case, a summary of available tools and services, and an outline of how these services can be used and integrated with other systems.

Recommendations: This is a restricted access report for JISC providing recommendations for future research and development of tools and services for automatic metadata generation.

Synthesis Report: This is intended for anyone with an interest in Automatic Metadata Generation and assumes no previous knowledge of the topic. It provides an overview of the topic and is recommended as an introduction to the other reports.

Specialist Reports: These reports were commissioned as part of this study and provide an overview of a particular topic in automatic metadata generation including recommendations for further work by each of the report authors. The reports are on:

- Subject metadata
- Name metadata
- Geospatial metadata
- Factual metadata
- Bibliographic metadata
- Usage metadata
- File format metadata
- Integrating automatic metadata services
- Automatic language translation of metadata

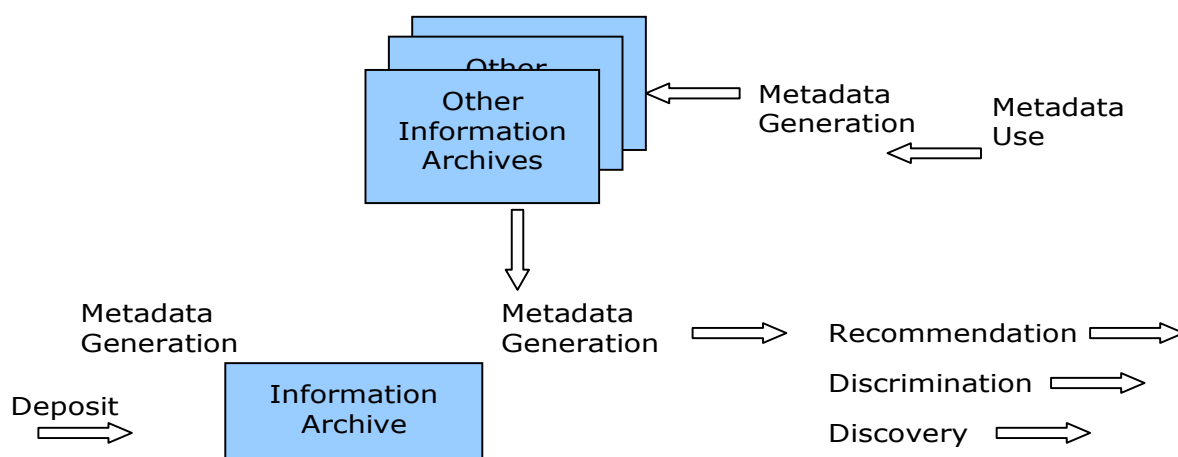
We wish to record our thanks to the authors of these commissioned reports. Their reports have played a major part in assembling this report.



Approach

Metadata is important. The traditional view of metadata is that it is an aid to resource *discovery*. However, our access to information is now so ubiquitous that discovery is more often about filtering important information from the multitude in which it is swamped. In this sense it may be better to describe metadata as an aid to *discrimination*. Metadata supports discrimination by providing additional information, often about relationships that helps clear a path to discovery.

A further use of metadata is for *recommendation*. While discovery and discrimination are based on activities and choices dictated by the seeker, recommendation systems push information based on the activities and choices of others. As a result they can effect discovery of resources the seeker might never have otherwise considered. So, in this study we have considered that metadata will be used for *discovery*, *discrimination* and *recommendation* and the use cases in the Synthesis Report cover all of these purposes.



Of course, metadata is often gathered at the time of ingest. But additional metadata can also be generated at the time of discovery. This additional metadata could be further details about authors, automatic translation of metadata into other languages, links to other versions of the same resource, indications of the popularity of the resource, to name but a few.

Much of this metadata, though not all of it, could also have been generated at the time of ingest. It is important to consider whether metadata should always be gathered as early as possible in the process so that the work of producing it can be carried out once and maximise the benefits. Or, should metadata be generated on the fly as it is demanded by users, which offers a great deal of flexibility. In cases where there is a licence cost associated with automatic metadata generation it is necessary to consider if just-in-case metadata generation (at the time of ingest) is costly and the metadata may never be used, or if just-in-time metadata generation is better because the cost is applied where user demand justifies it.

