Automated Metadata

A review of existing and potential metadata automation within Jorum and an overview of other automation systems

Authors

Kenny Baird

Jorum Team

Date

31st March 2006

Version

1.0

Status

Final

Authorised

Signed off by JISC and Intrallect July 2006

JISC

EDINA MIMAS

JISC-designated national data centres

Contents

1. Executive Summary

The benefits of automated metadata are savings in time and human resources, and in certain cases, an increased level of consistency when compared against human-created metadata. Many organisations are looking at automated metadata systems to reap these benefits. This is evidenced by the large number of projects and companies who are creating programmes which automate metadata.

However, the increased application of systems and process to automate metadata will not result – in the foreseeable future at least – in the obsolescence of the human in the metadata creation process. Whereas a computer can read, say, an IMS Manifest file and record all references of technical formats much faster than a human, a skilled cataloguer is able to make judgements on the practical application of the described resource within a learning environment in ways a computer cannot. Therefore, much of the metadata which can be automatically generated relates to its technical properties, repository users will typically need more subjective metadata to enable them to asses their retrieval results. The ideal situation is the two approaches to metadata creation working in tandem, with as much automated as possible to allow cataloguers to spend greater time on creation of metadata that cannot reasonably be expected to be automated. Concise, accurate recordings of technical properties and other elements which are consistent and overly mundane to warrant repetition on creation (such as vCard details), allied with human catalogued entries will provide the discovery user with both an overview of a resources' properties and limitations, and also allow the user to make a quick judgement on the relevance of the retrieved resource for a particular educational/learning context.

This report comprises three main strands; an analysis of what Jorum is currently doing, and an initial assessment of what it could do. It also presents an overview of other some systems, services and products which claim automated metadata generation as either a direct or indirect output of their aims.

The Jorum repository is powered by intraLibrary (currently at build version 2.4), provided by Intrallect[1]). It automates several elements of the Jorum Application Profile, and though there is scope for increased automation, there comes a point where return on investment may become prohibitive. The development of algorithms and processes to allow the software to attempt automation of the elements more suited to human completion needs to have benefits off set

[1] Intrallect: http://www.intrallect.com

against factors such as system resources and developmental costs.

While analysis of other automation systems has yielded benefits and suggestions, there is no system which has been identified through this report's limited scope, which has significantly more automation, or provides higher levels of accuracy than intraLibrary. However, as this is an emergent landscape, work should be done periodically to revisit this area of the report to ensure that any benefits or lessons published can be assessed and possibly utilised by Jorum. One area where this could be applicable is in the scoping of the Jorum Client tool, and this is discussed further in section 6.1.2.

More work also needs to be done in terms of taking the recommendations listed in this report and concept checking their viability from a system / programming viewpoint. This further work then needs to be cross checked against further research into the usefulness of this metadata for end users. As part of a separate research strand, work is being done within Jorum to map metadata elements to what users search for, and prior to the recommendations made in this automated metadata report being implemented, they should be cross checked against the data from this further research to derive a cost benefit proposal. This should be then offset against other implications of the recommendations, such as server capabilities and process performance of the Jorum software. If a recommendation is made in this automated metadata report, but a consequence of its implementation is an increased slowdown of the contribution process because of increased system requirements, then the recommendation should not be implemented.

Similarly, there would be little point in automating metadata entries if evaluation showed that these entries were not used for research discovery - unless it was found that unclear guidelines were the reason for this disuse. For instance, a recommendation made in this report is the mapping of element 4.1 technical format to 4.4.1.2 Name. However, if this either results in a slowdown of the contribution process due to the computing power needed to do this mapping, or people do not search for 4.4.1.2 Name, then the recommendation should not be implemented.

2. Scope and Introduction

2.1 Scope

This report will look at the model Jorum currently uses for automatic metadata generation to identify further metadata elements that could easily and effectively be automated. It will examine the benefits and drawbacks of automated metadata processes, as well as looking at systems which are available to automate the metadata generation process. Due to resource limitations, this report does not provide a comprehensive overview of systems providing automatic metadata generation. Indeed, identifying a comprehensive list of repositories or tools related to automatic metadata generation for learning objects is in itself a sizeable task, and would be a useful output from the JISC Digital Repositories Programme.

Further to this, the focus of the report is on metadata for resource discovery and providing information to users rather than for other purposes such as preservation or advocacy and further work would be required to address other metadata use cases regarding automation.

This report does not examine the usefulness of the UK LOM Core for resource description or discovery.

This report has been produced independently of Intrallect; while informal consultation over automation of metadata has occurred, claims made about intraLibrary are made solely on the Jorum team's experience and understanding of the software. Recommendations made here for Jorum which would involve further developments to intraLibrary should be passed to Intrallect for comment.

2.2 Definitions

2.2.1 Automated Metadata

Automatic metadata generation is an output of machine processing and, other than engineers developing the software and Contributors or cataloguers initiating the generating process, has no other human interaction.[2]

[2] Greenberg, J., Spurgin, K. & Crystal, A. Functionalities for Automatic-Metadata Generation Applications: A Survey of Metadata Experts' Opinions. International Journal of Metadata, Semantics, and Ontologies. 2006, *Vol. 1, No. 1, 2006* 3 http://www.inderscience.com/storage/f121932106117458.pdf

2.2.2 Single File

A single file is defined in this report as a single asset contributed to Jorum, such as a word document or PDF.

2.2.3 Content Package

A content package is defined in this report as a collection of resources packaged together and conforming to the IMS Content Package specification v1.2[3]

Within Jorum, both single files and content packages have extensive metadata records related to them. This metadata conforms to UK LOM Core format but is also accessible, both directly from the system and also remotely via OAI–PMH in Dublin Core (DC) and UK LOM Core.

Throughout this report, reference to metadata is made on the assumption the metadata is in UK LOM Core[4] format.

2.3 Keywords

Automated, content package, Jorum, Jorum Application Profile, learning object, metadata, repository, resource, resource discovery, retrieval, UKLOM CORE.

2.4 Introduction

Currently, Jorum uses intraLibrary (build v2.4), configured with the Jorum Application Profile (JAP) which is based on the UK LOM Core:

> *The UK LOM Core is essentially an application profile of the IEEE 1484.12.1 - 2002 Standard for Learning Object Metadata that has been optimised for use within the context of UK education*.[5]

Under the current set up within intraLibrary, metadata is recorded at the top level, rather than at organisation, resource or asset level. The only exception to this rule is the recording of the element 4.1 technical format which is undertaken by cataloguers to record multimedia assets within a content package. In this case, cataloguers do look at asset level to record technical format. For instance, if a ten page content package had 9 pages of HTML, and one page of HTML and a QuickTime file, the cataloguers should record both HTML and QuickTime within the technical format field. However, it is recorded at top level.

[3] IMS: http://www.imsglobal.org/content/packaging
[4] CETIS: http://metadata.cetis.ac.uk/specs/#UK_LOM_Core
[5] Jorum Application Profile: http://www.jorum.ac.uk/docs/word/japv1p0.doc

If a content package is imported with metadata at asset level, the system will not currently index this for searching, so in the present version of the Jorum repository software, searches cannot be made on this asset level metadata. The system will import and export this metadata. In future versions of the software, functionality will be developed to allow searching at asset level of metadata.

2.4.1 Overview of the Jorum Contributor and Jorum User Service

Jorum Service-in-Development is split into two services – a Contributor Service and a User Service.

The Contributor Service (see http://www.jorum.ac.uk/Contributors/index.htm for more detailed information) allows resources to be uploaded, subject to the Jorum Depositor agreement (http://www.jorum.ac.uk/docs/word/JORUM_Deposit_Licence_11_07_05.doc), catalogued and then subsequently shared for the benefit of the community. To date, Contributors come from both JISC funded projects (for example X4L Phase 1, and, imminently X4L Phase 2) as well as Further and Higher Educational Institutions who are keen to share their own materials. Contributors need to be named on a deposit licence in order to deposit resources within Jorum. As part of the contribution process, a Contributor needs to access the service with a dedicated Athens account which is created for them when they register for the Contributor Service. Account creation and Contributor set up is handled as part of the Jorum support process.

The User Service (see http://www.jorum.ac.uk/user/index.html for more detailed information) allows users to search, browse, preview and download resources. Once downloaded, subject to the Jorum User Sub Licence Agreement (http://www.jisc.ac.uk/index.cfm?name=coll_jorum_user_sub), users can reuse and repurpose materials they've found in Jorum. Once an institution has signed the licence, staff within it can self register using their Athens details to obtain access to the Jorum User service.

Both services are free and both are for staff working within FE or HE institutions within the UK.

The two services are discrete. A member of staff whose institution has signed the appropriate licences can be either a Contributor and a User, or else a Contributor or a User. There is also no direct relationship between contributing and download. Therefore, a user could conceivably download every single resource within Jorum without contributing a single one back in return.

2.4.2 Metadata creation

Resources which are contributed to Jorum can either have their metadata added directly within the system, using a series of editors (metadata, classification and rights), or can be imported with existing metadata already attached within the IMS Manifest file – created through an application like RELOAD or similar.

A point to note about actual metadata creation is that;

> *metadata creation is intimately connected to institutional processes and individual behaviors* [6]

This presents a problem for Jorum, which as a repository for the whole of the UK is taking resources from a variety of institutions. Therefore, there is likely to be a variety of processes and behaviours and an inconsistency in human generated metadata creation. To address this inconsistency, Jorum has cataloguers who implement Jorum cataloguing guidelines for metadata across the repository, but to assist this and also in the interests of transparency, Jorum should produce clear guidelines on metadata creation for Contributors. Currently this is done as a component of the training guides and support documentation, but a separate document explaining the rationale behind each metadata element completed by Contributors would be useful, and the creation of this document is therefore listed as a recommendation in this report.

Currently, we have a contribution model which requires very little metadata input from Contributors. A fuller description of elements which the Contributors need to complete is listed under section 3.1.

Informal feedback thus far elicited from a Contributors from both X4L phase 2, and institutions keen to share resources but not involved in any directly funded project has been that Contributors want to add more than the mandatory minimum elements required by Jorum. This is evidenced by support calls about metadata elements. It should be stressed that this feedback is to date informal, and further investigation is being undertaken as part of the Review of the Jorum Workflow Report due in July 2006. It may well be that these Contributors, who contributed at the beginning stages of the Jorum Contribution Service, can be regarded as those in the community who are most eager to embrace sharing. Thus it may well be that more formal feedback from a

[6] Crystal, A. & Greenberg, J., Usability of a metadata creation application for resource authors, 2005, Library and Information Science Research 27(2), http://ils.unc.edu/~acrystal/crystal_greenberg_2005_LISR.pdf

wider section of the learning and teaching community elicits a rather different response to manual metadata completion and, therefore, an evaluation of this activity may provide further requirements for automated metadata.

