



RepoMMan Project

D-D13

Automatic generation of object metadata

Richard Green and Chris Awre

August 2007

Version 1.1 October 2007



The RepoMMan Project

Project Director:	Ian Dolphin, Head of e-Strategy, University of Hull (i.dolphin@hull.ac.uk)
Project Manager:	Richard Green (r.green@hull.ac.uk)
Technical Lead:	Robert Sherratt (r.sherratt@hull.ac.uk)
Repository Domain Specialist:	Chris Awre (c.awre@hull.ac.uk)

The Repository Metadata and Management Project (RepoMMan) at the University of Hull is funded by the JISC Digital Repositories Programme. The project is being carried out by the University's e-Services Integration Group (e-SIG) within Academic Services.

Authors' note: This document incorporates parts of our document 'R-D11 Report on metadata needs for the University of Hull Digital Repository'.

1 Introduction

The RepoMMan project is researching and developing a tool to facilitate the use of a Fedora-based Institutional Repository for private development work. At the University of Hull we see a repository not just as a web space in which to deposit finished outputs of one sort or another, rather a working tool which can support users in the development of such objects from conception to completion and possible publication.

It is our belief that the success of an institutional repository, as seen from the outside world, is heavily dependent upon the quality of metadata associated with the digital objects that it holds. Effective metadata underpins effective discovery, location and potential re-use of digital objects. Metadata can take various forms, including descriptive, administrative and preservation metadata. Manual generation of this metadata has revealed difficulties in ensuring that a full and accurate metadata record is stored with the object, limiting future use. The creation of quality metadata can be both time consuming and laborious.

The RepoMMan project investigated the extent to which automatic population of object metadata might be feasible and this report is one of the outcomes from that work.

2 Overview

We consider that there are three types of metadata that need to be addressed in setting up a repository at the University of Hull:

- descriptive metadata which deals with the digital content of an object (sometimes called 'resource discovery metadata')
- administrative metadata which deals with administrative and technical matters
- preservation metadata which helps inform the potential long-term preservation of an object (this and administrative metadata are not necessarily disjoint categories)

Descriptive metadata might usefully be further subdivided into

- descriptive metadata about the digital content of an object
- metadata about the 'author(s)' of an object
- metadata about the context of an object

Purists may argue against the final two bullet points, but at the University of Hull we feel that this additional information can aid in the discovery of an object and to the extent of that purpose is justified.

We shall deal with each of the three main types of metadata in turn and examine how they might be derived automatically or, failing this, dealt with in a manual fashion. The intention of both approaches would be to prevent an author, or possibly repository administrator, needing to re-type the same details time and again.

For the purposes of the RepoMMan project, the automatic derivation of descriptive metadata was intended largely as a proof-of-concept and thus practical application of the work extended only to the population of a Dublin Core¹ based template. This work will be further developed for deployment at the University and it is intended that a much richer schema should be used

¹ See: Dublin Core Metadata Initiative (2005) *DCMI Metadata terms* At: <http://dublincore.org/documents/dcmi-terms>

from which simpler schemas can, if necessary, be populated by cross-mapping.² Work undertaken towards the end of the RepoMMan project was already showing some of the difficulties of working with no more than the simple Dublin Core metadata schema.

3 Descriptive metadata

3.1 Descriptive metadata about the content of a digital object

Automatic generation of descriptive metadata for text objects

It is perhaps ironic that the logical structure of this report places 'descriptive metadata about the content of a digital object' as the first of the categories to be considered because it is without doubt the one that has been the most difficult to deal with.