Some metadata, particularly usage metadata, cannot be created at ingest and is constantly updated.



Business Cases

Cost

Gathering metadata can be expensive and time-consuming. Although the time taken for cataloguers to create or refine metadata is not often published some indications of the effort involved are available.

In 2003¹ an exercise to improve metadata which was perceived to be of insufficient quality by using cataloguers to re-edit 2,500 metadata records took about 550 hours and cost around £6,500, or about 13 minutes or £2.60 per record to apply quality assurance to existing metadata.

In 2005² a study of self-archiving by researchers found "The median time for metadata entry is 5 minutes and 37 seconds per paper. The average is 10 minutes 40 seconds owing to the long tail of the distribution." and "the average and median number of keystrokes per record is 1500 and 970 respectively. (This is an overestimate because the paper title and abstract are frequently cut and paste from an online catalogue, or from the paper itself.)"

Intute has a considerable track record of creating metadata and calculates the average time to create a new metadata record as 22.6 minutes³. As Intute also discovers and selects its resources prior to cataloguing the overall task takes an average of 31 minutes.

An unknown, but often perceived, cost is that the effort involved in self-archiving acts as a barrier to deposit of digital resources into information archives.

Benefits

The headline benefits of automatic generation of metadata are:

- Saving of time and effort by those carrying out cataloguing
- Reduction of an important barrier to deposit
- Enhanced quality and consistency of metadata
- Improved discovery, discrimination or recommendation of resources as a result of richer metadata
- Increased "linkage" between people, their digital resources, and related resources making it easier to find collaborators, build communities and exploit synergies

¹ "Building Quality Assurance into Metadata Creation", Jane Barton, Sarah Currier, Jessie M. N. Hey, International Conference on Dublin Core and Metadata Applications, Seattle, 2003 (<http://dcpapers.dublincore.org/ojs/pubs/article/view/732/728>)

² "Keystroke Economy: A study of the time and effort involved in self-archiving", L.Carr and S.Harnad, Technical report, ECS, University of Southampton, 2005 (<http://eprints.ecs.soton.ac.uk/10688/>)

³ Personal communication. Jackie Wickham, Intute




Tools and Services


This list of available tools has been subdivided into the six categories that are used in the other reports of this study.

The first three categories are well served by available tools and services while the last three are emerging technologies and few tools or services exist, at least for the academic domain.

Automatic recognition/extraction tools and services

A previous JISC project, MetaTools⁴, carried out a detailed evaluation and comparison of several of the tools mentioned below. The tools contained in the MetaTools report are highlighted with a  beside their name.

Name	Termine
URL	http://www.nactem.ac.uk/software/termine/
Description:	Termine identifies terms in submitted text. The terms are recognised by the C-value method and acronyms recognised by AcroMine. C-value term extraction requires a fair amount of text to produce reasonable termhood scores, as these rely on the key terms occurring multiple times.
Conditions of use:	Open access for non-commercial purposes. Batch service available for files larger than 2Mb and SOAP service available for integration

Name	Yahoo Term Extraction Service 
URL	http://developer.yahoo.com/search/content/V2/termExtraction.html
Description:	A service in which text is submitted and an XML record listing extracted terms is returned.
Conditions of use:	An open access service which is rate-limited. Each user (IP address) is able to submit up to a fixed number of queries each 24 hours. The number varies with application ID. You can apply for unique applications IDs.


Name	GeoDoc
URL	http://www.gogeo.ac.uk/cgi-bin/mdres.cgi
Description:	The GeoDoc Metadata Editor Tool offers the functionality to support partial automation of geospatial metadata creation and


⁴ <http://www.jisc.ac.uk/media/documents/programmes/reppres/metatoolsfinalreport.pdf>



	publication on a geo-portal. GeoDoc also supports a range of standards for the export of metadata records into XML and PDF files.
Conditions of use:	Requires a Shibboleth (UK Access Management Federation) account to log in.