2.4.3 Jorum account set up and configuration

Once a Jorum deposit licence has been signed and processed, the depositor's or depositors' details are entered into the system and the Jorum administrator configures their Contributor accounts. Within the system, there are users, groups, workflows and templates. Each Contributor has logins to accounts which belong to one group. Typically, all members of a project team which are named in a deposit licence will belong to the same group. This process generates data that is used for automated metadata entries during the contribution process.

It should be noted that should users be accessing Jorum via shibboleth in the future that some data for their user profile could be automatically picked up, reducing the manual or self registration data entry requirements for the user or Jorum administrative team.

Group creation and configuration is handled manually by one of the Jorum support team.

Once the group and users have been created, and users, special templates and workflows assigned to the group, the Contributor is notified that their account is active and they can start the contribution process. Several steps within the contribution process are automated and these are outlined and examined below.

2.4.4 intraLibrary Metadata Templates

The system can use metadata templates (written in xml and configured by Jorum administrators within the system) to populate metadata elements automatically on upload. Future builds of the Jorum repository software will expand this functionality, allowing Contributors to create templates for themselves, which should significantly increase the amount of automatic metadata generation

Templates can either be applied at import, or, if no template is selected, they can be applied manually by the Contributor from within the metadata editor on a case by case basis as required.

Fig 2.4.3.3 shows a selected extract of a template; a complete template is shown in appendix 1 From the default template:

- <general>

- <title>
<langstring xml:lang="en">{ObjectFilename}</langstring>
</title>
- <catalogentry>
<catalog>{RepositoryId}</catalog>
- <entry>
<langstring xml:lang="x-none">{ObjectId}</langstring>
</entry>
</catalogentry>
<language>{UserPreferredLanguage}</language>
</general>
(edited)
- <rights>
- <copyrightandotherrestrictions>
- <source>
† <langstring xml:lang="x-none">LOMv1.0</langstring>
† </source>
- <value>
† <langstring xml:lang="x-none">Yes</langstring>
† </value>
† </copyrightandotherrestrictions>
- <description>
† <langstring xml:lang="en">See the JORUM Terms & Conditions at
http://www.jorum.ac.uk/licences/JORUM_RepurposeNoRepublishTandCv1p0.html</l
angstring>
† </description>
† </rights>

It is apparent from this extract and appendix 1 that entries are being automatically generated on application of the template and these are made in General (title, language, catalogue entry, date), Lifecycle, (role, entity, date), Metametadata, (Role, Entity, date Scheme,) Technical (format, size and location), and Rights (6.2 and 6.3 Description).

Templates can be customised for groups' particular needs. For instance, one X4L Phase 2 project, Learning 2 Learn, use a template to automatically apply the same classification to every resource which they upload. We use several templates within the system to allow groups to automatically generate the metadata which is best suited to their requirements.

This is clearly only a useful feature of the system when a project has outputs which all have at least one classification in common. However, where this is the case, informal feedback elicited from the Contributors indicates that this feature is useful in saving time.

3. Overview of automated metadata systems

As noted above, numerous projects and services are available which automate metadata to varying degrees and standards. Current resource constraints make it impracticable to examine them all within this report, but detailed examination of a wide range of automation tools would be useful. A good starting point would be an expansion of the resource lists available at http://dublincore.org/tools/

The reason for this research into automated metadata generation is given by Arms et al thus:

> *Simple algorithms plus immense computing power often outperform human intelligence*[7].

Nonetheless a useable, interoperable system requires more than just harnessing this power: two major problems in the developing landscape of automatic metadata generation tools are that:

> *Automatic techniques are rarely exploited. It seems that experimental research findings – specifically, the development of sophisticated automatic indexing algorithms focusing on resource content, semistructured metadata, and knowledge representation systems – have yet to be fully incorporated into the current automatic metadata generation applications* […]
>
> *Applications are developed in isolation, failing to incorporate previous as well as new advances, partly because of the absence of standards or recommended functionalities guiding the development of metadata generation applications. A standard set of functionalities could inform the development of more robust automatic metadata generation applications.* [8]

[7] Arms, W.Y. Automated Digital Libraries: How Effectively Can Computers Be Used for the Skilled Tasks of Professional Librarianship? *D-Lib Magazine*, *6* (7/9). http://www.dlib.org/dlib/july00/arms/07arms.html

[8] Greenberg, J., Spurgin, K. & Crystal, A. Functionalities for Automatic-Metadata Generation Applications: A Survey of Metadata Experts' Opinions. International Journal of Metadata, Semantics, and Ontologies, http://www.inderscience.com/storage/f121932106117458.pdf *Vol. 1, No. 1, 2006* 3

The second point is only partially valid in the context of Jorum, which is currently only using metadata which does exist within a standard, because while it is difficult to identify and assess software which claims to automate metadata within a UK LOM Core framework, there are indeed software applications which do so.

4 Workflow and automation processes and procedures within Jorum

Note. This explanation is based on the assumption that Contributors are using version 0.8 of our workflow. As the workflow is customisable and different groups require different workflows for their different contribution models, it is not the case that all groups of Contributors use this workflow. However, it is the most commonly used, and contains all the elements which make up workflows in the system.

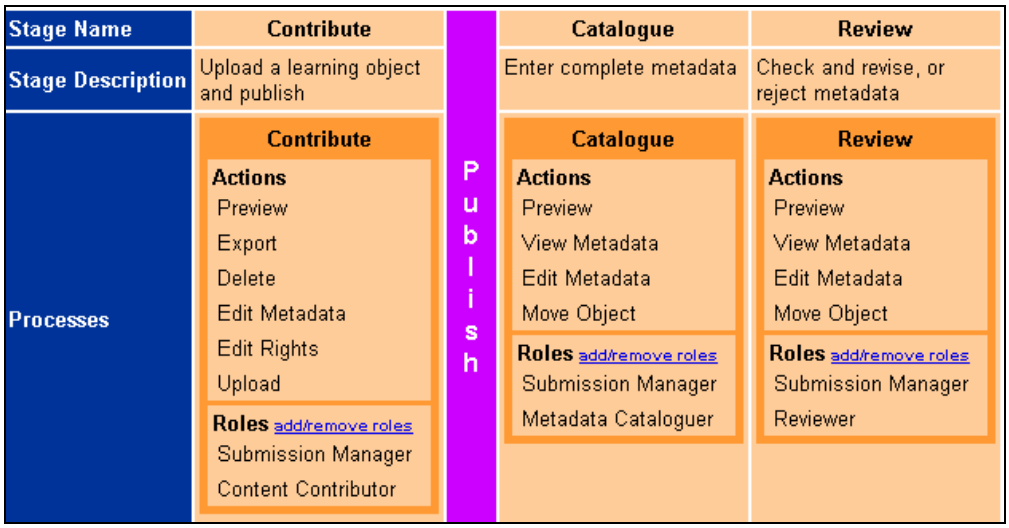

Fig 4.1 Screenshot of Jorum Workflow 0.8

As demonstrated in the screenshot above, and also in more depth on a Jorum Training video[9], the workflow model contains three stages – a contribution stage, a catalogue stage and a review stage. In workflow 0.8, the object is published between the contribution and cataloguing stage, but feedback from some Contributors is that they would like a review stage prior to contribution. However, the configuration of these workflows does not affect which elements are automated by intraLibrary so for the purposes of this report, automated elements will be grouped in the stages as they are typically encountered in following workflow 0.8.

[9] Jorum workflow video: http://www.jorum.ac.uk/contributors/chelp/index.php#Training

4.1 Metadata completed by Contributors

On upload, the following metadata elements need to be completed:

1.2 Title

1.4 Description

2.31 Role

2.3.2 Entity

3.2.1 Role

3.2.2 Entity

6.2 Subject to Copyright

6.3 Copyright and other restrictions

9.2 Classifications

As well as listing third party rights holders (expressed in ODRL)

Examination of the elements completed by a Contributor, their automation and the associated issues is made in Section 5.

Completion of these metadata elements represents the minimum requirement for Contributors; there is no maximum.

Once entries for these elements have been made, the Contributor can publish the resource and progress it through the workflow to the cataloguer stage. From the publish stage onward, the resource is available to other users via browsing taxonomies or via retrieval from metadata searches. Therefore, users can retrieve and download resources which are published but with an incomplete UK LOM Core Metadata record attached to them. Investigation into the frequency of this occurring will be undertaken in the forthcoming Workflow Review Report.

4.2 Cataloguers

At present, there are 12 Metadata cataloguers. They are subject specialist information professionals from the Resource Discovery Network (RDN)[10] who have experience of cataloguing both web (virtual) and learning objects. The cataloguing system has been designed to ensure that the complete JACS[11] and LearnDirect[12] subject classifications are covered but one cataloguer may have responsibility for more than one sector. For instance, a single cataloguer has responsibility

[10] RDN: http://www.rdn.ac.uk
[11] JACS Classification: http://www.hesa.ac.uk/jacs/completeclassification.htm
[12] LearnDirect Classification: http://www.learndirect-advice.co.uk/provider/standardsandclassifications/classpage/

for cataloguing philosophy, English Lit and Lang, and history, and medical resources.

Once a cataloguer reserves a resource, they complete the following

1.5 Keyword
3.2.1 Role
3.2.2 Entity
3.2.3 Date
4.1 Format
5.1 Interactivity Type
5.10 Description
5.11 Language
5.2 Learning Resource Type
5.3 Interactivity Level
5.5 Intended End User Role
5.6 Context
5.7 Typical Age Range
5.8 Difficulty
9 Classifications:

- Discipline using JACS, LearnDirect taxonomy
- Idea using Policy Themes taxonomy
- Educational Objective using Pedagogic Terms taxonomy
- Educational Level using Educational Level classifications

In addition, cataloguers check existing entries, to ensure compatibility with Jorum cataloguing guidelines[13] (for instance, removing subjective terms from the description).

As outlined above, the bulk of the cataloguing work is concerned with the educational properties of the resource. Cataloguers make judgements on these elements based on an assessment of the resource and also by comparing the resources to others which they have catalogued. This process is time intensive and subjective and automating many of the elements would prove difficult.

[13] Jorum Cataloguers' Handbook: http://www.jorum.ac.uk/docs/pdf/JORUM_Cataloguers_Handbook.pdf (2.5mb)

4.3 Reviewer

Once a cataloguer has finished completing entries, or checking existing entries made by both the original Contributor, they progress the resource through to the last stage of the workflow, review. Again, while a Reviewer is working on a resource, it is still available for users to retrieve and download. Currently a single Reviewer undertakes this task.

A Reviewer reviews 25% of catalogued resources. The Reviewer can then either progress the resource, so that it is no longer in a workflow, or reject it back to an appropriate stage of the workflow.

5. Current automation of metadata elements

Below is a list and explanation of elements which can be fully or partially automated by intraLibrary.

5. 1 Identifiers

The system uses identifiers which are created automatically by the system. One benefit of having the identifier elements automated is that we can hide them from the users to prevent confusion or Contributors overwriting the automatically generated entry.