At the beginning of the project, desk-based research found a large number of websites that addressed the issue of automatically generating such metadata as keywords for an unseen text object; a few of these sites actually went further to suggest that their authors had some sort of tool under development to fulfil this function; a very, very few offered some sort of tool for download. In fact it proved almost impossible to identify an available, effective, open-source tool to meet this need. The closest that we came was to identify 'Kea' from the New Zealand Digital Library. This is clearly a useful tool but requires a controlled vocabulary to be really effective. Hull's vision of a repository used for development work across many uses and disciplines did not sit well with this approach. Later we were introduced to 'Data Fountains',³ a collaboration between the University of California (Riverside), the National Science Digital Library in the US, and the US Institute of Museum and Library Services. One component of the Data Fountains toolset, the iVia metadata tool, proved to be close to our needs and was eventually used in our proof-of-concept work.

Initial work with Data Fountains was conducted using their on-line test site.⁴ To quote from the site home page:

"NSDL Data Fountains is a tool for automatically and/or semi-automatically discovering and describing Internet resources about a particular topic. It explores for and finds new significant Internet resources, generates metadata records, and extracts rich full-text in order to help build or augment collections. In addition it is used to generate metadata for and extract rich full text from PDF and Postscript (ps and ps.gz) documents. NSDL Data Fountains is a service for NSDL projects. An account is required for use."

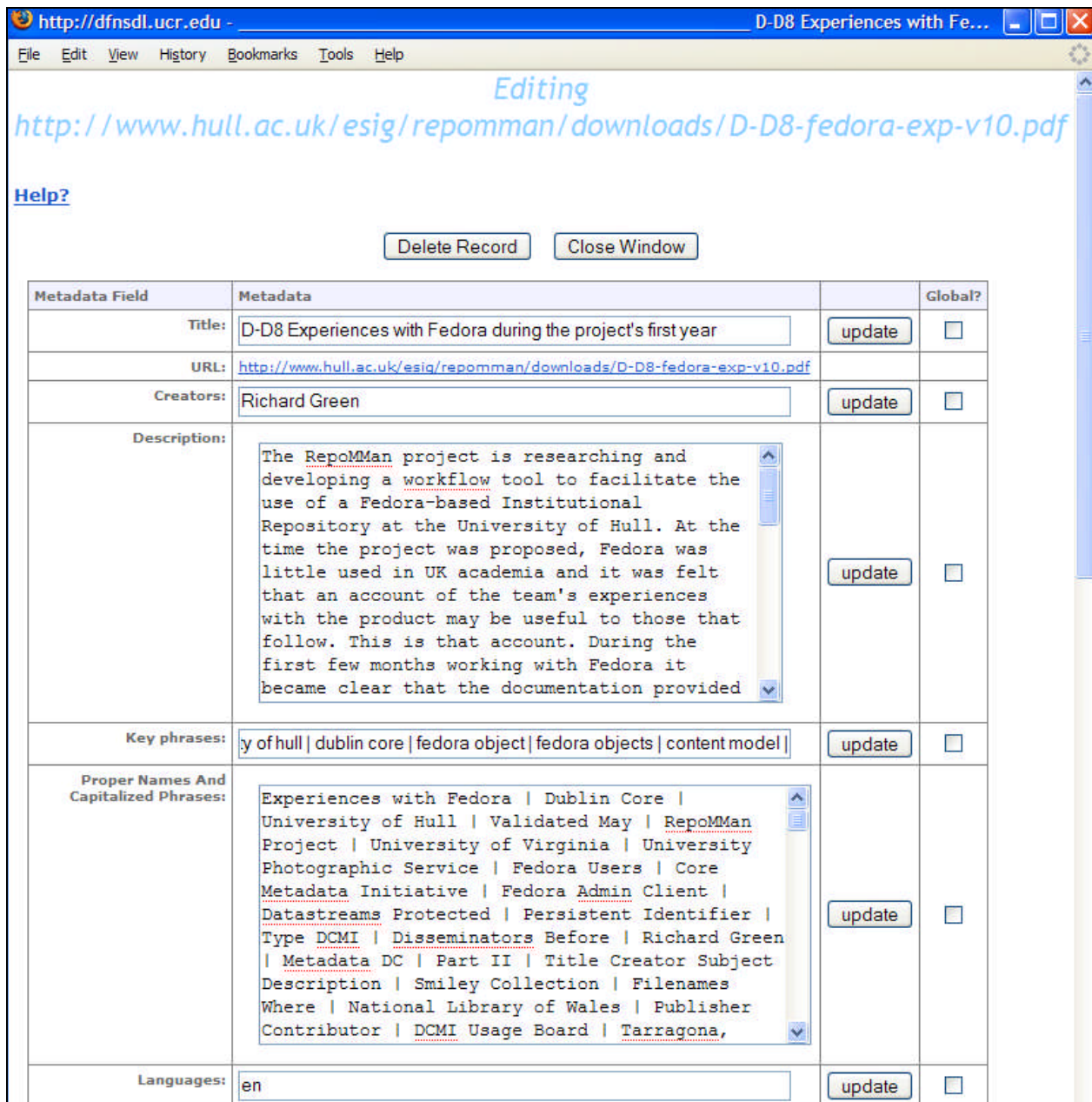
On the surface of it there seems to be only a small amount of overlap between the functionality offered by Data Fountains and the need of the RepoMMan project to derive descriptive metadata about an unseen text file. In fact, one of the components of the Data Fountains toolset does just this: the web page suggests that it only works for PDF and Postscript documents whereas it also works for html pages and, in a later version, for Microsoft Word documents.