Name	GeoParser
URL	http://edina.ac.uk/projects/geoxwalk/geoparser.html
Description:	The GeoParser demonstrator is a tool that allows users to upload web pages, text files, metadata records, xml etc., which can then be parsed for geographical names. These are then checked against GeoCrossWalk to obtain explicit geographical coordinates for the location referred to, in order to "geo-tag" the uploaded document.
Conditions of use:	If you are interested in using the GeoParser Demonstrator, please get in touch with the GeoCrossWalk Team: gxw@ed.ac.uk

Name	SamgI (Simple Automated Metadata Generation Interface) 
URL	http://www.ariadne-eu.org/index.php?option=com_content&task=view&id=40&Itemid=56
Description:	A suite of software tools for extracting metadata automatically.
Conditions of use:	Available as open source. Downloadable from Sourceforge. Can be run in stand-alone mode or as a web service.

Name	DC-DOT 
URL	http://www.ukoln.ac.uk/systems/tools/dc-dot/
Description:	DC-Dot is a Dublin Core metadata generator. In its original form it retrieves a web page, analyses it, and generates Dublin Core metadata. DC-dot was substantially updated in the early spring of 2009 with the intent to: <ul style="list-style-type: none"> • Update the level of compliance with coding standards • Improve reusability • Increase the number of input filters (ie. input file formats) • Increase the number of output formats (metadata formats) • Improve the accuracy/relevancy of suggested metadata



Conditions of use:	The original version of DC-Dot is free to use. Conditions of use of the updated version are not yet announced.
--------------------	--

Name	iVia DataFountains 
URL	http://ivia.ucr.edu
Description:	A suite of digital library tools that include a metadata extractor that reads a URL, downloads the web page content, and attempts to extract its Title, Creator, Keywords and Description using natural language processing.
Conditions of use:	Available as open source code.


Name	paperBase
URL	http://www.ukoln.ac.uk/systems/tools/paperbase/
Description:	PaperBase is a formal metadata extraction system that makes use of Bayesian statistics and a hidden Markov model (HMM) approach to extract relevant facts from the full text of documents. These facts include author, title, number of pages, place of publication and so forth.
Conditions of use:	Not yet known.

Name	DROID
URL	http://droid.sourceforge.net/wiki/index.php/Introduction
Description:	DROID was developed by the National Archives. DROID allows files and folders to be selected from a file system for identification. DROID identifies the format of the file. After the identification process had been run, the results can be output in XML, CSV or printer-friendly formats.
Conditions of use:	DROID is available for download under an open source licence

Name	JHOVE (JSTOR Harvard Object Validation Environment)
URL	http://hul.harvard.edu/jhove/
Description:	JHOVE (JSTOR Harvard Object Validation Environment) was developed by JSTOR and Harvard University Library. The tool can be used validate and characterise identified formats. The



	JHOVE distribution includes the following standard modules: AIFF, ASCII, BYTESTREAM, GIF, HTML, JPEG, JPEG 2000, PDF, TIFF, UTF-8, WAVE, XML.
Conditions of use:	JHOVE is available as downloadable open source code under a LGPL licence

Name	Metadata Extraction tool ⁵ (KEA )
URL	http://meta-extractor.sourceforge.net/
Description:	The Metadata Extraction Tool is designed to automatically extract preservation-related metadata from digital files and output that metadata in a standard format (XML) for use in preservation activities. The Tool includes a number of 'adapters' that extract metadata from specific file types including: Images: BMP, GIF, JPEG and TIFF; Office documents: MS Word (version 2, 6), Word Perfect, Open Office (version 1), MS Works, MS Excel, MS PowerPoint, and PDF; Audio and Video: WAV and MP3; Markup languages: HTML and XML.
Conditions of use:	The Tool is written in Java and XML and is distributed under the Apache Public License (version 2)

Authoritative source tools and services

In addition to the services mentioned below there are a number of commercial providers building databases of academic's names from publications. These include Thomson Reuter's ResearcherID⁶, Elsevier's Scopus Author Identifier⁷, ProQuest's ScholarUniverse⁸. Other commercial services that combine name information with publications are provided by services such as Symplectic's Publications Management System⁹ and Textensor's PublicationsList¹⁰.