5. 2 Title

The system can either pick up the title from the imported IMS Manifest file if the imported object is a content package, or, with a single file, the system automatically generates an entry for title based on the filename of the object. Using the default metadata template, this approach of using the filename as a title is implemented regardless of the type of resource uploaded.

Having intraLibrary automatically dropping the file extension – thus changing "An introduction to Ohms Law.doc" to "An introduction to Ohms Law" - could potentially be a development which aided Contributors. However, this would require that they named their resources with the title. Many objects have been uploaded into Jorum with titles like:

"LeedsUCM_Lecture1_Introducing Upland Catchment Management and the Leeds UCM Series"

Therefore, merely dropping a file extension from an uploaded file would not be universally successful in ensuring that automatically generated titles were accurate and as such, the Contributor and Cataloguer would still need to check this element. However, it should conceivably be an easy change to implement and would in many instances be useful.

Titles are checked by cataloguers and in the example title cited above, would be changed if the Contributor submitted the resource without changing the title.

5.3 Language

The system automatically assigns the default value of EN from ISO 639:1988 to the language elements 1.3 language, 3.4 language. A value for element 5.11 Language is not made.
The process of assigning an entry to elements 1.3 and 3.4 is part of the process of applying a template and this value could be changed as required to other languages such as FR.

5.4 Roles

For elements 2.31 Role and 3.2.1 role, the system automatically generates values. For element 3.2.1 this is "creator". This automation illustrates a problem with generic automation, as the entry would be incorrect where a record to be reviewed by the Metadata Reviewer is generated as "creator"; the entry should be validator.

5.5 Metadatascheme

intraLibrary automatically generates entries for the field 3.3 Metadatascheme –

- LOM 6.2
- IMS 1.2.1
- JORUM

Again, as these are elements which consistently have the same entry, automation of them is a useful feature which saves time in the creation of a LOM compliant record.

5.6 vCard

intraLibrary uses vCards[14] to record information on those whose work on a resource needs to be recorded, as well as those whose work on the metadata record for the resource also needs to be recorded. The vCard details of Contributors and Metadata Cataloguers: Name, Email and Organisation, are populated by the system from the account profile information. This account profile information is taken from the deposit licences which are typically completed by the project managers from the institutions contributing to Jorum.

However as the vCards are customisable, altering of the automatically generated metadata will mean an inconsistency between the metadata record and the information stored on the user within our Jorum Admin tool and the intraLibrary database which lists the affiliation, names and email addresses for each Contributor.

vCards are used both for 2.3.2 Entity and 3.2.2 Entity entries.

5.7 Dates

Dates are entered automatically by the system. The system assigns the date within the system as date entries for Lifecycle and Metametadata entries.

[14] Internet Mail Consortium: http://www.imc.org/pdi/

The date elements can be easily manually changed within the metadata editor.

When a cataloguer records his / her contribution to the metadata record, they need to manually add the date on which they make the contribution, so dates are not fully automated within the system.

5.8 Technical Format

The UK LOM Core recommends MIME types based on IANA registration (see RFC2048:1996) to record technical formats. intraLibrary makes an entry for technical format automatically. For a single file, intraLibrary makes an entry based on its examination of the resource at import. As multiple resources have to be submitted as a content package, it assigns it the MIME Type application/zip entry. This is another limitation of current automation within Jorum and as such, needs an extra process to ensure correct data extraction is made and recorded; as part of the cataloguing process, the assigned cataloguer should extend this entry to include multi-media entries. However, this is work the system could potentially do, and do better than a cataloguer - if a cataloguer is using Internet Explorer without required plug ins to view multi-media content, they will not know what the technical format of the plug in is, therefore be unable to record it.

5.9 Size

If the contributed resource is a single file, then the system determines its file size on import and automatically makes an entry for the element 4.2 Size. However, if the contributed resource is a content package, the system assigns it the compressed file size for this element. UKLOMCORE guidelines state that *"The size must refer to the uncompressed size of the resource".*

How much use this is, is debatable –the size indicated is actual size of the resource downloaded by the user which is a good guide for estimating download times. However, the uncompressed size may in some instances prove to be more useful for making judgements for whether the resource will pass restrictions to play in a VLE. For example, the default maximum object size setting for Moodle is 2MB.

5.10 Rights information

In addition, Contributors also need to complete third party rights information, which is stored as ODRL (http://odrl.net) separate to the UK LOM CORE metadata but is exported as part of an IMS

manifest file. Within the system, a URL listing terms and conditions of use (see the Jorum Terms & Conditions at http://www.jorum.ac.uk/licences/JORUM_RepurposeNoRepublishTandCv1p0.html) is automatically entered into element 6.3 Copyright and other restrictions. This URL entry currently over-writes any existing entries in the IMS Manifest file.

This approach has been taken because the UKLOMCORE does not make adequate provision for recording rights information. Element 6.3, Copyright and other restrictions, for instance, is a single element in which two separate entries, i) copyright and ii) other restrictions should be made. This is a poor data model. ODRL is a much richer, and more flexible expression language to record rights information. ORDL is also the preferable language should Jorum move to multiple licences.

The system uses profile information to automatically generate the user's institution as the default third party rights holder, but users can easily add others. Once the user has listed these rights holders, their entries are written to the IMS Manifest files thus;

```
- <o-ex:rights xmlns:o-ex="http://odrl.net/1.1/ODRL-EX" xmlns:o-
    dd="http://odrl.net/1.1/ODRL-DD" xmlns="http://www.intrallect.com/DRM">
- <o-ex:offer>
- <o-ex:context>
  <o-
    dd:reference>http://www.jorum.ac.uk/licences/JORUM_RepurposeNoRepublish
    TandCv1p0.html</o-dd:reference>
  <o-dd:name>JORUM RepurposeNoRepublish Licence</o-dd:name>
    </o-ex:context>
- <o-ex:party>
- <o-ex:context>
  <o-dd:name>University of Edinburgh</o-dd:name>
    </o-ex:context>
    </o-ex:party>
- <o-ex:permission id="JORUMmodify">
  <o-dd:display />
  <o-dd:print />
  <o-dd:play />
  <o-dd:execute />
  <o-dd:modify />
  <o-dd:excerpt />
  <o-dd:annotate />
  <o-dd:aggregate />
  <o-dd:move />
  <o-dd:duplicate />
  <o-dd:backup />
  <o-dd:install />
  <o-dd:delete />
  <o-dd:verify />
  <o-dd:restore />
  <o-dd:uninstall />
```

```
    <o-dd:save />
  - <o-ex:constraint>
    <o-dd:transferPerm o-dd:downstream="equal" o-ex:idref="JORUMmodify" />
      </o-ex:constraint>
      </o-ex:permission>
  - <o-ex:constraint>
  - <o-dd:purpose>
  - <o-ex:context>
    <o-dd:name>Educational Purposes Only</o-dd:name>
    <o-
      dd:reference>http://www.jorum.ac.uk/licences/JORUM_RepurposeNoRepublish
      TandCv1p0.html</o-dd:reference>
      </o-ex:context>
      </o-dd:purpose>
  - <o-dd:purpose>
  - <o-ex:context>
    <o-dd:name>No Commercial Use</o-dd:name>
    <o-
      dd:reference>http://www.jorum.ac.uk/licences/JORUM_RepurposeNoRepublish
      TandCv1p0.html</o-dd:reference>
      </o-ex:context>
      </o-dd:purpose>
      </o-ex:constraint>
  - <o-ex:requirement>
  - <o-dd:accept>
  - <o-ex:context>
    <o-dd:remark>User agrees to use this object under the terms and conditions
      stipulated in the JORUM licence found at
      http://www.jorum.ac.uk/licences/JORUM_RepurposeNoRepublishTandCv1p0.h
      tml</o-dd:remark>
    <o-dd:reference>http://www.jorum.ac.uk/odrl-
      licenses/RepurposeNoRepublish/1.0</o-dd:reference>
      </o-ex:context>
      </o-dd:accept>
      </o-ex:requirement>
      </o-ex:offer>
      </o-ex:rights>
```


5.11 Classifications

The classification entry that is mandatory need only be a general guide, as the purpose of it is to notify the cataloguers of the broad subject area of the resource. In documentation, Contributors are instructed that they only need make a single classification using either the JACS or LearnDirect subject taxonomies. By default, classification is made by Discipline.

All cataloguers are informed when any resource is published by the cataloguers, and cataloguers use the subject classification to decide whether the resource falls into their area of subject expertise. Thus, for instance, a resource with the classification JACS:/Medicine and Dentistry

would be ignored by all but those cataloguers who were specialists in Medicine and Dentistry. Classification is made by default by discipline although no specific entry is made by the default template. There is currently no way in the system to stop illogical classifications (eg, making a subject classification by educational level), but cataloguers will check and amend these entries.

5.12 Summary of automated elements

Table 5.1.2 displays the Metadata entries, by section, and shows the extent to which the elements within the sections are either automated, not automated, or whether an initial automation is made.

	Automated	Not automated	Initial automation
General	2	5	2
Lifecycle	2	2	1
Metametadata	6	0	1
Technical	2	7	1
Educational	0	11	0
Rights	2	0	0
Relation	0	5	0
Annotation	-	-	-
Classification	0	3	1

Table 5.1.2

Two things to note about the above data are that

a) it ignores container elements which just headings and contain no metadata
b) annotations are not included as they are not part of the contribution metadata model.

Partial automation is defined as where an entry is made, but would need to be checked – for instance assigning the document file name as a title.

6 Potential for change

As demonstrated in the previous section, intraLibrary automates much metadata. However, while there are areas where further automation could be executed - thereby reducing time taken to catalogue resources, and aiding retrieval and resource discovery - there is also a limit whereby any more work to automate further elements would prove counter–productive. The work required to allow the system to automatically make entries for the more subjective elements which the cataloguer currently completes or checks, such as 5.7 Typical Age Range, would be both financially expensive, and require extensive checking by cataloguers to ensure the entry made is logical. And having an experienced cataloguer check that an existing entry is valid is almost as time consuming as having them make the entry themselves.

Developing the functionality to make subjective entries is likely to be a prohibitly expensive return on investment, even if it were deemed possible by Intrallect Ltd.

Objective elements, such as size, date and format are much easier for a computer to automate than more subjective elements such as 5.8 difficulty.

However, there are options for further developments in automated metadata within Jorum and these are outlined below.

6.1 Mappings

It may be possible to map selected metadata elements to one another, with the effect that completion of one element results in the automatic completion of the elements to which it is mapped. Several mappings are outlined below.

6.1.1 Keyword / Description to classification

It should be possible for the system to take keywords and then make classification entries based upon them via means of a method such as an SQL look up and insert query.