Pointing the website version of the tool at a document early in our work produced the following, which is only the top part of a longer page:

² See R-D11 Green R & Awre C (2007) Report on metadata needs for the University of Hull Digital Repository

³ See <http://dfnsdl.ucr.edu>

⁴ See again <http://dfnsdl.ucr.edu>



The tool has identified within the document at the given URL: a title, an author (creator), a paragraph of description, a set of key phrases and the language of the document. Additionally it has extracted, as further potentially useful metadata, all the proper names and capitalised phrases that it identified. Not seen on this screenshot it also attempts to identify an abstract, a set of Library of Congress (LC) Subject Headings, a set of LC classification terms or numbers, a broad subject category and a file type.

This output greatly excited us because we recognised our document within the metadata. Clearly here is a tool with potential.

Over a period of time we installed a local version of the iVia metadata tool and added functionality to expose it as a web service which is capable of doing this analysis within the RepoMMan workflow and transferring the metadata to the RepoMMan tool where elements of it can be inserted into the metadata template available to a user. The user is then able to accept or edit the metadata to be associated with the material being stored.

Data Fountains is designed on a grand scale and natively uses large 64-bit computers. The version locally installed in Hull now runs on a much more modest 32-bit machine but its performance appears to be perfectly adequate for our purposes. We should acknowledge, at

this point, the prompt and courteous help given us over a period of weeks by members of the Data Fountains development team as we struggled to implement our local installation.

Beyond the lifetime of the RepoMMan project, the University of Hull will be further investigating the potential use of the iVia tool as a key part of its repository workflow.

Automatic generation of descriptive metadata for non-text objects

The iVia metadata tool is designed exclusively to work with forms of text object. We should consider also the automatic production of descriptive metadata for non-text objects. This was not an area that the RepoMMan explored. That said, we are not aware of available, open source tools that would, for instance, automatically provide keywords describing a picture or a dataset. We should be happy to be corrected.

Other approaches

If an automatic system is not available, there would seem to be no choice but to have an author (or other human agent) type the descriptive metadata for an object on a case by case basis. We have not come across any approach that currently avoids the need for this work. There has been a lot of interest in text mining as a tool that can be used for equivalent purposes, and there has been success in extracting concepts from medical texts and abstracts. Text mining works best when there is a substantial body of content with which to work, and suits wide collaborations that can provide this. It was not apparent whilst undertaking RepoMMan that text mining was an option available to individual institutional repositories, though we are continuing to monitor developments to identify advantages where they occur.

