CENTRE FOR DIGITAL LIBRARY RESEARCH

cdlr.strath.ac.uk

Evaluation of metadata workflows for the Glasgow ePrints and DSpace services

R. John Robertson

June 2006

1 Executive Summary

The institutional repositories at the University of Glasgow which began as part of the DAEDALUS project have developed into an integral part of Glasgow University Library's services. Using both EPrints.org and DSpace, they provide access to, and permit management of, the University's academic digital assets.

This evaluation analyses and comments on the metadata workflows of these services, their support for metadata quality, and how changes in purpose, which have accompanied their transition from project to service, have influenced the repositories. This evaluation will be of benefit not only to DAEDALUS but also to other institutional repositories facing the transition from project development to operational service.

The metadata workflows supporting the management and retrieval of ePrints offer a number of paths for metadata creation – each of which has seen shifts in their relative importance as the purpose of the repository has evolved and become clear. The management and retrieval of other academic content in the DSpace service is entirely mediated by repository staff and follows a basic workflow. The quality of metadata in both services has been maintained through staff training and the ongoing involvement of professional cataloguers.

The strengths of both repository services lie in their clarity of purpose, utilisation of appropriate software to support those purposes and their successful integration into Glasgow's institutional context.

Although they also present a significant opportunity, the new challenges faced by the repository services arise from the emerging involvement of non-specialists in the creation of records and their potential involvement in the administration of sections of the DSpace repository.

To address these challenges, the repository services will have to maintain their clarity of purpose, monitor metadata quality, capitalise on opportunities for efficiency, and continue to significantly engage in advocacy and user training.

2 Introduction

2.1 Aims

This short study evaluates metadata quality assurance and workflow for the DAEDALUS project, and the ongoing service that has developed from it. The study considers the viability of a range of approaches to populating the DAEDALUS institutional repositories, together with their workflows. This work was carried out against a backdrop of previous research on repository workflows carried out in projects such as Mandate (http://mandate.cdlr.strath.ac.uk/) and Stargate (http://cdlr.strath.ac.uk/stargate/).

2.2 Objectives

Specific objectives of the study were:

- to examine and analyse the workflows developed by the DAEDALUS project;
- to examine and analyse changes in the workflows as DAEDALUS has progressed from project to service;
- to utilise the results to recommend metadata workflow models that will ensure that metadata created will be of sufficient quality to support the functionality required of the University of Glasgow's Institutional Repository Services in terms of information discovery and retrieval, repository management and interoperability with other systems;
- to assess workflow issues related to the capture of technical and administrative metadata required for digital preservation;[1]
- to identify the critical points at which metadata content that must be acquired from academic staff may be captured and deposited in the repository during the publication process;
- to assess the sustainability of the developed workflow models;
- to identify associated training needs and necessary support mechanisms for those creating metadata and managing repositories within the University of Glasgow.

2.3 Outputs

The primary outputs of the study are:

[1] Defining such technical and administrative metadata is outwith the scope of this evaluation; it will rather suggest when, where, and how such metadata might be created within the DAEDALUS repositories' workflows.

- a report on the use and effectiveness of the workflow models developed by the Glasgow ePrints Service and the Glasgow DSpace Service;
- recommendations on best practice for the continued use and development of such models (including the “self-deposit model” and the “mediated model”).

2.4 The Phases of DAEDALUS

For the purposes of this study, there are two distinct phases of DAEDALUS to be evaluated:

- the JISC-funded project phase during which the focus has been to deliver the project’s key aims and objectives (August 2002 – July 2005);
- the service phase which will have a broader remit in line with long term institutional goals relating to open access, the Research Assessment Exercise, and research excellence.

The study is both a summative evaluation of the workflow element of the DAEDALUS project and a formative evaluation of the workflow for the ongoing DAEDALUS service in both the medium and long term.

2.5 Acknowledgements

The author would like to thank the DAEDALUS’ repository team, in particular William Nixon and Morag Greig, for their cooperation with, and generous availability to, this evaluation. The evaluation builds on preliminary work by Jane Barton.

3 Background

The DAEDALUS project was funded under the JISC Focus on Access to Institutional Resources (FAIR) Programme and ran from August 2002 until July 2005.

> "The [FAIR] Programme was inspired by the Open Archives Initiative and sought to make funding available to investigate the mechanisms and technologies required to enable the deposit and disclosure of digital materials. Specifically, it sought to examine how the increasing body of digital materials produced by Higher and Further Education institutions might be made more accessible to the wider community. Many institutional resources are either consigned to print publications or hidden across a multitude of computers within an institution, when they could be shared more effectively. The Open Archives Initiative has been found to be valuable, if possibly limited, in addressing this, whilst the place where resources are deposited, the repository, has come to the fore as a key part of the successful management and delivery of these resources."[2]

Within this programme DAEDALUS' main aims were as follows:

- to establish and populate a range of OAI-compliant digital collections at the University of Glasgow using a range of different OAI-compliant pieces of software;
- to act as a catalyst for cultural change and ongoing discussions about the crisis in scholarly communication within the University of Glasgow and the wider community;
- to disseminate experiences and findings to the wider community through reports, workshops, exemplars and guides to best practice in the development of these services;
- to liaise with other members of the e-prints and e-theses cluster to share experiences, co-host events as appropriate and ensure co-ordination within the cluster;
- to investigate digital preservation issues in conjunction with the SHERPA project (http://www.sherpa.ac.uk).