Name	Sherpa/Romeo
URL	http://www.sherpa.ac.uk/romeo/api.html
Description:	This service provides authoritative information on journal title, ISSN and publisher's name, although there are also some special searches, such as for RoMEO "colours". The basic principle is that an application makes an HTTP request to the API, which returns an XML stream with the search results. The

⁵ <http://meta-extractor.sourceforge.net/>

⁶ <http://www.researcherid.com/>

⁷ <http://info.scopus.com/etc/authoridentifier/>

⁸ <http://www.scholaruniverse.com/>

⁹ <http://www.symplectic.co.uk/products.html>

¹⁰ <http://publicationslist.org/>



	query specification is given in the URL's parameters.
Conditions of use:	<p>The information provided through the SHERPA/RoMEO API is given freely to interested parties for their re-use, although the following conditions apply:</p> <ul style="list-style-type: none"> • The information provided by the SHERPA/RoMEO service is not for commercial re-use. • It would be appreciated if the SHERPA/RoMEO logo could be incorporated somewhere on a relevant public page which uses SHERPA/RoMEO information and make an active link to http://www.sherpa.ac.uk/romeo.php • The SHERPA/RoMEO logo is available in various sizes, and can be downloaded from http://www.sherpa.ac.uk/images/romeologos.html Wherever possible, please include the logo on the page where the information is presented. • If the SHERPA/RoMEO information is re-interpreted or modified, then we ask for the following text to be used in addition to the logo in an appropriate place: "This information is derived from the SHERPA/RoMEO database and has been modified for use here."

Name	Name Authority Pilot
URL	http://130.88.120.172:8080/
Description:	This service provides authoritative names for individuals and institutions. A sample implementation is provided at http://names.mimas.ac.uk/script-test/ . The service can return records in several formats including JSON, RDF, MarcXML and Names format XML.
Conditions of use:	This pilot project is openly available to encourage feedback. Conditions of use of any future service have yet to be determined.

Name	PRONOM
URL	http://www.nationalarchives.gov.uk/PRONOM/Default.aspx
Description:	<p>PRONOM holds information about file formats, and the software products which can process (read, write, identify etc) each format. Information related to the file formats, such as documentation about them, their compression types, character encoding schemes and intellectual property rights is also held. A full description of the individual fields used by PRONOM is available in the system documentation.</p> <p>PRONOM is working with the Global Digital Format Registry</p>



	(administered by Harvard University Library) to reate the Unified Digital Formats Registry
Conditions of use:	Pronom is commonly used as a web application but is also available as a web service (used, for example by DROID).

Translation tools and services

Translation tools and services are those which translate data from one form to another. Some translation services are available to automatically translate metadata from one language to another but these have not been considered here.

Name	GeoCrosswalk
URL	http://www.geoxwalk.ac.uk/
Description:	GeoCrossWalk is JISC funded middleware implementing a digital gazetteer service and server for the UK academic Higher and Further Education community. The rationale behind the service is that there is currently no unified entry point to assist in geographic searching within the existing academic network, as each information provider/service adopts different geographic coding conventions (some use postcodes, others placenames, some grid references etc.). GeoCrossWalk is designed to make geographic searching transparent by 'crosswalking' these different geographies as illustrated below.
Conditions of use:	As GeoCrossWalk is based on Ordnance Survey data it is subject to licensing restrictions although JISC has agreements in place to permit use within HE institutions.

Name	GeoNames
URL	http://www.geonames.org
Description:	This service converts geospatial information from one form to another on a global basis. The GeoNames geographical database contains over eight million geographical names and consists of 6.5 million unique features whereof 2.2 million populated places and 1.8 million alternate names. All features are categorized into one out of nine feature classes and further subcategorized into one out of 645 feature codes.
Conditions of use:	The GeoNames geographical database is available for download free of charge under a creative commons attribution license. It is also available as a set of web services.