3.2 Descriptive metadata about the author(s) of a digital object

Metadata about an author should be just that, and there are those who would argue strongly that such information has no place in the metadata for an object other than by providing the name of its creator and/or author. However, others will argue that a minimal amount of metadata about the author might help the process of search and discovery around his or her works. This was a problem that we struggled with at the University of Hull and eventually, after a number of discussions with colleagues in the UK and overseas, we decided to hold a minimal amount of information:

- unique University ID
- name
- date of birth
- gender
- e-mail address

University ID: this is the key field for use with IT systems within the University. This is a unique identifier for an individual in perpetuity; it is not necessarily the same as other 'unique identifiers' that might be ascribed to an individual, for instance a staff or student number. It is potentially useful where a person changes their name because of marriage or other reasons.

Name: this will be held in a standard format (eg Larkin, Philip) matching that of the normal repository name authority (where appropriate the Library of Congress, in other cases the University should consider creating its own name authority files in the absence of a UK service).

Date of birth: this will be held in a standard format (1953-02-18). Apart from other uses, the date of birth will be a useful tool to resolve potential name conflicts outside the University system where the unique university ID is not applicable.

Gender: this may be useful to users where a forename is ambiguous (eg Jean, a girl's name in the UK, but a boy's name in France) or when a user wishes to do a specific search by gender, perhaps as part of a research exercise.

e-mail address: there was much discussion of this field but a consensus that it should be included. This would form the basis for potential contact and University staff e-mail addresses are already freely available from its public website.

Potentially, this metadata set may need repeating for multiple authors. The RepoMMan Project did not specifically address how to deal with generating multiple sets of entries potentially across multiple institutions.

For most of our users, for most of the time, this information is unchanging and it is precisely this sort of metadata that we do not wish our users to have to retype on each occasion they add metadata to a repository object that they have created. We see two viable approaches to providing this information automatically: taking the information from the network context in which the user is working, or taking an 'enter once, re-use many times' approach.

The University of Hull will surface the RepoMMan tool in one of two ways: either through the University Portal⁵ or through its preferred VLE, Sakai.⁶ In the short term it will be possible to derive a number of these items of metadata from the software environment that they provide. For instance, the portal stores a personal name for use as soon as a user logs in. In the longer term the University plans to make use of a central Identity Management System (IdMS). The IdMS, when fully implemented, will be able to provide all the metadata listed above.

Many institutions will be unable to source personal metadata in these ways. In that case we would recommend holding the metadata in a user's personal repository area, or elsewhere, so that it can be called upon and entered into a metadata set as required. This presumably requires providing users with a form on which to enter the metadata once, with the possibility of revisiting it to amend what has previously been typed. We have tested essentially this approach, though in a slightly different context. For the purposes of testing, we stored the metadata in a specific, and easily identifiable, Fedora object in the user's private space. When it came to a user entering metadata, the workflow searched for such an object and, if found, placed the metadata content into the appropriate fields of a larger form. This seems to us a simple approach that we could commend and we shall return to it in the next section.

It is recognised that there are a number of solutions available for recording metadata about people, for example FOAF, vCARD, eduPerson, DC-Agents, IMS LIP and HR-XML. An analysis for the PORTAL project in 2003 reviewed existing schemas and specifically compared the use of eduPerson and IMS LIP for use in describing users within a portal environment. This report favoured the use of eduPerson and the institutional portal in use at the University of Hull makes use of a subset of attributes that map onto equivalent attributes in the eduPerson schema. Since this report, eduPerson has become more widely adopted due to its use within the UK Access Management Federation for the exchange of attributes as part of Shibboleth authorisation.

It is important to consider the purpose of gathering metadata about an author alongside metadata for an object. It is clear that it is important to capture who created the object, and the discussion above highlights why the other fields selected were chosen. The purpose is to provide context for the object to aid search and discovery (of which more will be discussed in the next section) and it is not the intention to replicate full user records with each object. As stated, the institutional portal makes use of a subset of eduPerson attributes, and these will be used to capture fuller information about the user.