The DAEDALUS project plan lists a number of types of digital materials to be exposed through the developed repository services:

- published and peer reviewed academic papers;
- pre-prints and grey literature;
- doctoral theses;
- research finding aids;
- university administrative documents.

[2] Chris Awre (2005) *FAIR Synthesis Introduction*. JISC Available from: http://www.jisc.ac.uk/fair_synthesisintro.html.

Although DAEDALUS undertook these aims and objectives in the context of a project, the staffing allocation and intent throughout the project has also reflected a longer-term desire to integrate and embed the development of the repository into the University's library services.

The project's development of repository services at the University of Glasgow and the issues it encountered has been thoroughly documented and disseminated by the project team. A comprehensive list of their presentations, reports, and publications is available at: http://www.lib.gla.ac.uk/daedalus/papers/index.html and http://www.lib.gla.ac.uk/daedalus/docs/index.html.

Although the focus of this study is explicitly on the evaluation of the *metadata* workflows, this cannot take place without some overlap with, and consideration of, DAEDALUS' workflows in general. The effectiveness and efficiency of the repositories' workflows (both metadata and general) is vital to ensuring the creation of high-quality metadata (necessary to support effective information retrieval and asset management) without overly burdening the academics and library staff involved. Inefficient or inadequate workflows can produce lesser quality metadata and lead to problems for the delivery of services– these problems can both directly affect information retrieval results and indirectly exacerbate social and cultural problems in the adoption of new repository services (if users invest time and effort in supporting a service they expect it to work). If it fails to work well, a negative view of it will develop that may impact on service adoption by academics and institutional support.

One major focus of the project was an advocacy campaign to obtain information about, and copies of, published peer-reviewed papers from the University's academics. This required that the repository's procedures for ingesting and supporting access to this content be highly developed. One effect of this is that the workflows for this asset type (managed using the EPrints.org software) matured quickly during the project.[3]

[3] Although effective, the comparable practices for other asset types are comparatively less mature.

4 Methodology for the evaluation

The methodology used in the study had three components:

- a consultation exercise to gather information on policy, rationale, and practice regarding workflows in the DAEDALUS ePrints and DSpace services;
- a review of current theory and practice relating to workflows, and particularly metadata workflows, within digital repositories;
- an evaluative analysis of the metadata workflows used within the DAEDALUS ePrints and DSpace services.

The consultation exercise was carried out through a review of the project's documentation, a series of meetings with the repository staff, and an examination of the systems themselves. The review of current theory and practice was informed by ongoing research on institutional repositories and workflows conducted by the Centre for Digital Library Research (CDLR) within Mandate (http://mandate.cdlr.strath.ac.uk/), Stargate (http://cdlr.strath.ac.uk/stargate/), and other projects.

The methodology used to evaluate the metadata workflow within DAEDALUS (the last component of the evaluation) was based on the workflow design framework (*figure 1* below) developed as part of previous work undertaken by the CDLR. Within this context, the evolving purpose of the repositories, the development of their workflows, the workflows themselves, and the quality of the resulting metadata was evaluated, to produce a rounded assessment of the effectiveness and sustainability of the workflows in the context of their institutional setting and their wider information environment.