Mapping an un-catalogued description to classification would present more effort, as a list of excluded words would have to be compiled for the system to cross check against prior to classification. For instance, mapping an example description of "This object contains information on gerunds and other grammar rules" would produce erroneous classifications for "information" in "Information technology" nodes, and "other"- as "other" appears as the lowest node in taxon paths in JACS ("Others in education" as the bottom node in the "Education" node for instance)

Mapping keywords would be more accurate, given Contributors complete metadata in line with Jorum standards, as keywords are subject specific. This would require revision to the contribution model, though, as element 1.5 keyword is not currently a mandatory field for Contributors. Offset against the increase in workload this amendment would entail for Contributors would be the benefit of not having to make a classification.

It should also be noted that, to date, the majority of records created by Contributors have keywords completed within the metadata record.

Further research needs to be undertaken to evaluate the usefulness of the classification systems as a discovery mechanism for retrieval as a whole. From this research, a decision made on whether the work needed to map keyword to classification is justified. Recent research from the University of Southampton discussed on the JISC Digital Repositories Mailing list[15] suggests that only 7.5% of downloaded resources during a sample period from Southampton Institutional Repository were the result of Browse by Subject Hierarchy command. However it's worth noting that Browse by Subject Hierarchy is only one application of classification entries and therefore detailed classification work may be required for other purposes such as exposure of resource metadata via OAI against a particular subject classification or a classification driven RSS feed.

In light of these considerations, a mapping of classifications to keyword has also been considered. Contributors have to make at least one subject classification using either JACS or LearnDirect taxonomies to notify the relevant cataloguer that a resource is available, and from this classification, keywords could be generated. But this is likely to derive less valuable metadata if the lower nodes of the JACS taxonomy are converted into keywords as the lowest node in each tree is a generic "others in" node.

[15] JISC Digital Repositories Mailing List: http://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind0603&L=JISC-REPOSITORIES&P=R3628&I=-3

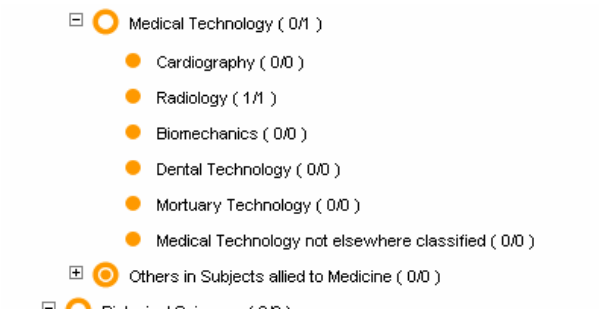

Fig 6.1.1.a) Screenshot of JACS taxonomy

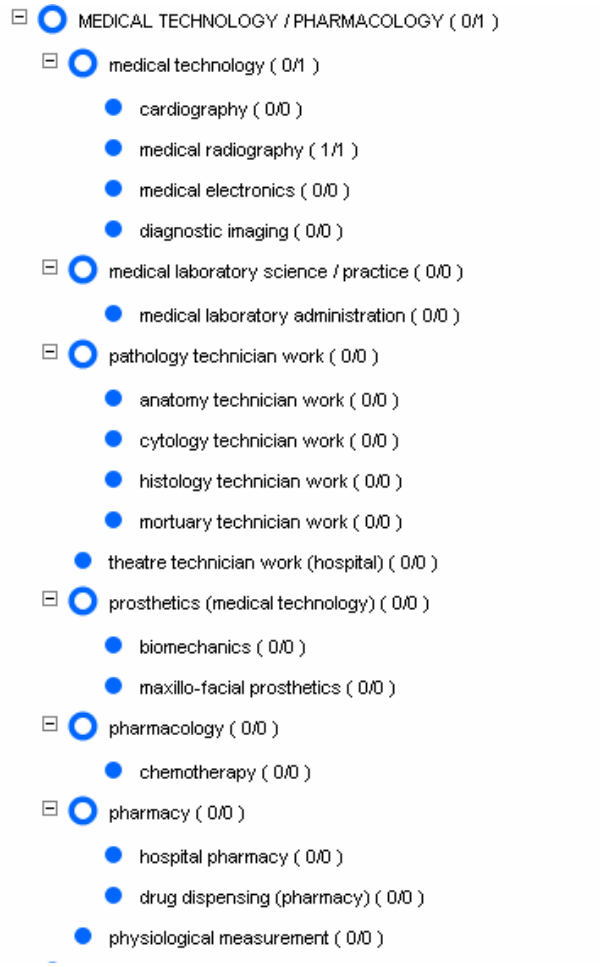

Fig 6.1.1.b) Screen shot of LearnDirect Taxonomy

As part of the mapping process between classification and keyword, duplicate terms would have to be removed, to prevent multiple entries of the same term as a keyword.

For instance, in the Metadata record for the resource taken from a Jorum resource and shown below, there would have to be a clean up script to prevent “English” and “Literature” appearing multiple times.

lom:string language="en">AREA STUDIES / CULTURAL STUDIES / LANGUAGES / LITERATURE</lom:string>
- <lom:string language="en">LITERATURE</lom:string>
<lom:string language="en">English literature</lom:string>
<lom:string language="en">English literature of specific periods</lom:string>
="en">AREA STUDIES / CULTURAL STUDIES / LANGUAGES / LITERATURE</lom:string>
<lom:string language="en">LITERATURE</lom:string>
<lom:string language="en">English literature</lom:string>
<lom:string language="en">English literature: specific authors</lom:string>
="en">Linguistics</lom:string>
- < <lom:string language="en">English studies</lom:string>

<lom:string language="en">English Literature</lom:string>
- < <lom:string language="en">English Literature by period</lom:string>

Fig 6.1.1.2 Edited taxonomy

Converting classifications into keywords would result in a large number of generated keywords. While the Jorum Cataloguers' Handbook recommends no more than 6 keywords are generated by cataloguers- based on ROLLMAP[16] recommendations- this is because of both the resource constraints which come with employing human cataloguers and also guidelines from other projects which suggests that after 6 keywords, relevance of subsequent keywords declines.
Keywords should be taken as phrases, so, in the example above, "English Literature" should be entered as a single term. This is to avoid confusion which would arise when phrases like "medical technology" are broken down into what would be overly vague keywords.

Mapping UKEL classifications to keywords would not be a viable use of resources as there is not a logical extraction of keywords which can be derived from the UKEL taxonomy.

6.1.2 Mapping of 1.8 Aggregation Level to 7 Relation

At present, none of the elements in Section 7 Relations of the UK LOM CORE are used by Jorum. Neither is 1.8 Aggregation Level, as the notes in the Jorum Application Profile explain:

> *As there is no consensus as to how to categorise aggregation level, use of this element is not recommended unless there is a specific rationale for exploring how aggregation can be described.*[17]

Although interestingly, Aggregation Level is one of the 22 mandatory elements within the Curriculum Tagging Tool.[18]

It may be possible in scoping the Jorum client tool, to allow using the relation section of the LOM to define relationships between resources created using the tool. For instance, several Contributors have uploaded both entire learning objects, and also disaggregated components of them into Jorum. It may be possible to include in the scope of the Jorum Client Tool the option to automatically link created resources to the previous resource created. If this was done using the Jorum Client Tool, it may be useful to allow linking between created resources via relation fields to point end users to other components of disaggregated resources. If this scenario was adopted, a

[16] RDN/LTSN LOM application profile (RLLOMAP): http://www.rdn.ac.uk/publications/rdn-ltsn/ap
[17] Jorum Application Profile: www.jorum.ac.uk/Contributors/chelp/japv1p0.html
[18] Curriculum Tagging Tool: http://www.curriculumonline.gov.uk/SupplierCentre/taggingtool.htm

further automation could be made for aggregation level of the aggregate content package. For instance, if Jorum formalised guidelines for use of 1.8 Aggregation Level that a resource with less than five components being rated as "1" within the element, 6-10 as "2", 11-15 as "3" and anything over 15 as "4", the Jorum client tool could automate the element based on it recording components of the resource.

Jorum would need to both specify the functionality required to make this mapping possible, as well as formalising use of aggregation level. Jorum is presently scoping requirements for the X4L Phase 2 Projects to use Relation elements to map entire content packages to separately uploaded disaggregated component assets of these packages.

Ferl[19] uses Relation elements within the LOM to allow users to search related resources.

6.1.3 Mapping Technical Format to Operating System / Browser

It should be possible to map technical format to Operating System (OS) or browser name, although it would need a group competent in the various versions of these to build on the work which is available on Wikipedia on browser compatibility.[20]

In addition to the question of whether these elements are used within Jorum as search terms, a further problem with these elements is their controlled vocabularies which are obsolete. Linux, Firefox and JAWS for instance, are not listed as browsers within the controlled vocabularies the UK LOM Core currently recommends for this element. Firefox had an 8.71% Market share in July05 [21] - the second highest market share behind Internet Explorer (86.56%) (NB, versions of browsers not stated). The OS controlled vocabularies list does not make provision for the rise of web and multimedia access via mobile devices.

Because Jorum can determine what OS and Browsers are used to access the Jorum website, it is worth considering attempting a combination of filtering of OAI Harvested resources and mapping against these elements as a search option to allow users to specify resources which matched their technical set up when they search via OAI. (For example, if a user visited the Jorum website on a browser which could not display Java, the OAI query returned could exclude any resources catalogued as containing Java components if the user selected this option).

The anticipated results would probably not highlight or exclude any notable resources - given that

[19] Ferl: http://ferl.becta.org.uk/index.cfm
[20] Wikipedia: http://en.wikipedia.org/wiki/Comparison_of_web_browsers
[21] Clickz Network: http://www.clickz.com/stats/sectors/traffic_patterns/article.php/3520661

most browser / OS combinations are Internet Explorer and Microsoft Operating Systems, but it would allow Jorum to showcase the functionality it has in presenting metadata via OAI – PMH.

6.1.4 Mapping Dual Source Vocabularies

At present, we ask cataloguers to make multiple entries from dual source vocabularies for element 5.2 Learning Resource Type and 5.6 Context. For 5.2 Learning Resource Type Jorum uses the UK LOM Core vocabulary twinned with the RDNLTSN vocabulary. For Element 5.6 Context, Jorum uses the UK LOM Core vocabulary twinned with the UKEC vocabulary. Each element should contain at least one entry from each vocabulary. These procedures are intended to allow metadata records within Jorum to exploit the richer expressions within the alternative vocabularies, thus providing more relevant information for users. At the same time, using UK LOM Core in addition to these vocabularies increases interoperability between systems and repositories which use only the default UK LOM Core vocabularies.

There are several common points of reference, especially between the UK LOM Core and UKEC vocabularies, and so mapping between these vocabularies should be straightforward. Dependent on the system limitations, mapping could be done at either import stage or during the metadata generation process within the system's Metadata Editor.

For instance, in 5.6 Context, if a Contributor of cataloguer selects "Higher Education" from the UKEC Controlled vocabulary, an entry should also automatically be made for the corresponding UK LOM Core value "Higher Education" as there is a direct mapping between these values.