It was concluded that the limited metadata listed above was appropriate for the purpose at hand and extended solutions were not required. The development of people metadata

⁵ The University uses uPortal. See: <http://www.uportal.org>

⁶ See: <http://sakaiproject.org>

standards will be monitored for general institutional usage, and attributes added to digital objects in the repository where appropriate as part of additional descriptive metadata.

3.3 Descriptive metadata about the context of a digital object

Contextual metadata profile for research

In August 2006, the JISC-funded StORe (Source-to-Output Repositories) project at the Edinburgh Research Archive produced a report surveying the research use of repositories.⁷ The project was considering ways in which “repositories of published reports and papers [could] interact with the repositories of source data from which, in general, they are derived.” Thus researchers who were questioned in this context had a very broad view of metadata needs. Respondents were asked specifically about metadata they might attach to data sets (rather than to reports or papers relating to the datasets):

“The term **metadata** refers to the information or labels that you use to identify and describe your data. The principal purpose of metadata is to make it easy to recognise, access and retrieve data. A familiar example of this would be a library catalogue, which is a collection of metadata records that in their simplest form might contain details about the title, author and date of publication of each item in the library. So, assigning metadata enables one to understand what is contained within a set of data, when and how it was created and by whom, and the relationships it may have to other data. By selecting from the following options, please would you indicate what types of metadata you consider it important to assign to your data.”⁸

Suggested in the StORe project survey were the following; figures show the percentage of respondents who thought the field important:

- author or data creator name(s) (83.3%)
- project title (78.8%)
- subject keywords (69.2%)
- project description (67.6%)
- title of data set (65.8%)
- date of data creation (64.2%)
- format (eg pdf or HTML) (53.8%)
- project identifier (45.1%)
- dates of project (42.4%)
- publisher (28.9%)
- funding source (20.4%)

Whilst some of these items of metadata are clearly essential, others of them would serve to provide a richer context that might help discovery of a digital object resulting from the research. Equally, the majority of these terms would provide enriched metadata for a paper or report just as much as for a data set. If we accept that this is a valid list to work with, two questions arise: where the metadata might be obtained and how?

Source and feasibility of contextual metadata for research

In an ideal world, it would be nice to envisage an on-line directory of research-related information on which the RepoMMan tool could draw, in a similar manner to deriving personal metadata from an IdMS. Such a directory might go beyond storing details of a current project into more general areas such as research interests, associated researchers and so on. At the time of writing, systems at the University of Hull do not support this approach and so it will be necessary to consider an alternative solution, which we, and others in a similar situation can

⁷ Pryor G (2006) *Survey of Researcher Use of Repositories* StORe Project, University of Edinburgh

⁸ *op cit* p24

use. Given that derivation of this metadata will be managed by a Web Service, it will be relatively straightforward to provide an alternative service that derives it in a different way.

Many of the metadata fields noted in the preceding section will be largely static across the timespan of a particular research project. It may be appropriate, therefore, to have a researcher fill these in, either on a personal or group basis, the first time they are needed, but then store them for use with future, related digital objects. The RepoMMan tool could interrogate the user's repository area for such information and use it to pre-populate some elements of the metadata on subsequent occasions. Further refinement of the approach would allow multiple metadata sets to be held relating to different projects, allowing the user to choose the appropriate one as needed, possibly from a drop-down list. This is the approach we shall adopt in the short to medium term at the University of Hull and is merely a variation on the method described above for local storage of personal metadata.

A small number of the metadata fields do not lend themselves to this approach:

- *subject keywords* - whilst some of these may be regarded as static, others will need to be derived on a file-by-file basis. As noted above, the RepoMMan Project has successfully implemented a system for the automatic extraction of keywords from textual information, but it is unlikely that the same will be possible for numeric or other material.
- *title of dataset (or paper)* - we might reasonably assume that this can be automatically extracted from a text, but for other material it may need to be user-provided
- *date of (data) creation* - this field might easily be pre-populated with "today's date" but the user would need to be able to override this.
- *format* - the RepoMMan tool will assign a mime-type to an object when it is first created and this information can be automatically transferred to the metadata field; indeed it can be argued that it is better not to show this to the user for potential editing because a machine test is more likely than a human to identify and describe the format correctly.