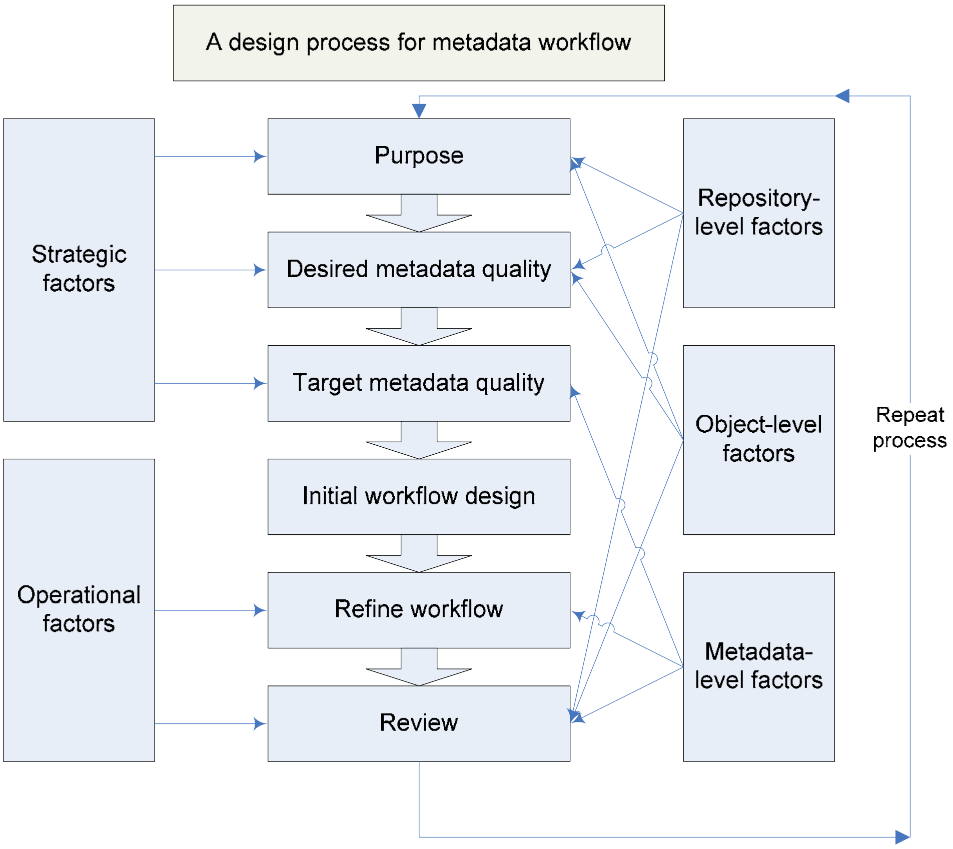


Figure 1: Design framework for metadata workflow in a distributed information environment

Although intended to support repositories as they set up their workflows, the metadata workflow design framework also provides a structure to support the evaluation of a workflow. The central column of the framework provides a simple overview of the stages of the process of designing a workflow and how those stages are affected by different internal and external factors. Strategic and operational factors, on the left-hand side of the model, define the purpose of the repository, the resources available to carry that purpose out, and the local context in which that workflow is being designed. The factors on the right-hand side represent external influences on the workflow (such as interoperability considerations, software functionality, and choice of base metadata schema).

One implication for DAEDALUS immediately apparent from the design framework is that, with different software packages are in use, there will be a metadata workflow for each service (since each software package brings different external factors into play at repository- and metadata-levels). There may also be multiple workflows within each service to support assets of different types or purposes; since they may then have different metadata requirements or involve different staffing commitments (a workflow to deposit and create metadata for an e-thesis may look different to one for a finding-aid – the thesis may require input from, and information about, the supervisor, the finding aid may require input from an archivist and overview information about the collections it details). A further implication, given the likely shift in the strategic and operational factors influencing a repository between a development project and an active service, is that the workflows for the project phase and the service phase are very likely to be different (for example, there may be a reduction in the service's budget, it may be asked to manage new types of asset, or to support new functions such as an RAE audit).

What this means in respect of the ePrints element of the study, is that there are eight identifiable workflows within the service. Four distinct workflows in the project phase – one for each of the different types of deposit- and four distinct workflows in the service phase – reflecting the changes in the context of each of the types of deposit. These two sets of workflows allow not only the analysis of the workflows themselves but also the opportunity to highlight workflow issues and metadata quality issues arising in the transition from project to service.

The DSpace situation is different in so far as the DSpace service has matured less quickly. Although there are a number of workflows present within the service, they are less distinct and there are fewer changes between the project and service phases. It is expected that as the use of DSpace repository continues to develop significant workflow changes will emerge. With this in mind, the DSpace element of the study utilised the design framework to support formative analysis of this service.

Note that, within institutional repositories which contain the full text of papers as well as their metadata, metadata workflow is to an extent bound up with the workflow for capturing a suitable version of the paper (or other type of asset). The DAEDALUS ePrints repository contains both metadata and full text, and the associated workflows are closely integrated. As such, it has been useful to consider this broader concept of workflow throughout.

5 Key findings from the consultation exercise

5.1 Phase 1: Project DAEDALUS ePrints

5.1.1 Purpose

In exploring the provision of academic e-content, DAEDALUS elected to manage published peer-reviewed articles, conference papers and book chapters (refereed publications) separately from other e-content; EPrints.org software was chosen to provide the required functionality.

The project's role as a community exemplar and advocate for the development of repositories in other institutions led to an emphasis on gaining a critical mass of articles as soon as and as quickly as possible. Acquiring content, however, proved to be a slow process and one of the effects of this was incorporation of metadata-only records into the ePrints service – these records contained information about the final article and linked to publisher's copies where possible, but did not provide access to an author's final post-print or other open access version of the paper. Such records allowed academics without access to the final version of their paper to participate in the service. This allowed for a more rapid population of the repository and an early examination of issues that tend to be more evident in mature and populous repositories (e.g. the need for authority files – controlled lists of preferred or mandatory terms for certain metadata fields).