6.1.4.1 Element 5.2 Learning Resource Type

The most obvious mapping here is between "Exam" in the LOM vocabulary and "ExaminationTest" in the RDNLTSN vocabulary

6.1.4.2 Element 5.6 Context

Suggested mappings between the vocabularies for element 5.6 Context are as follows:

UKEC:	Can be mapped to	LOMv1.0:
"nursery education"		"school"
" primary education"		"higher education"
"secondary education"		"training
"sixth form college"		"other"
" further education"		
" higher education"		
" vocational training"		
" continuous professional development"		
" community education"		

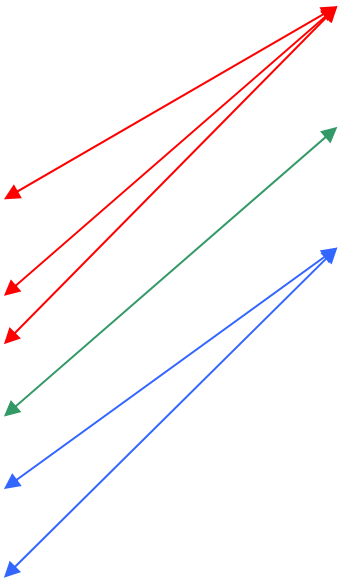

Table 6.1.4.2 – Mappings between UKEC and LOMv1.0 vocabularies

Key

Red – Maps Educational entries

Green - Maps training entries

Blue - Maps other entries relevant to Jorum

NB – no mapping of nursery, primary and secondary education are needed, as Jorum is for FE and HE use only, so it is highly unlikely that these elements will be used in Jorum.

6.2 Other potential areas of automatic metadata generation

Aside from the mappings outlined above, other areas of the contribution process present further opportunities for the automation of metadata to further streamline the process. These are outlined in the section below;

6.2.1 Automatic retrieval of keywords

If a Microsoft Office document is contributed, it should be possible to extract metadata from the metadata stored within the Word document.

Products are already available which claim to automatically generate the following metadata for Word (as well as similar metadata elements for StarOffice)

> *Application name, Title, Author, Subject, Keywords, Template, Comments, Last author, Revision number, Number of pages, Number of paragraphs, Number of lines, Number of words, Number of characters, Number of bytes, Number of notes, Number of slides, Manager, Company, Category, Security flags, Creation date, Last accessed date, Last print date, Edition time, get file type.*[22]

Generated entries as described above could provide a rich seam for exploitation, provided the extracted information is the sort of information which is searched on within the system. While extracting the number of lines or characters within a word document should prove easy, the usefulness of this metadata may be very low and presenting all of this metadata within a resources' record would likely prove an information overload.

Another factor which needs to be borne in mind regarding extraction from within Office documents is that the generation is dependent on content creators using these headings and tags within documents correctly.

Automatic extraction from other types of document, such as .pdf or HTML should also be possible, with the same caveats applying.

[22] Soft Experience: http://peccatte.karefil.com/software/Software.html.

6.2.2 Element 4.1 Technical Format

Examination of the manifest or document type could provide the opportunity to increase the numbers of entries made under 4.1 Technical Format.

An issue related to this would be the potential for multiple entries for things like spacer.gif. In discussion with members of CETIS, it was stated formats like spacer.gif should be recorded as, where included, they are valid components of a package. However Jorum uses this element to provide users with an overview of the technical components of the resource, and as such, a balance needs to be struck to ensure that listings are useful. Automatic generation of file types would need to be set up, therefore to prevent multiple entries for things like spacer.gifs; this would most likely confuse a user rather than provide them with an overview of the major components of a resource.

Therefore, when a package is ingested by the system, or at a pre-determined event stage within a workflow which produces minimum delay for the Contributor, it should make format entries after cross checking against a list of formats to ignore e.g. spacer.gif.

As an aside, this element is based on RFC2048:1996 which is flawed as it was last updated in 2001 – for instance the entry "…shockwave-flash" is a single entry listing two, different software programs.

6.2.3 Size

As mentioned in section 5.8, the system does not record the uncompressed size. While this is non UK LOM CORE compliant, the effort involved in changing this set up would not justify the returns. Merely listing the compressed file size means that file sizes across the repository are inconsistent – as different file compression programs, their settings, and the type of object being compressed all affect the size reduction when they are compressed. However, the purpose of this element in Jorum is intended to allow users to estimate the time it would take to download the object. In this instance, providing the compressed size would actually be of more use than the uncompressed size, given that it is the compressed folder which is downloaded and so there should be no work undertaken to adjust the automatically generated entry intraLibrary makes for file size.

7 Overview of other automation systems

7.1 Introduction

The conclusions drawn from both investigation into Jorum and other system matches that recorded by JISC's Joint Programme Meeting on: Digital Preservation & Asset Management and Digital Repositories, namely;

> *Much metadata could and should be automatically generated through machine processes. Descriptive and detailed subject and/or pedagogical metadata cannot be automated*[23]

As mentioned in section 2.1 Scope, because of resource limitations, this report does not claim to provide a comprehensive listing of other systems which partially or completely automate metadata generation. Jorum is aware of other work, such as projects using DSpace and the JISC ePrints UK project to explore added value services such as automated subject classification and name authority services[24], and the work RepoMMan is doing:

> *assess the feasibility of automated population of object metadata within an authenticated environment by (a) the extraction of descriptive metadata from simple digital objects and (b) drawing contextual metadata from existing institutional sources such as a portal profile or an enterprise directory via a Personal Metadata Profile and related profiles mapped to appropriate metadata schema. The feasibility of this will be assessed against the needs of potential users in the fields of research, learning and administration.*[25]

These projects will likely develop more detailed lists and evaluations as part of their reporting process, and use this evaluation to inform both practical applications in usage of automatic metadata as well as theoretical information process mappings.

7.2 Disclaimer

The research undertaken into the systems under section 7.2 is limited, and carried out at a high level without detailed knowledge of the systems. Any errors or omissions are the responsibility of the Jorum team, and any claims made about the software must be taken on the basis of limited testing undertaken by the Jorum team.

---

[23] Joint Programme Meeting : Digital Preservation & Asset Management and Digital Repositories Monday 10th October 2005 http://www.ukoln.ac.uk/repositories/digirep/index/Programme_meeting_10-10-2005_summary

[24] http://www.jisc.ac.uk/uploaded_documents/digital-repositories-review-2005.pdf, p5

[25] RepoMMan: Repository Metadata and Management, http://www.hull.ac.uk/esig/repomman/

The following systems; Automatic Metadata Generation Framework, Marvel, DC-Dot, IVIMEDS and Amazon were considered as potentially relevant to Jorum's needs.

7.2.1 Automatic Metadata Generation Framework:

http://ariadne.cs.kuleuven.be/amg/TryIt.jsp

This service allows users to upload files or specify a URL for automatic metadata generation in a variety of formats; Ariadne, DC or IEEE LOM.

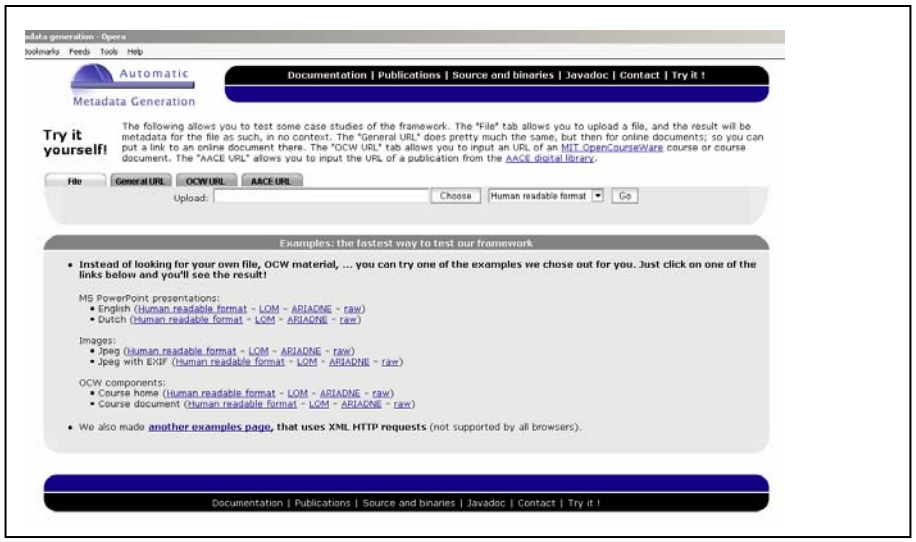

Fig 7.2.1.1 Screenshot of Automatic Metadata Generation Framework system

According to the website;

The idea behind our framework is that learning object metadata can be derived from two different types of sources. The first source is the learning object itself; the second is the context in which the learning object is used. Metadata derived from the object itself is obtained by content analysis, such as keyword extraction, language classification and so on. The contexts typically are learning (content) management systems (like Blackboard) or author institution information. A learning object context provides us with extra information about the learning object that we can use to

define the metadata.[26]

The project behind the framework system have done detailed work into automation of metadata and their outputs have informed their mapping of metadata extraction which is presented in the form of this diagram:

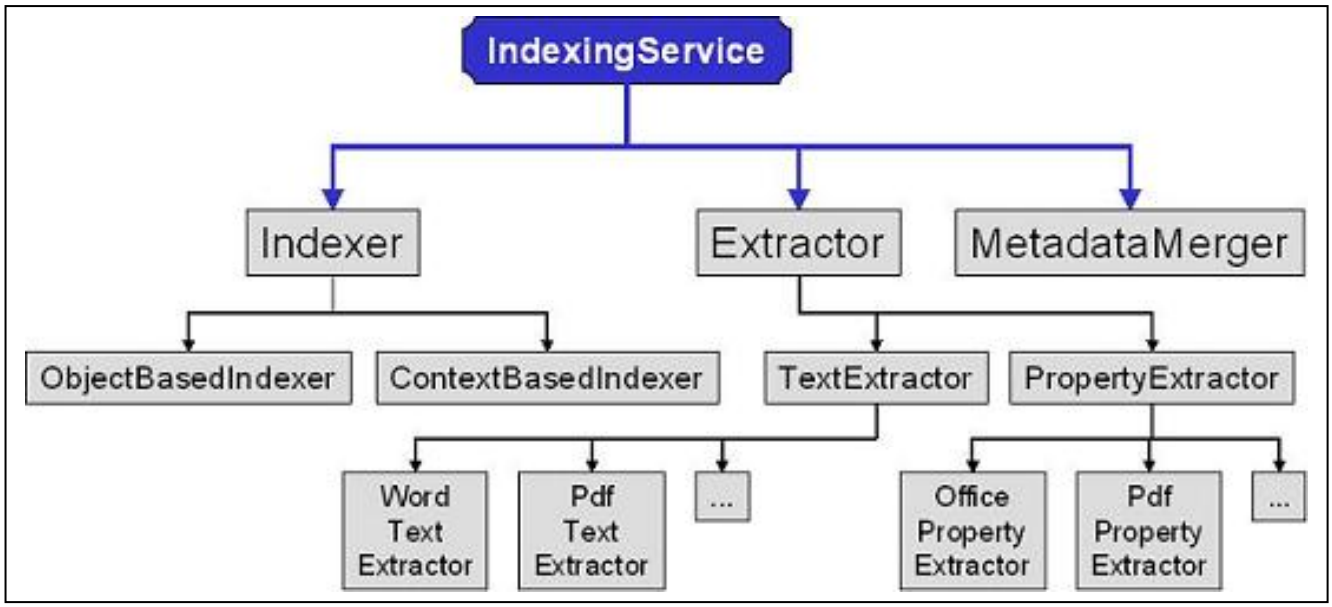

Fig 7.2.1.2 General Structure of the Automatic Indexing Framework

The project is most beneficial as a model for metadata extraction, and the resources on the website provide useful information for further exploration of automated metadata generation.