The research metadata profiles would themselves be a digital object or objects within the user's repository area. As in the previous section, our approach would be to make them easily identifiable to our workflow engine so that a drop-down list can be created of any or all found.

The approach outlined for researchers (either individually or as groups) can be extended to other users. At Hull we shall be extending it to provide contextual profiles also for work in the Learning and Teaching (L&T) field and for administration. Again it is likely that a particular user may need a number of context profiles covering, perhaps, their involvement in different taught courses or different areas of administration. It is possible that a user can have profiles covering one, two or all three of these areas (research, L&T, administration).

Contextual metadata profile for Learning and Teaching

Discussions with those in Hull's L&T community and others have elicited the following as possible contenders for contextual metadata elements around their work:

- Courses/degrees associated with the object
- Position of object contents in overall course structure
- Educational level/intended audience
- Format/technical requirements for use
- Associated assessment
- VLE used
- Associated discussions/forums
- Copyright/rights
- Title
- Description
- Identifier
- Creation date

- Language
- Material type

It may be possible to extract some of these elements from the materials themselves in an automated fashion, however other elements probably require 'manual' entry and may, or may not, lend themselves to a stored profile approach

Source and feasibility of contextual metadata for teaching and learning

Contextual metadata for L&T as it relates to a particular person is likely to be rather less static across different digital objects than in the case of research metadata. Whilst a researcher might only be engaged in only one research project at a time, it is not uncommon for members of the L&T community to be dealing with more than one course at once. It is thus likely that multiple profiles will need to be managed.

Contextual metadata profile for administration

Interviews with administrators at the University have helped formulate a range of possible metadata fields that could usefully be associated with "official" documents such as policies and committee papers:

- Committee/role/section
- Document owner
- Document creator
- Rights⁺
- Identifier⁺
- Title⁺
- Keywords⁺
- Description⁺
- Creation date⁺
- Language⁺
- Status (draft, final etc)
- Format⁺
- Freedom of Information Act (FOI) status
- Preservation status
- Management flag

A number of these, marked ⁺, are central to the metadata that would be associated with almost any digital object and will not be dealt with further here. Others, though, do fall into the category of contextual metadata associated with the author and some others pose interesting problems when looked at from the perspective of some administrative users.

Administrators potentially introduce a particular challenge that falls outside the scope of the RepoMMan project but which should be mentioned here, and it is one that we shall shortly have to address at the University of Hull. Administrators are often working as a proxy for other individuals or groups. Whilst a particular person might well be the scribe of a set of minutes, it can be argued that the 'author(s)' are the members of the committee that (s)he services. Both contextual and personal metadata will need to be dealt with carefully in such a situation.

Discussion with Thornton Staples from the University of Virginia resulted in the following recommendations relating to the first three items above: committee, owner and creator. In the case of a University policy document, the University should be deemed owner and creator. In the case of committee papers, the committee as a body collectively created and therefore owns the document. Individuals can be listed as "contributors" and the committee secretary(ies) should be listed as contributors with this specific role.

Source and feasibility of contextual metadata for administration

As with the cases above, it may be that specific document creators in the administration field will be able to generate one or more metadata profiles that can stand as the basis for pre-populating metadata and contextual metadata fields associated with their outputs. However, other potential administrative users may be creating a digital object as a proxy for another person or group and it may be more difficult, therefore, to extract appropriate metadata for the object from the user context in which it is generated. Equally, there may be occasions when these people generate objects in their own right and the user context is therefore the logical source of some metadata.