As a result of the incorporation of metadata only records there was a shift in the purpose of the project's repository from providing, where possible, open access to the University's published output to providing a record or showcase of all of the University's published output, together with open access to a subset of this. A consequence of this shift was that the project undertook to include publications retrospectively; i.e. project staff would mediate the creation of metadata records for an academic's back catalogue of papers and where possible also mediate the deposit of open access versions of them.

5.1.2 Metadata quality

The requirement to be a community exemplar and a showcase for University of Glasgow publications, led the project to aim to optimise metadata quality. To ensure this it chose to include specifically-trained staff in the editing of author self-deposit records and the creation of mediated deposit records, and involve the library's Bibliographic Services team in the classification of all records.[4]

The metadata from the ePrints and DSpace services is exposed for harvest via the OAI-PMH by both Glasgow's local pilot search service (http://daedalus.lib.gla.ac.uk:83/pkpharvester/harvester/index.php) and international services such as OAIster (http://oaister.umdl.umich.edu/o/oaister/). The metadata is also available to Google Scholar (http://scholar.google.com/) and Elsevier's Scirus service (http://www.scirus.com/). The local harvester, using PKP harvester (http://pkp.sfu.ca/?q=harvester), provides access to the entire academic e-content provided by DAEDALUS and provides different browse access points including repository and content type.

5.1.3 Workflows

Initially, the metadata and content workflows were based on the self-archiving model, in which academics deposit and describe their own papers. However, it became apparent early in the project that although there was considerable support among academics for an institutional ePrints archive, very few were sufficiently motivated to archive their own papers. In order to ensure the timely delivery of the project, a mediated model was introduced, in which project staff deposited and described papers on behalf of academics. This mediated model was supported by a concerted advocacy programme, but, even so, very few actual papers were forthcoming. To achieve a body of full text content and associated metadata within project timescales, it was often necessary to use the publisher's electronic copy (or even, in a small number of cases, to scan from print) and to use project resources to create the metadata.

The final metadata and content workflow for the project phase is documented in some detail by the project in *Populating the Glasgow ePrints Service: A Mediated Model and Workflow* (Nixon and Greig, 2005) but is outlined in figure 2. It can be seen that, within the ePrints service there are four distinct workflows: one for self-deposit; one for departmental bulk deposit; one for mediated deposit for small numbers of articles and one for the mediated deposit of large numbers of articles. Much of the metadata creation and checking is carried out by the project's administrative assistant (this was not initially expected but has worked

[4] Records are classified using Library of Congress Subject Headings.

well). The Library’s Bibliographic Services department then classifies the record before accepting it into the repository.