Metadata quality validation services

There are no systems for metadata quality validation currently available to UK HE Institutions. A system for institutional repositories in New Zealand administered by KRIS (Kiwi Research Information Service)¹¹ provides regular quality reports on the metadata in NZ institutional repositories.

Activity aggregation services

An "activity" record is metadata that records someone's use of a resource. The most commonly quoted examples are the recommendations used by Amazon of the form "people who bought that book also bought these..." Work is at a very early stage in developing "activity" records for aggregations and sharing. Guidelines¹² have been produced by the JISC MOSAIC project. The type of activity data that could be shared includes:

- Local library circulation data
- Reading lists
- Circulation & Inter-lending modules of Library Management systems (LMS)
- Virtual Learning Environments (VLE)
- OpenURL resolver / ERM derived information
- Institutional and Open Education Resource Repositories

Relationship services

Although no academic relationship services have been established yet there are several general systems that can act as models.

- DBpedia¹³ offers structured information from Wikipedia and includes more than 2.6 million things, including at least 213,000 persons and 328,000 places. It includes 609,000 links to images and 3,150,000 links to external web pages as well as 4,878,100 external links into other RDF datasets.
- Freebase¹⁴ is a very large open database that can be queried through web services¹⁵ which as well as full text searching also includes geosearch¹⁶.

Perhaps the closest academic example is CiteSeer¹⁷ which not only includes searches of publications but can search the citations and show the context of the citation in the original publication. However, a truly useful academic relationship service would go beyond publications to include authors and research projects or perhaps learning resources and courses that use them.

¹¹ <http://nzresearch.org.nz/index.php/reports/indexMetadataQualityReports>

¹² http://library.hud.ac.uk/wikis/mosaic/images/2/2d/Mosaic_data_collection_-_A_guide_v01.pdf

¹³ <http://dbpedia.org/>

¹⁴ <http://www.freebase.com/>

¹⁵ http://www.freebase.com/docs/web_services

¹⁶ <http://www.freebase.com/docs/geosearch>

¹⁷ <http://citeseer.ist.psu.edu>



Examples of services using these tools

Some examples of services which are already making use of the types of services described above are:

- Go-Geo¹⁸: The Go-Geo geospatial information service offered by Edina shows how a graphical interface can be used to search for geospatial information. Its advanced search uses a Gazetteer and GeoCrossWalk to allow geographic locations to be provided as place names, postcodes, regions or map coordinates.
- Intute Repository Search¹⁹: The IRS offers two innovative ways of searching across multiple research repositories and displaying the results in ways that reveal relationships between the clusters of keywords that are found by text-mining systems. These services include a visual cluster-map view of related concepts or terms.

Implementing Services

In order to make use of the tools and services described above some technical skills are required. We have deliberately used the term “tools” to refer to software that you can obtain and install on your own servers, and “services” to refer to software on remote servers to which you can send files and queries and from which you can obtain a response. These services are generally designed to be used on a machine-to-machine basis so the service is accessed using a system-independent web services interface (usually REST or SOAP) and results are return in a system-independent form (usually in XML or JSON). However, RESTful web services are also easy to use from a web browser and they often offer a web site on which they can be tested.

This means that the skills required to implement any of these services vary a great deal:

- RESTful web services can be quickly tested to determine whether or not they will be useful and can be implemented by people with limited technical skills.
- SOAP-based web services require a higher degree of technical capability.
- Tools which need to be installed, and may need to be adapted, require the highest level of skills as they will need both system administrator and programmer skills which may be language-specific. Even then, some tools may be easy to use and may have a significant community of developers offering support while others may be more difficult with little support.

As part of this study a specialist report “Integrating Automatic Metadata Generation into Deposit Workflows” has been prepared by Richard Green at the University of Hull

¹⁸ <http://www.gogeo.ac.uk/>

¹⁹ <http://www.intute.ac.uk/irs/>



who has experience of these approaches through the RepoMMan and REMAP projects. That report is available from the same location as this report.



This work is licensed under a [Creative Commons Attribution 2.5 License](https://creativecommons.org/licenses/by/2.5/)