Clearly, one needs to consider what is the best 'default' case in these situations. Identifying the metadata is not a problem *per se*, rather the issue is what is the appropriate metadata to use. In addressing this and how to best generate the metadata profiles one opportunity and one threat require attention. The opportunity of developing profiles fits well with emerging trends in personal identity management (e.g., through developments such as OpenID), and this trend will be capitalised upon where favourable. The threat emerges from the generation of profiles that lead to duplication of and multiple copies and/or versions, an issue that will have to be dealt with via regular review of profiles and their usage.

4 Administrative metadata

We have used the term 'administrative metadata' to describe here metadata relating to technical and to administrative matters. This should be kept clearly distinct from the last section above relating to metadata generated by or about administrators.

Technical metadata

As with descriptive metadata, dealt with in section 3, it is possible to identify a number of open source tools available on the web which claim to be able to interrogate the potential 'content' of a digital object (the 'digital payload, if you will) and to provide a range of technical details about it. Probably the best known of these is JHOVE⁹ and it is that tool that the RepoMMan project has investigated in some detail and which will eventually form part of the University's 'production' system.

Ultimately, it is likely that the University's administrative metadata will comprise the appropriate elements from the following list:

```
<pid>
<identifier>
<source>
<adminrights>
  <policy>
    <access>
    <use>
<technical>
  <image>
  <text>
  <audio>
  <video>
  <geospatial>
  <stats>
```

Here is not the place to deal with the syntax and possible sub-structure of all these elements: this is dealt with in a separate booklet¹⁰. However it is appropriate to look at some of these elements in more detail and consider where and how the data might be derived.

⁹ See: <http://hul.harvard.edu/jhove/>

¹⁰ Green, R (2006) University of Virginia Metadata Schema, RepoMMan Project, University of Hull

pid

The persistent identifier of a Fedora object is determined at the time of its first ingest into the repository. Allocation of pids is essentially machine-controlled so that duplicates cannot be issued and in Hull's implementation a user of the repository will have no way of over-riding this. In a Fedora repository it is possible to know what pid will be allocated to a new object before the event and so to include it in the metadata.¹¹

identifier

The identifier in the administrative metadata datastream is a reference to the location of the digital payload within the extended Fedora system. Whilst the requirement is for a filename only, the Hull proposal for naming repository-related files will make it clear where the file should sit inside the repository file hierarchy. This element can be filled in automatically as the object is generated.

source

<source> is one of the few elements here that is not susceptible to automatic generation. If this element is to be populated, it is likely that it will have to be filled in 'by hand' although it may be that, for a text object, the iVia metadata tool may be able to identify something. Whether it could do so reliably is not something that we have tested.

adminrights

There are two important components to <adminrights>, those relating to access and those relating to use.

'Access' determines who is able to 'see' the object within the repository. The entry here will be drawn from a simple list including:

- No access
- Staff only
- Students and staff
- Public

'Public' access, of course, allows access to staff and students of the University too. In practice access can be controlled in a number of ways but where it needs to be restricted the most common will be to limit access to a group or groups defined within the University Identity Management System - the highest levels of these being 'staff' and 'students'.

'Use' details any licensing arrangements that there might be for use of an object's payload. Again this may be drawn from a simple list, but more complex entries could be made.

<adminrights> is the other set of elements that cannot easily be derived automatically on entry although defaults could be provided until human intervention populated the elements appropriately. The defaults would necessarily be 'no access' and 'no use permitted'.

technical

The technical metadata can be derived from the output of the JHOVE tool. The tool will be called as a service by the RepoMMan workflow tool to provide technical information about the payload for an object. It is primarily designed to cope with textual and image content, but will return minimal data for some other file types. Of the six possible tags in the technical section it is likely that only one would apply to any given object. Our repository structure envisages

¹¹ This is achieved through an API call 'getNextPID'

that where an author would like to 'link' say a text and an image, these would be created as separate objects - each with its own metadata - with a parent object which asserts a relationship between them. This atomistic approach gives rise to so-called 'complex objects' which actually comprise a related set of simpler ones. It is not uncommon for the parent object to hold metadata in common for all its children. It was outside the scope of the RepoMMan project to investigate how this might be done (either manually or automatically).