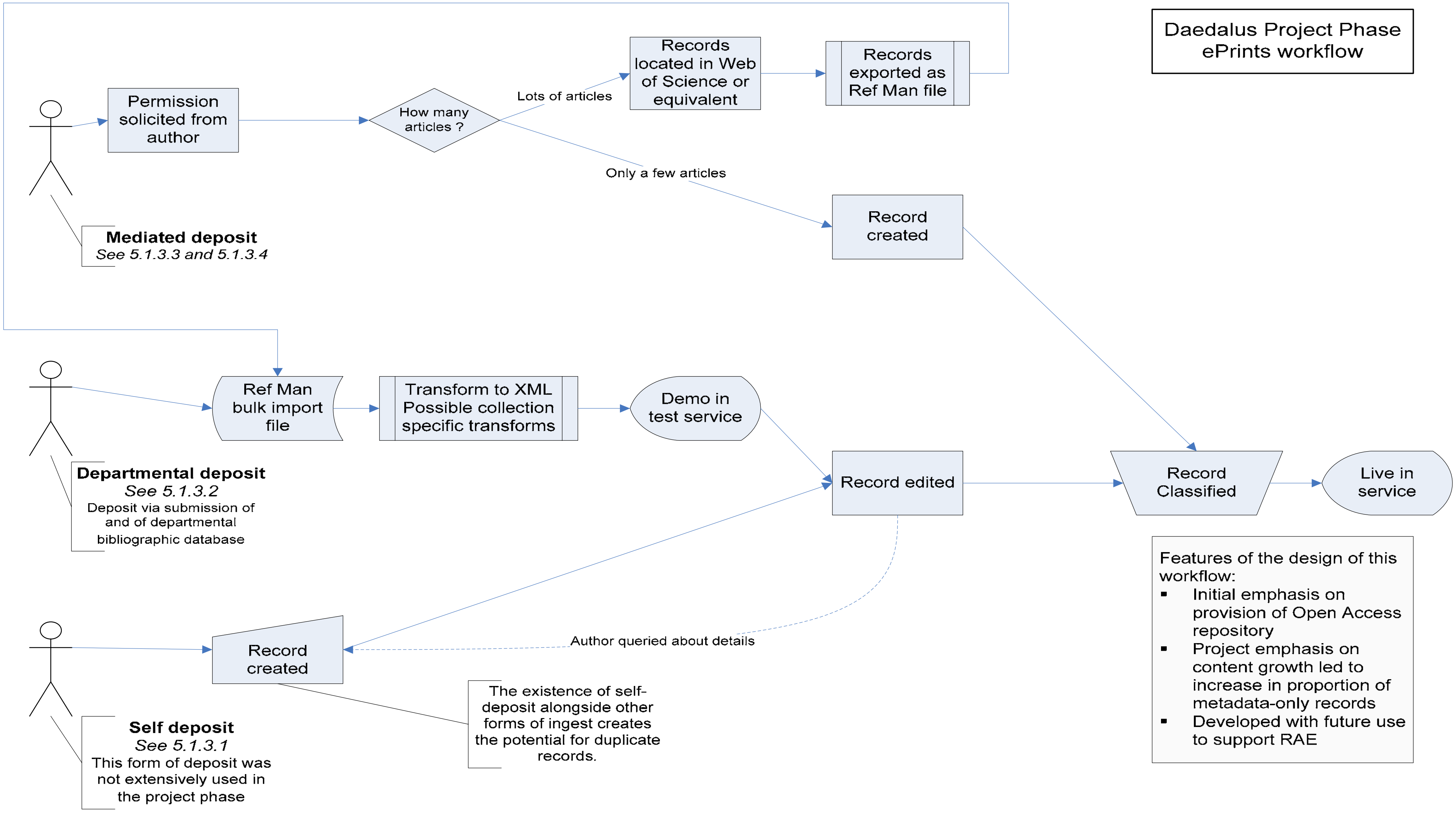

Daedalus Project Phase ePrints workflow
Permission solicited from author
How many articles ?
Lots of articles
Records located in Web of Science or equivalent
Records exported as Ref Man file
Only a few articles
Record created
Mediated deposit
See 5.1.3.3 and 5.1.3.4
Ref Man bulk import file
Transform to XML Possible collection specific transforms
Demo in test service
Departmental deposit
See 5.1.3.2
Deposit via submission of and of departmental bibliographic database
Record edited
Record Classified
Live in service
Features of the design of this workflow:
▪ Initial emphasis on provision of Open Access repository
▪ Project emphasis on content growth led to increase in proportion of metadata-only records
▪ Developed with future use to support RAE
Record created
Author queried about details
The existence of self-deposit alongside other forms of ingest creates the potential for duplicate records.
Self deposit
See 5.1.3.1
This form of deposit was not extensively used in the project phase

Figure 2 DAEDALUS project phase ePrints workflow

5.1.3.1 The self-deposit workflow

In this workflow, authors register with the service and create record(s) for their article(s). The fields they can create are restricted and the record is edited and classified later by library staff. Within DAEDALUS, processing a self-deposit record for a single article takes almost as much time as creating a mediated record. From the University's perspective therefore the additional time taken by academics in the self-deposit workflow should be included as an additional cost.

The self-deposit approach has seen very little activity and, after the initial period, was not encouraged in the project phase. It is regarded as producing small numbers of records that require significant editing. Although paper deposit is strongly encouraged, this workflow does not require it. Authors can create a metadata-only record.

5.1.3.2 Departmental mediated deposit

In this workflow, bibliographic records created within a department or research group's reference management system are imported into the ePrints system and the papers are added when academics are able to locate and provide a copy. This bulk import of records takes advantage of the existence of publication lists throughout the institution, which are available in the University's site-licensed bibliographic software (Reference Manager or Endnote).

Perl scripts are used to convert Reference Manager files into xml files, which can then be imported into the ePrints software, automatically populating the database with batches of metadata records. The information provided varies considerably in quality, and work may be required to bring records up to standard. Where possible, the basic conversion scripts are edited to incorporate automated improvements in metadata quality, for example to provide default content for specific metadata fields (such as copyright statement, or publisher details for a batch of records from the same journal (described in Drysdale, 2004)). Within the project this has proved to be the most productive and efficient workflow. It capitalises on metadata creation efforts within departments and research groups, and permits a degree of automation.

5.1.3.3 Individual mediated single- or few- article deposit

This is the simplest workflow. In this, permission to deposit an article, or a few articles, is sought from the authors and where possible a copy of the paper is obtained. The records for these items are created by an editor and then passed to bibliographic services for classification.

5.1.3.4 Individual mediated bulk deposit

In this workflow a semi-automated approach has been developed which parallels the departmental mediated deposit. An individual or department (without a bibliographic database) requests that a selection of articles is entered into the system. The references for these articles are located by project staff in Web of Science or an equivalent service (http://scientific.thomson.com/products/wos/) and exported as a Reference Manager file. This file is then imported as if it was a departmental mediated deposit.