Testing the resource by uploading this report (saved as Metadata.doc) produced some interesting results.

<technical>
† <format>**application/msword**</format>
† <size>**1567744**</size>
- <requirement>
- <orComposite>
- <type>
† <source>**LOMv1.0**</source>

[26] Automated Metadata Generation Framework: http://ariadne.cs.kuleuven.be/amg/Intro.jsp 2006

† <value>**operating system**</value>
† </type>
- <name>
† <source>**LOMv1.0**</source>
† <value>**multi-os**</value>
† </name>
† <minimumVersion />
† </orComposite>
† </requirement>
† </technical>
- <educational>
- <interactivityType>
† <source>**LOMv1.0**</source>
† <value>**expositive**</value>
† </interactivityType>
- <interactivityType>
† <source>**LOMv1.0**</source>
† <value>**expositive**</value>
† </interactivityType>
- <learningResourceType>
† <source>**LOMv1.0**</source>
† <value>**narrative text**</value>
† </learningResourceType>
- <learningResourceType>
† <source>**LOMv1.0**</source>
† <value>**narrative text**</value>
† </learningResourceType>
- <intendedEndUserRole>
† <source>**LOMv1.0**</source>
† <value>**learner**</value>
† </intendedEndUserRole>

(This XML has been edited for brevity)

The linking between technical format and type is interesting, but on further investigation, it appears that this entry is generated even when an exe file is uploaded for assessment:

<technical>
<format>application/x-dosexec</format>
<size>1083289.6</size>
<requirement>
<orComposite>
<type>
<source>LOMv1.0</source>
<value>operating system</value>
</type>
<name>
<source>LOMv1.0</source>
<value>multi-os</value>
</name>
(Edited Code)

However, .exe files are not multi-OS files as they are specific to the windows family of Operating Systems, and will not, for example run on Linux.

Similarly, on upload and examination of an Apple OS specific file, the same relationship is generated by the software:

<format>application/x-stuffit</format>
<size>1357824</size>
<requirement>
<orComposite>
<type>
<source>LOMv1.0</source>
<value>operating system</value>
</type>

<name>

<source>LOMv1.0</source>

<value>multi-os</value>

(Edited code)

Further investigation also casts doubt on the reliability of the educational sector of the metadata. On upload of a blank MS Word document the following automatically generated metadata was returned;

<educational>

–

<interactivityType>

<source>LOMv1.0</source>

<value>expositive</value>

</interactivityType>

–

<interactivityType>

<source>LOMv1.0</source>

<value>expositive</value>

</interactivityType>

–

<learningResourceType>

<source>LOMv1.0</source>

<value>narrative text</value>

</learningResourceType>

–

<learningResourceType>

<source>LOMv1.0</source>

<value>narrative text</value>

</learningResourceType>

–

<intendedEndUserRole>

<source>LOMv1.0</source>

<value>learner</value>

</intendedEndUserRole>

</educational>

Whereas it's reasonable to assume that all MS Word documents will, by their nature, contain Narrative Text, it is not a valid assumption that they will always be intended for learners.

7.2.2 MARVEL

http://www.research.ibm.com/marvel/demos.html

MARVEL is an initiative from IBM which attempts to automatically create metadata from multimedia files. It attempts to derive semantic concepts from audio and visual content within multimedia files and these semantic concepts are then used as a retrieval mechanism.

Fig 7.2.2 – Screenshot of Marvel interface

IBM cite the rationale for MARVEL thus:

> *Manual labeling of multimedia content is extremely human resource and cost intensive and typically requires in excess of ten times greater time spent per unit time of video, e.g., one hour of video requires ten hours of human effort for complete annotation. Furthermore, manual labeling often results in incomplete and inconsistent annotations*[27]

Attempts to use the software were problematic and it is unclear whether this was the result of the browser / OS set up on the designated PC, network problems, poorly understood instructions, or a combination of these and other factors.

[27] MARVEL website http://www.research.ibm.com/marvel/details.html

The cost saving benefit of automatic metadata extraction from multimedia resources, coupled with the potential to classify, and, thus make retrievable, huge amounts of multimedia content means that this project should be monitored closely so that outputs from it which are made available can be incorporated into Jorum and other services.

7.2.3 DC-Dot

http://www.ukoln.ac.uk/cgi-bin/dcdot.pl?n=0&guesspublisher=yes

DC-Dot deals solely with automatic metadata creation for web resources;

> *This service will retrieve a Web page and automatically generate Dublin Core metadata, either HTML <meta> tags suitable for embedding in the <head>.</head> section of the page or RDF*[…] *Optional, context sensitive, help is available while editing*[28].

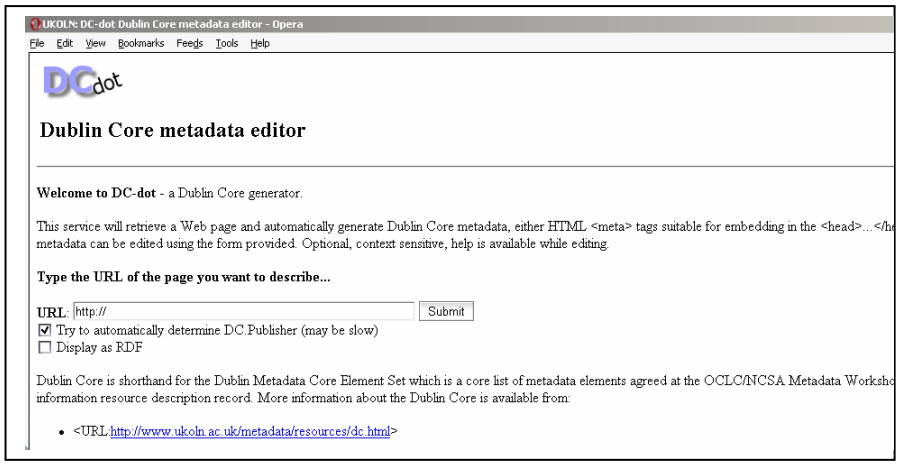

Fig 7.2.3.1 Screenshot of the DC-Dot Metadata Editor

To test the service, this document was saved as an html file by using MS Word 2003. This file was then examined and the conversion feature within the system was used to generate an XML file of IEEE LOM.

An interesting point to note was that the system generated 125 keywords, although it is highly likely that this test was unfair to the extent that most websites contain much less information on a single page than a single document does. Indeed, testing the software on the webpage on which it was hosted (http://www.ukoln.ac.uk/metadata/dcdot) produced a much more manageable set of metadata. Conversely, as it can only index a single page per submission, it may mean a high workload for a contributor who wanted to catalogue an entire web site. The Jorum cataloguers' guideline is not to create more than 6 Keywords, after which relevance declines. Keywords such as "information" and "format", generated on submission of this document as a web page, would increase search retrieval within Jorum, but would greatly diminish relevance. This shows that while a computer can generate metadata more quickly than a human, a human is better placed to concept check the generation.

[28]UKOLN: http://www.ukoln.ac.uk/metadata/resources/dc/

The software interface includes the functionality for the user to edit and refine metadata once it has been automatically created, and, while this is useful, it does reinforce a previous lesson which is that while computers can generate certain types of metadata quickly, they are not yet capable of consistently generating useful, complete, metadata records without human editing; it would be an interesting exercise to evaluate the time benefits of a workflow consisting of automatic creation and human review versus a workflow which only included human metadata creation.

7.2.4 IVIMEDS

http://www.ivimeds.org

IVIMEDS is included here because automated metadata need not only be concerned with record creation, but also as a tool to enable users to make judgements on the usefulness of retrieved resources, and here IVIMEDS uses automatically generated metadata effectively. Great efforts are now being made in collating secondary metadata, to help inform users' opinions on the resources they find via browsing or searching, and IVIMEDS has secondary metadata based on developments on secondary metadata from flickr and Amazon[29]

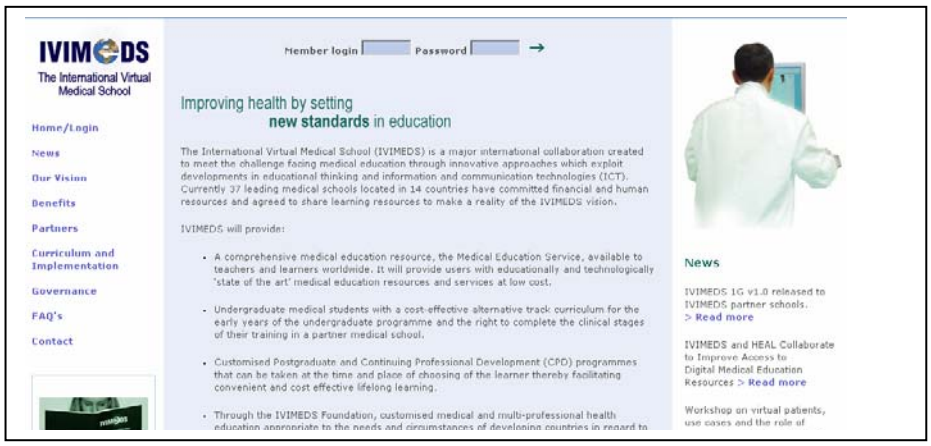

Fig 7.2.4.1 Screenshot of IVIMEDS

IVIMEDS defines itself thus:

> *The International Virtual Medical School (IVIMEDS) is a major international collaboration created to meet the challenge facing medical education through innovative approaches which exploit developments in educational thinking and information and communication technologies (ICT). Currently 37 leading medical schools located in 14 countries have committed financial and human resources and agreed to share learning resources to make a reality of the IVIMEDS vision.*[30]

IVIMEDS displays comments and reviews from users on resources. It also displays automatically gathered statistics, such as how often each resource has been downloaded. This allows users to see more popular resources, which may create a self perpetuating process, but it nevertheless is a clear way for users to make an initial assessment of the usefulness of the resource.