The reader may still be a little unclear as to what is meant here by 'technical metadata'. An example for a simple pdf object and for a jpeg image object may serve to make this clearer:

```
<technical>
  <text>
    <filesize type="bytes">57009</filesize>
    <character>UTF-8</character>
    <mimetype>application/pdf</mimetype>
  </text>
</technical>

<technical>
  <image>
    <filesize type="bytes">4366264</filesize>
    <format>
      <mimetype>image/jpeg</mimetype>
      <compression>6</compression>
      <colorspace type="JPEG">RGB</colorspace>
      <spatialmetrics>
        <imagewidth type="pixels">2592</imagewidth>
        <imagelength type="pixels">1944</imagelength>
        <sourceX type="pixels">2592</sourceX>
        <sourceY type="pixels">1944</sourceY>
      </spatialmetrics>
      <bitspersample>8,8,8</bitspersample>
      <samplesperpixel>3</samplesperpixel>
    </format>
  </image>
</technical>
```

In fact the wealth of information provided by JHOVE can potentially be put to other uses. We shall deal with preservation metadata separately but it is worth noting here that the information returned by JHOVE from a digital camera picture includes all its EXIF 2.2 information¹² which might contribute to other sections of metadata.

5 Preservation metadata

We shall deal with preservation metadata only briefly as it was not an area with which the RepoMMan project had intended to engage. It is, however, part of the focus of the University's follow-on REMAP project¹³ which will extend the ideas developed by RepoMMan into the areas of records management and preservation.

A specific workpackage within the REMAP project will be identifying and assessing relevant services available to assist with preservation activities in general. In the context of services that can be used to automatically generate metadata for inclusion alongside the relevant digital objects, though, two are worthy of mention here. The National Archives has developed the DROID format identification tool¹⁴. This is a downloadable tool that can be used to perform automated batch identification of file formats, and the information can be associated with the digital objects concerned to aid future preservation. The DROID makes use of the information

¹² See: <http://en.wikipedia.org/wiki/EXIF> for a further explanation of EXIF data

¹³ See: <http://www.hull.ac.uk/remap>

¹⁴ See: <http://droid.sourceforge.net/wiki/index.php/Introduction>

contained within the associated PRONOM registry of file format information. The CRiB¹⁵ service based at the University of Miño in Portugal is available via Web Services. It evaluates digital objects for optimal migration options, and generates reports that can be embedded in preservation metadata.

Suffice it here to say that a common, if blunt, approach to preservation metadata is to store the entire output of JHOVE in the digital object. For a pdf file, for instance, this metadata can run to dozens of pages of A4 paper. It is likely that the University of Hull will adopt this approach in the short term but refine it in due course. JHOVE was mentioned in the previous section and it will thus be realised that there is often an overlap between technical and preservation metadata. The REMAP project will need to consider how best to deal with this.

6 Summary

At the outset of the RepoMMan project we were curious to know how much automated metadata production might reasonably be employed around the digital objects that we intended to produce. University systems and the repository itself offer a number of mechanisms whereby administrative metadata might be assembled. Having found JHOVE to deal with technical metadata the team was somewhat despondent that a similarly effective tool could not be found for descriptive metadata. Indeed, as we talked to colleagues world-wide, it became clear that the search for such a tool was somewhat akin to that for the Holy Grail. The discovery of 'Data Fountains' and, in particular, the iVia metadata tool gives us hope that some effective descriptive metadata might indeed be generated automatically.

At the time of writing, August 2007, the iVia metadata tool has only recently been deployed against the RepoMMan system and it is as yet too early to comment on the quality of the metadata that it is producing. An examination of this is shortly to be undertaken by the RepoMMan team jointly with the DEST-funded Arrow Project¹⁶ in Australia.

¹⁵ See: <http://crib.dsi.uminho.pt/>

¹⁶ See: <http://www.arrow.edu.au>