In the project phase, this approach enabled the rapid population of the repository with individuals' metadata. However, in the same manner as the departmental deposit, this approach had the effect of decreasing the proportion of records with full text available. Although this approach is more efficient and more accurate alternative than manual data entry, searching for and finding the reference remains time-consuming.

Both these forms of mediated deposit have been boosted by an extensive advocacy programme, which has encouraged keen individuals to participate ahead of wider departmental involvement.

5.2 Phase 2: The University of Glasgow's ePrints service

5.2.1 Purpose

There have been two significant changes in the purpose of the ePrints service as it has developed from project phase to operational phase. The first of these changes is the shift of the service back to an emphasis on open access provision. The second is that the use of the service to support the Research Assessment Exercise (RAE), as anticipated in the project phase, is no longer considered likely.

The first change in purpose has emerged in response to the increasing proportion of metadata-only records in the service and a reassessment of what the service is there for. This reassessment has concluded that the service is for the open access provision of refereed publications (as opposed to simply metadata) – the original purpose of institutional repositories.

The second change in purpose has emerged in response to two general trends: the slow uptake of author archiving in institutional repositories, and the RAE requirement to submit the publisher's final version of the paper. Given the slow rate at which institutional repositories are being populated and the increasing reluctance of publishers to allow the inclusion of their final publisher version in repositories, it currently looks unlikely that an institution will be able to provide access to the four best papers of each academic via an open access institutional repository.[5]

For the ePrints service this change in purpose has led to a decision not to accept metadata-only records in future, except in exceptional circumstances. As academics continue to have difficulty providing their final post-print, the volume of refereed publications the workflow has to support has decreased. This has allowed appropriate project phase workflows to continue as before, has freed up resources to allow some degree of advocacy to continue, and will also allow the project to focus on supporting the development of a culture of dual submission – submitting a copy to the repository when submitting the article to the publisher. An implication of this change is that it will not be a primary aim of the service to provide a comprehensive retrospective listing of an academic's papers.

Another factor influencing the purpose of the service is that the departmental databases, which were helpful in populating the ePrints project phase, are generally perceived by academics as the primary bibliographic reference point. It is unlikely that departments would abandon local control over this data and it is these departmental-level services, rather than the ePrints service, which academics are using for the added value services – such as the creation of publication webpages, and monitoring of RAE submission – that have been perceived as incentives for participation in institutional repositories. This switch in emphasis has reinforced DAEDALUS' renewed focus on open access provision, a shift that has been

[5] Even publishers sympathetic to the open access movement generally now prefer that authors deposit post-prints and not publisher pdfs.

further encouraged by the impending mandate from the UK Research Councils that publications related to work they fund be available via open access.

5.2.2 Metadata quality

The reduction in the overall number of records being classified - brought about by the decommissioning of the metadata-only option - has helped ensure that it has been possible to maintain the quality of the metadata produced by the service. Moreover, as the service appears to have been well integrated with core library business and is perceived as part of the University's normal business, there is good reason to suppose that there will no need to compromise the current level of metadata quality in the foreseeable future.

5.2.3 Workflows

5.2.3.1 Departmental deposit and the individual mediated bulk deposit

Both of these workflows were key components for the bulk import of records in the project phase and contributed significantly towards attaining critical mass of records in the repository. With the post-project phase decommissioning of metadata-only inclusions, these workflows, which have tended to increase the proportion of metadata-only records, have been retired.