[29] Presentation by David Davies on IVIMEDS at CETIS Metadata SIG, Open University, 09/05 http://metadata.cetis.ac.uk/sig_meetings/OUSept2005/daviddavies.makingmetadatawork.ppt PowerPoint 9.5mb

[30] IVIMEDS: http://www.ivimeds.org/home.html

One thing to note is that the usefulness of the number of times the object has been downloaded is probably diminishing as transport speeds via Broadband or SuperJANET increase – with decreased cost and effort of download comes an increase in willingness to download resources.

A more useful statistic would be "Number of times this resource has been run in a VLE", but obtaining this statistic would obviously be effort intensive, even if obstacles regarding privacy and data protection were overcome. Indeed, a specification for secondary metadata for Learning Objects, similar to data defined by the COUNTER project[31] would be very useful as there is at present a gap in metadata provision – the UK LOM Core aids only resource discovery, but Learning Objects are intended to be reused as well as used, and at present there is no standard for capturing reuse effectiveness.

[31]Counter Project: http://www.projectcounter.org

7.2.5 Amazon

http://www.amazon.co.uk

Although it is not a Learning Object Repository, Amazon is worth investigating because of the innovative way metadata is used, and the fact that these innovations are now being picked up and used by repositories (see section 7.2.4 on IVIMEDS).

Amazon has many useful features on automated secondary metadata- an area of great interest for providing users with alternative routes to resource discovery- which would be of benefit to Jorum.

Presently, Jorum uses annotations and star ratings features to allow users to comment on retrieved resources. This brings benefits, both to the contributor who can see feedback on their resource, and also to other users, who can see what others in the community think of the resources, as well as descriptions on how they've been reused or modified.

Amazon takes this approach a stage further, by allowing users to create accounts. This allows interested parties to see all reviews by one reviewer. Amazon automatically generates a page based on rated reviews to power its Top Reviewers feature.

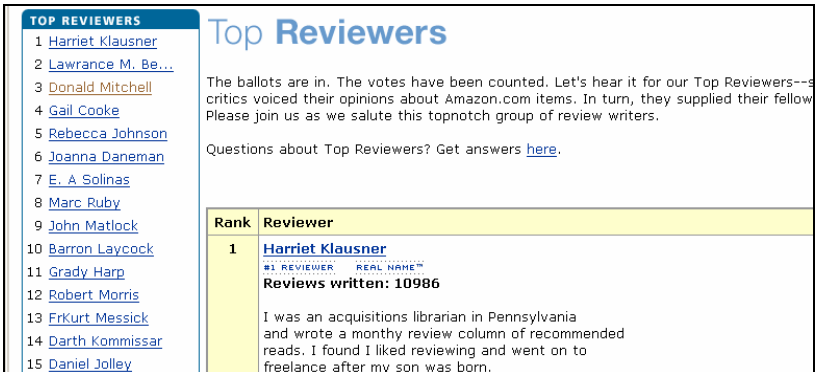

Fig 7.2.5.1 – Amazon's Reviewer page

Fig 7.2.5.2 – Amazon has the functionality to see other reviews by the same reviewer

Amazon also utilises Metadata to automatically link resources together. This takes the form of their "Customers who bought books by X also book books by Y", as shown in Fig 7.2.5.3;

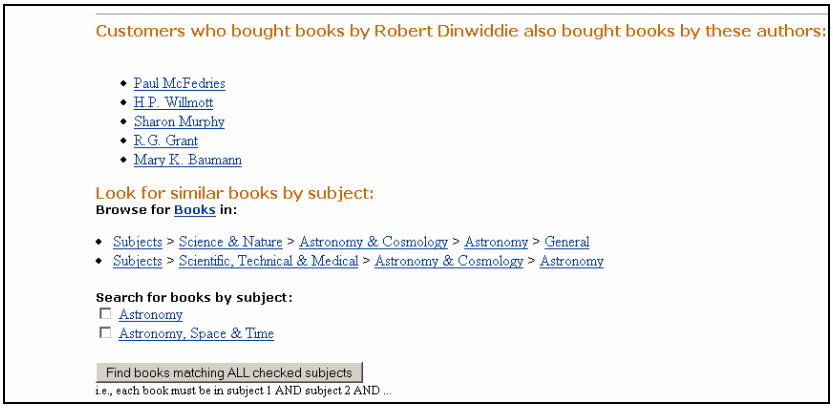

Fig 7.2.5.3 – Metadata generated by linking purchased resources

This feature is an interesting way to link resources together and is an informative method of resource discovery. The huge spend Amazon has on R&D ($132m in the fourth quarter of 2005)[32], and the highly competitive commercial environment in which it operates means that there are likely to be benefits of further work into automated metadata by looking at Amazon for use cases of secondary metadata.

8 Recommendations

There are several developments which could potentially be of benefit to Jorum. However, as with everything, there is a caveat; as this report was written independently of Intrallect, they have had no input into the recommendations which arise from this report. Their initial opinion should be sought prior to work being undertaken to evolve these recommendations into functional requirement specifications.

Similarly, others involved in the emerging arena of automated metadata and its associated systems would be better placed to expand on the initial overview of these systems. Were Jorum to keep an "automated metadata watch" on this work, it is likely that this would uncover developments in systems and procedures which could be advantageously adopted by Jorum.

Recommendations are outlined below and fall into two main areas. These are 7.1 System Requirements, which outline improvements which could be undertaken to improve how Jorum uses intraLibrary, and 7.2 Procedures which outline areas where Jorum could improve its procedures to

[32] Business Week: http://www.businessweek.com/technology/content/feb2006/tc20060203_222987.htm

enhance the Jorum workflow via automation.

A recommendation not to do something is as valid as a call for change. Thus two recommendations are to preserve the current practice within Jorum, even though this initially appears to limit the extent and accuracy of automated recommendation.

8.1 System recommendations

- Having a process which ignored file extensions such as .doc in the titles may aid Contributors by increasing the likelihood of file titles mapping to resource titles, and thus the feasibility of this should be examined with Intrallect.
- No work should be undertaken to make element 4.2 size UK LOM Core compliant, as the resources required to do so outweigh the benefits. As it is, it can be argued that having the zipped size gives the user a better indication of download times anyway.
- Work should be done to allow automatic generation of increased and more accurate instances of technical format than are currently generated.
- Automated generation of metadata based on the relationship between technical format and 4.4.1.2 Name is recommended. This would be subject to analysis by Intrallect software engineers that this generation would not be overly resource intensive for the software. Element 4.4.1.2 Name is not currently used but it would be a useful tool to be able to browse on via OAI.
- Mapping UKEC and RDNLTSN values to UK LOM Core vocabulary entries, as highlighted in section 6.1.4 would be a time saving device for either the Cataloguer or Contributor, therefore it is recommended that this should be undertaken.
- Automation and non-editability of vCard entries would help the synchronisation process while maintaining consistency of the users' vCard details in Jorum and the information they submitted during the signing of the Jorum Deposit Licence. Therefore this is recommended.
- Mapping keyword to classification should be undertaken. Research that shows that classification is not a conduit to resource discovery is not comprehensive enough for us to discount this mapping.
- Research between Jorum systems analysts and programmers is recommended to further consider the viability of including automatic generation of Relation elements within the Jorum Client Tool.

8.2 Procedural recommendations

- It is recommended that production of a brief guide for Contributors is undertaken to explain the rationale of metadata elements used. This will address research which indicates

metadata creation is highly institutional dependent

- While automatic extraction of metadata from MS Office, PDF or HTML documents is desirable, a balance needs to be struck to ensure that the user is not presented with an overly large metadata record which contains spurious metadata (such as the number of characters in a MS Word document). This ties in with the next point, namely;
- Further work, as part of the Jorum workflow evaluation, should be made on what elements users are retrieving resources on, as well as examining in more detail the proportion of submitted resources into Jorum which have more than the minimum required metadata elements completed. This will allow us to assess both the usefulness of current cataloguing guidelines, as well as establishing how prepared the community is to create metadata themselves.
- And any further work which Jorum undertakes should strive to address these concerns with usability and the balance between asking the Contributor to complete metadata and maintain a contribution process which is not overly burdensome, especially as the Contributors to Jorum will cover the spectrum in terms of technical expertise.
- The current workflow model which has a process of metadata creation followed by a process of cataloguing and then review should remain in place, and cataloguers should at least concept check selected entries which the system makes. Whereas it is not necessary or practical to require cataloguers to check vCard details or identifiers, it should still be a required step for them to check elements which aid resource discovery, even if the creation of these entries is automatic. This would not greatly impact on the cataloguers' workloads.
- As automatic metadata generation is an emerging technology, a watch should be kept on trends and developments from which Jorum could benefit.
- Consideration should be given for further work to investigate how the automation of metadata relates to areas other then resources discovery such as advocacy and preservation.

9 Bibliography

9.1 Mailing lists

JISC Digital Repositories List, http://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind0603&L=JISC-REPOSITORIES&P=R3628&I=-3

Jorum / RDN Cataloguers' Discussion List, http://www.jiscmail.ac.uk/lists/RDNJORUM.html

9.2 Journal and web articles

Joint Programme Meeting: Digital Preservation & Asset Management and Digital Repositories Monday 10th October 2005
http://www.ukoln.ac.uk/repositories/digirep/index/Programme_meeting_10-10-2005_summary

Arms, C., Dushay, N., Foulonneau, M., Hagedorn, K., Hutt, A., Hillmann, D., Lally, A., Landis, B., Redding, C., Riley, J., Shreeves, S., Ward, J. and Warner, S., Best Practices for Shareable Metadata – DRAFT, August 2005,
http://comm.nsdl.org/download.php/653/ShareableMetadataBestPractices.doc

Arms, W.Y. Automated Digital Libraries: How Effectively Can Computers Be Used for the Skilled Tasks of Professional Librarianship? August 2000, D-Lib Magazine, 6 (7/9)
http://www.dlib.org/dlib/july00/arms/07arms.html

Arms, W.Y., Dushay, N., Fulker, D.W. and Lagoze, C., A Case Study in Metadata Harvesting: the NSDL. October 2002, Library Hi Tech, 21 (2). http://www.cs.cornell.edu/lagoze/papers/Arms-et-al-LibraryHiTech.pdf

Arms, W.Y., Hillmann, D., Lagoze, C., Krafft, D.,Marisa, R., Saylor, J., Terrizzi, C. and Van de Sompel, H., A Spectrum of Interoperability: The Site for Science Prototype for the NSDL. January 2002, *D-Lib Magazine*, *8* (1). http://www.dlib.org/dlib/january02/arms/01arms.html

Berners-Lee, T, Metadata Architecture, January 1997,
http://www.w3.org/DesignIssues/Metadata.html

Crystal, A. & Greenberg, J., Usability of a metadata creation application for resource authors, 2005, Library and Information Science Research 27(2), http://ils.unc.edu/~acrystal/crystal_greenberg_2005_LISR.pdf