5.2.3.2 Individual mediated single-article deposit

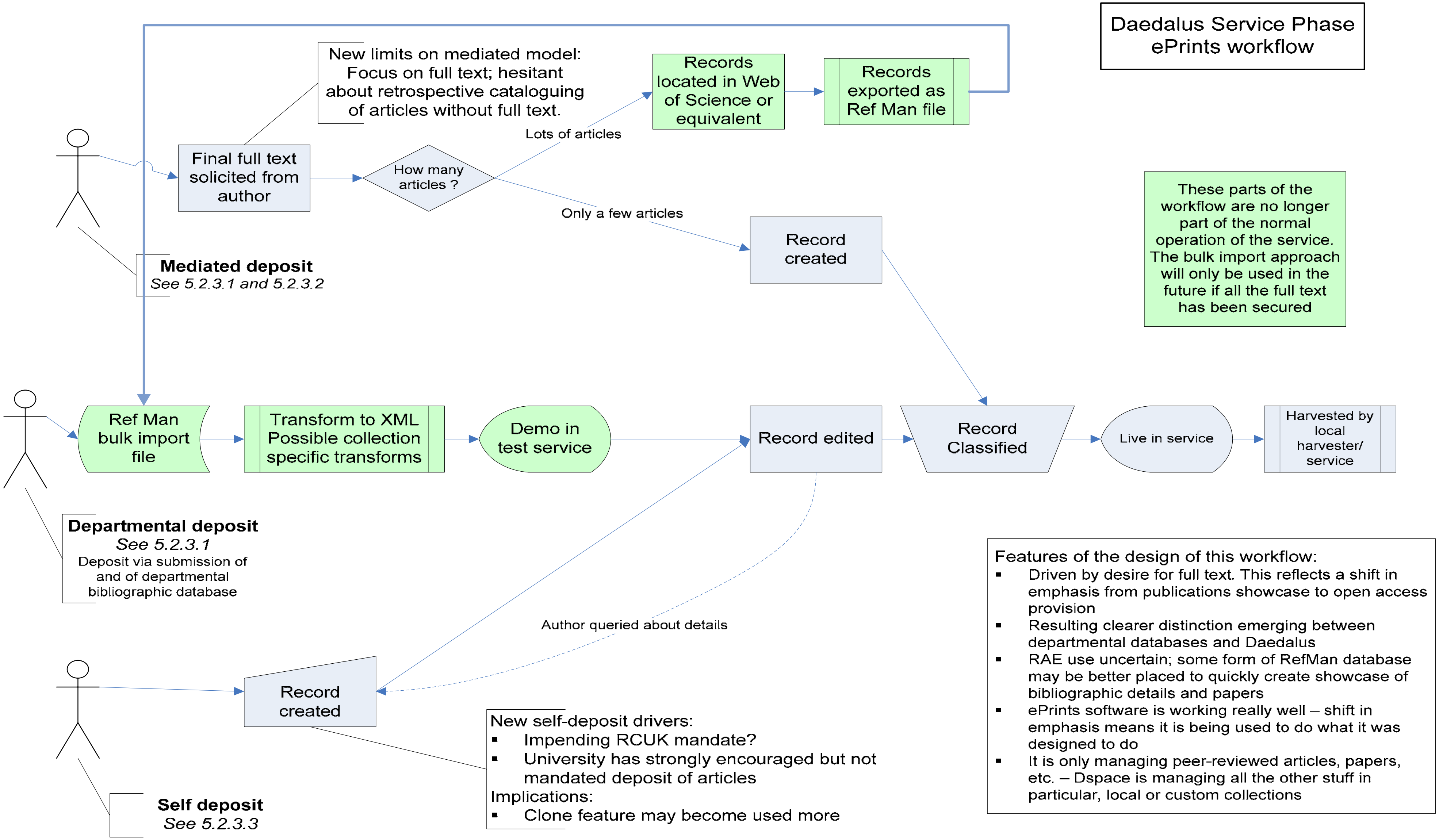
The individual mediated deposit is now the primary workflow of the ePrints service. It continues to be the simplest metadata creation process and to be closely tied to ongoing advocacy for the service within the University.

5.2.3.3 The self-deposit workflow

The self-deposit workflow may become increasingly important for the service. With the increasing prominence of open access, and with the possible Research Council mandate, academics are going to need to be increasingly concerned with where they publish and the value they place on participation in Glasgow's institutional repository is likely to increase.

As the use of the service increases, it is reasonable to hope that the quality of metadata produced by academics will improve, especially if they are made aware of the importance of metadata for retrieval via the advocacy programme and are also trained to produce high-quality metadata. One way of ensuring take-up of such training would be to address metadata creation in local bibliographic databases alongside metadata creation for deposit into the repository.

Unless a training programme is put in place and is successful, the increased emphasis on the self-deposit workflow is likely to result in a degradation of quality. More will be said about this in the analysis in section 7.


Daedalus Service Phase ePrints workflow
New limits on mediated model: Focus on full text; hesitant about retrospective cataloguing of articles without full text.
Records located in Web of Science or equivalent
Records exported as Ref Man file
Lots of articles
Final full text solicited from author
How many articles ?
Only a few articles
Record created
Mediated deposit
See 5.2.3.1 and 5.2.3.2
Ref Man bulk import file
Transform to XML Possible collection specific transforms
Demo in test service
Record edited
Record Classified
Live in service
Harvested by local harvester/ service
Departmental deposit
See 5.2.3.1
Deposit via submission of and of departmental bibliographic database
Author queried about details
Record created
Self deposit
See 5.2.3.3

These parts of the workflow are no longer part of the normal operation of the service. The bulk import approach will only be used in the future if all the full text has been secured

New self-deposit drivers:
- Impending RCUK mandate?
- University has strongly encouraged but not mandated deposit of articles

Implications:
- Clone feature may become used more

Features of the design of this workflow:
- Driven by desire for full text. This reflects a shift in emphasis from publications showcase to open access provision
- Resulting clearer distinction emerging between departmental databases and Daedalus
- RAE use uncertain; some form of RefMan database may be better placed to quickly create showcase of bibliographic details and papers
- ePrints software is working really well – shift in emphasis means it is being used to do what it was designed to do
- It is only managing peer-reviewed articles, papers, etc. – Dspace is managing all the other stuff in particular, local or custom collections

Figure 3 The DAEDALUS Service Phase ePrints workflow

5.3 DSpace at the University of Glasgow

5.3.1 Purpose

DSpace has been used by Glasgow to provide a software solution for the management of, and provision of access to, a variety of types of academic e-content. Its use has been explored within the project as a tool for the management of pre-prints, grey literature, doctoral theses, research finding aids, and university administrative documents. In essence, DSpace is being used for all the research and academic e-content not held in ePrints (i.e. content that isn't published and peer-reviewed).