Davies, D, Making Metadata Work, September 2005, CETIS Joint Metadata and Pedagogy SIG, Open University http://metadata.cetis.ac.uk/sig_meetings/OUSept2005/daviddavies.makingmetadatawork.ppt (9mb)

Greenberg, J., Spurgin, K. & Crystal, A. Functionalities for Automatic-Metadata Generation Applications: A Survey of Metadata Experts' Opinions. 2006, International Journal of Metadata, Semantics, and Ontologies. *Vol. 1*, *No. 1* http://www.inderscience.com/storage/f121932106117458.pdf

Greenberg, J, Metadata Generation: Processes, People and Tools, December/January 2003 BULLETIN of the American Society for Information Science and Technology Vol. 29, No. 2 http://www.asis.org/Bulletin/Dec-02/greenberg.html

Greenberg, J. Metadata extraction and harvesting: A comparison of two automatic metadata generation applications. 2005, *Journal of Internet Cataloging* , *6*(4), http://ils.unc.edu/mrc/automatic.pdf

Greenberg, J., Crystal, A., Robertson, W. D., and Leadem, E. Iterative design of Metadata Creation Tools for Resource Authors. 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice – Metadata Research and Applications. DC-2003: Proceedings of the International DCMI Conference and Workshop. http://www.siderean.com/dc2003/202_Paper82-color-NEW.pdf

Guy, M & Tonkin E, Folksonomies; Tidying up Tags, January 2006,D-Lib Magazine, 12(1) http://www.dlib.org/dlib/january06/guy/01guy.html

Jenkins C, Jackson M, Burden M, Wallis J. Automatic RDF Metadata Generation for Resource Discovery, Date unknown, University of Wolverhampton http://www8.org/w8-papers/2c-search-discover/automatic/automatic.html

Lagoze, C., Krafft, D.B., Payette, S. and Jesuroga, S., What Is a Digital Library Anyway? Beyond Search and Access in the NSDL. November 2005, *D-Lib Magazine*, *11* (11). http://www.dlib.org/dlib/november05/lagoze/11lagoze.html

Paynter, G. Developing Practical Automatic Metadata Assignment and Evaluation Tools for Internet Resources. Date unknown, The INFOMINE Project, University of California
http://ivia.ucr.edu/projects/publications/Paynter-2005-JCDL-Metadata-Assignment.pdf

9.3 Websites

Amazon: http://www.amazon.co.uk

Automatic Metadata Generation Framework:
http://ariadne.cs.kuleuven.be/amg
http://ariadne.cs.kuleuven.be/amg/TryIt.jsp
http://ariadne.cs.kuleuven.be/amg/Intro.jsp

Business Week:
http://www.businessweek.com/technology/content/feb2006/tc20060203_222987.htm

CETIS: http://www.cetis.ac.uk/list.html?SubjectContext=metadata

Clickz Network: http://www.clickz.com/stats/sectors/traffic_patterns/article.php/3520661

COUNTER: http://www.projectcounter.org/

Curriculum Online: http://www.curriculumonline.gov.uk/SupplierCentre/taggingtool.htm

DC - Dot: http://www.ukoln.ac.uk/metadata/resources/dc/

Dublin Core Tools: http://dublincore.org/tools/

Ferl: http://ferl.becta.org.uk/index.cfm

Flickr: http://www.flickr.com

IMS Global: http://www.imsglobal.org/content/packaging

Intrallect: http://www.intrallect.com

Internet Mail Consortium: http://www.imc.org/pdi

IVIMEDS: http://www.ivimeds.org/home.html

Jacob Neilsen: http://www.useit.com

JACS Classification: http://www.hesa.ac.uk/jacs/completeclassification.htm

JISC Digital Repositories review: http://www.jisc.ac.uk/uploaded_documents/digital-repositories-review-2005.pdf

LearnDirect Classification: http://www.learndirect-advice.co.uk/provider/standardsandclassifications/classpage

MARVEL: http://www.research.ibm.com/marvel/details.html

Multi OS Browser Test Website: http://www.aadmm.de/en/index.htm

ODRL Initiative: http://odrl.net

RDN/LTSN LOM application profile (RLLOMAP) http://www.rdn.ac.uk/publications/rdn-ltsn/ap

Resource Discovery Network: http://www.rdn.ac.uk

RepoMMAn: http://www.hull.ac.uk/esig/repomman/documents/index.html

Soft Experience: http://peccatte.karefil.com/software/Software.html.

Wikipedia: http://en.wikipedia.org/wiki/Comparison_of_web_browsers

All links valid March 2006

Appendix

1. **Exported XML file of UK LOM Core Metadata applied by intraLibrary on file upload completion.**

```
<?xml version="1.0" encoding="UTF-8"?>
<lom:lom xmlns:lom="http://ltsc.ieee.org/xsd/LOM"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://ltsc.ieee.org/xsd/LOM
http://ltsc.ieee.org/xsd/lomv1.0/lomLoose.xsd">
  <!--Generated by transforming IMS 1.2.1
metadata to IEEE LOM Metadata-->
  <lom:general>
    <!--General Section-->
    <lom:identifier>
      <lom:catalog>intraLibrary-OAI</lom:catalog>
      <lom:entry>oai:uk.ac.ed.ucs.bodach.jorum:897</lom:entry>
    </lom:identifier>
    <lom:title>
      <lom:string language="en">mdfinaldraftpriorjc.doc</lom:string>
    </lom:title>
    <lom:language>en</lom:language>
  </lom:general>
  <lom:lifeCycle>
    <!--Lifecycle Section-->
    <lom:contribute>
      <lom:role>
        <lom:source>LOMv1.0</lom:source>
        <lom:value>author</lom:value>
      </lom:role>
      <lom:entity>BEGIN:vcard
FN:Kenny Baird
ORG:University of Edinburgh
EMAIL:kenny.baird@ed.ac.uk
END:vcard</lom:entity>
      <lom:date>
        <lom:dateTime>2006-03-27T20:19:09.6Z</lom:dateTime>
      </lom:date>
```

```
    </lom:contribute>
  </lom:lifeCycle>
  <lom:metaMetadata>
    <!--Metametadata Section-->
    <lom:identifier>
      <lom:catalog>intraLibrary-OAI</lom:catalog>
      <lom:entry>oai:uk.ac.ed.ucs.bodach.jorum:897</lom:entry>
    </lom:identifier>
    <lom:contribute>
      <lom:role>
        <lom:source>LOMv1.0</lom:source>
        <lom:value>creator</lom:value>
      </lom:role>
      <lom:entity>BEGIN:vcard
FN:Kenny Baird
ORG:University of Edinburgh
EMAIL:kenny.baird@ed.ac.uk
END:vcard</lom:entity>
      <lom:date>
        <lom:dateTime>2006-03-27T20:19:09.8Z</lom:dateTime>
      </lom:date>
    </lom:contribute>
    <lom:metadataSchema>IEEE LOM 1.0</lom:metadataSchema>
    <lom:language>en</lom:language>
  </lom:metaMetadata>
  <lom:technical>
    <!--Technical Section-->
    <lom:format>application/msword</lom:format>
    <lom:size>1751040</lom:size>
    <lom:location>http://repository.jorum.ac.uk/intraLibrary/intraLibrary?command=open-
preview&amp;amp;learning_object_key=i1443n53826t</lom:location>
  </lom:technical>
  <lom:rights>
    <!--Rights Section-->
    <lom:copyrightAndOtherRestrictions>
      <lom:source>LOMv1.0</lom:source>
      <lom:value>yes</lom:value>
```

```
    </lom:copyrightAndOtherRestrictions>
    <lom:description>
      <lom:string language="en">See the JORUM Terms &amp;amp; Conditions at
http://www.jorum.ac.uk/licences/JORUM_RepurposeNoRepublishTandCv1p0.html</lom:string>
    </lom:description>
  </lom:rights>
  <lom:classification>
    <!--Classification Section-->
    <lom:purpose>
      <lom:source>LOMv1.0</lom:source>
      <lom:value>discipline</lom:value>
    </lom:purpose>
  </lom:classification>
</lom:lom>
```

NB; ODRL not applied until Contributor selects radio button for Jorum Deposit Licence

Comments and Corrections - Updated 02/08/06

Received 28/07/06: Michael Meire, developer on the "Automatic Metadata Generation Framework" Project. Katholieke Universiteit Leuven, Dept. Computer Science.

Corrections:
Several url links were broken and have been changed in the report:
http://memling.cs.kuleuven.ac.be/amg/tryIt.php changed to
http://ariadne.cs.kuleuven.be/amg/TryIt.jsp
http://memling.cs.kuleuven.ac.be/amg/documentation.php changed to
http://ariadne.cs.kuleuven.be/amg/Intro.jsp

The general access point to information on the "Automatic Metadata Generation Framework" Project was included in the bibliography (see p.51): http://ariadne.cs.kuleuven.be/amg

Comments – quoting Michael Meire directly:
p.12 of the report mentions 2 major problems in the area of automatic metadata generation. The second one is most interesting to us. It mentions the lack of standards or recommended functionalities for automatic metadata generation. This was one of our concerns too, and that's why we developed the "Simple Automatic Metadata Generation Interface" (http://ariadne.cs.kuleuven.be/amg/DesignSamgi.jsp). This is for sure not a standard, but rather our suggestion on what operations should be offered by a system that offers (automatic) metadata generation. We are by the way very much open to input/comments/suggestions on this.

p.38 mentions the fact that not all MS Word documents contain Narrative Text. Indeed, this is a default value that we assume within our system. More generally, we introduced the concept of "Conflict Handling" and "Confidence Values" within our framework, specifically to deal with the issue that certain values (default or non-default) are not always correct. The global idea is that, when suggesting a metadata-value, also a "degree of certainty" is given. For example we could have something like "We think it is Narrative Text, but we are only 20% sure".
The end-application, using our framework, could then for example decide not to use the values that have a degree of certainty below 50%.
More information on this can be found on our
http://ariadne.cs.kuleuven.ac.be/amg/DesignSamgi.jsp, and in the papers that are mentioned on
http://ariadne.cs.kuleuven.be/amg/Publications.jsp

Finally, I would like to add 2 *[sic]* concluding general comments about our work:

- We have been (and still are) giving much attention on re-engineering the first version of our system, and adding new features. The most recent version can always be tested using the Try-it section on our webpages.
- All code is put on Sourceforge. The structure of the code has been revisited to make it more easy to check out, compile and test the code yourself.
- If you use the Try-it section on our webpages, the amount and quality of generated metadata still is not always perfect at all. However this is in line with the philosophy behind our work: if you try to generate metadata for an object, looking only at that object in an isolated way, the generated metadata is limited. If you however also look at the context within which an object is used (eg a Learning Management System), and tailor the metadata generation process to this context, then the amount and quality can increase drastically.

  Our work has therefore focused very much on providing a general architecture where you can add (or plug-in) new pieces, for dealing with new contexts.

Contact: Any questions or comments to Michael Meire should be sent to: ariadne@cs.kuleuven.be

---