Partly because of this wider focus, and partly because of other factors such as the significant externally-driven evolution of the DSpace software during the project, and the unexpectedly high level of advocacy required for the ePrints aspect of the project, the DSpace service is, as yet, less mature than the ePrints service, and its purpose less tightly defined.

From the outset, the DSpace repository was seen as performing a variety of functions, an aim for which it is well-suited, since it permits the creation of a series of individual community-based collection-specific workflows, templates and policies. It is assumed that different local communities (e.g. departments) will want to use DSpace to support a preservation service, a publication service, an asset management service, an institutional records service, and so on, but the extent to which it has been possible to explore this aspect of its use remains limited.

This does not mean that there are not generic elements of its intended function. There will be service-wide implications to supporting some of these various community or collection specific functions (for example, storing and supporting the retrieval of learning objects in a sub-collection may require the addition of new metadata fields across the entire repository). However, there is an extent to which elements of this underlying function will remain unclear until the process of evolving from a centrally-managed service (as set up in the project) into a more locally-managed one is more fully developed. It is intended that some of the community developments required for the purposes outlined above will be created and managed at community or sub-community level. Unlike ePrints which has a purpose defined tightly at institutional level, DSpace at Glasgow is envisioned as a tool for departments to use for a wide range of research content. One aspect of this goal is the intent to capitalise on DSpace's ability to devolve system administration tasks to particular communities.

5.3.2 Communities and collections

An illustration of the comparative diversity of purpose in Glasgow's DSpace service emerges from a consideration of the currently existing communities. These include:

- The DSpace community established for the Economics Department, which has a collection of economics working papers produced by the department and published on their website. These papers have been added into the DSpace by DSpace staff capturing the papers from the department and metadata from the Library catalogue. In addition links to the DSpace record will be added from the catalogue.
- The doctoral thesis communities. Several departments, including computer science, have established collections of doctoral theses. To support the management of theses, the Tapir plugin (developed as part of the Theses Alive! Project http://www.thesesalive.ac.uk/) has been installed. This presents a workflow specific to theses and has added metadata elements to DSpace to support specific thesis management requirements.
- A community related to a specific conference. The Department of Slavonic Studies hosts a collection of a papers from a conference they organised. Although these are not strictly outputs of the University of Glasgow, hosting them as part of the DSpace service enhances the available content and enables academics to support the development of their academic community.

5.3.3 Metadata quality

As the DSpace service is currently using an entirely mediated model, the records created for the system are of as high a quality as those created for the ePrints service. When the service develops a more devolved model of record creation, it is probable that this will affect the quality or completeness of metadata produced subsequently. Certainly it is a possibility that must be considered—alongside any effect this may have on higher-level services which utilise this data. Moreover, producing a metadata training programme that can cover the wide range of functions envisaged – an important means of minimising a deterioration in quality – will be much more difficult for DSpace than for ePrints.

5.3.4 Workflows

One aspect of the differences in software design between ePrints and DSpace is that DSpace can support a more complex workflow. DSpace supports the involvement of four different actors/agents in the creation of a metadata record (Submitter, Reviewer, Co-ordinator, Metadata Editor). Workflows involving these actors can be implemented on a collection by collection basis. However, since DAEDALUS' instantiation of DSpace currently provides an entirely mediated service all of the above roles are carried out by DAEDALUS staff.

The workflow for the collection of economics working papers capitalises on work already done by the Economics department. The records are created by copying information about the working papers from the catalogue, and uploading the files from the department's website. The DSpace service has also established a number of collections of doctoral theses within their respective departments. When students opt to make their theses available electronically, they submit an electronic copy to the service. The thesis' metadata is added by Bibliographic Services in the Library, who will have already catalogued the physical thesis (held in the Special Collections department).

The workflows established for the service thus far only directly involve repository staff, but with the service's intent to devolve at least some collection creation, control, and population to academic departments and units, some of the workflows will become embedded in the normal working practices of departments and the library.

One aspect of the establishment of these communities still under consideration is whether the administration of collections and communities will be devolved. This devolution of responsibility would allow localised sub-administrators to manage user workflows, identities and permissions. Establishing this would reduce the workload on repository staff in the long term but would require significant training in the short term and the number of people with the ability to change aspects of the service could affect its quality (both in terms of information retrieval or asset management effectiveness and its perception as a professional service).

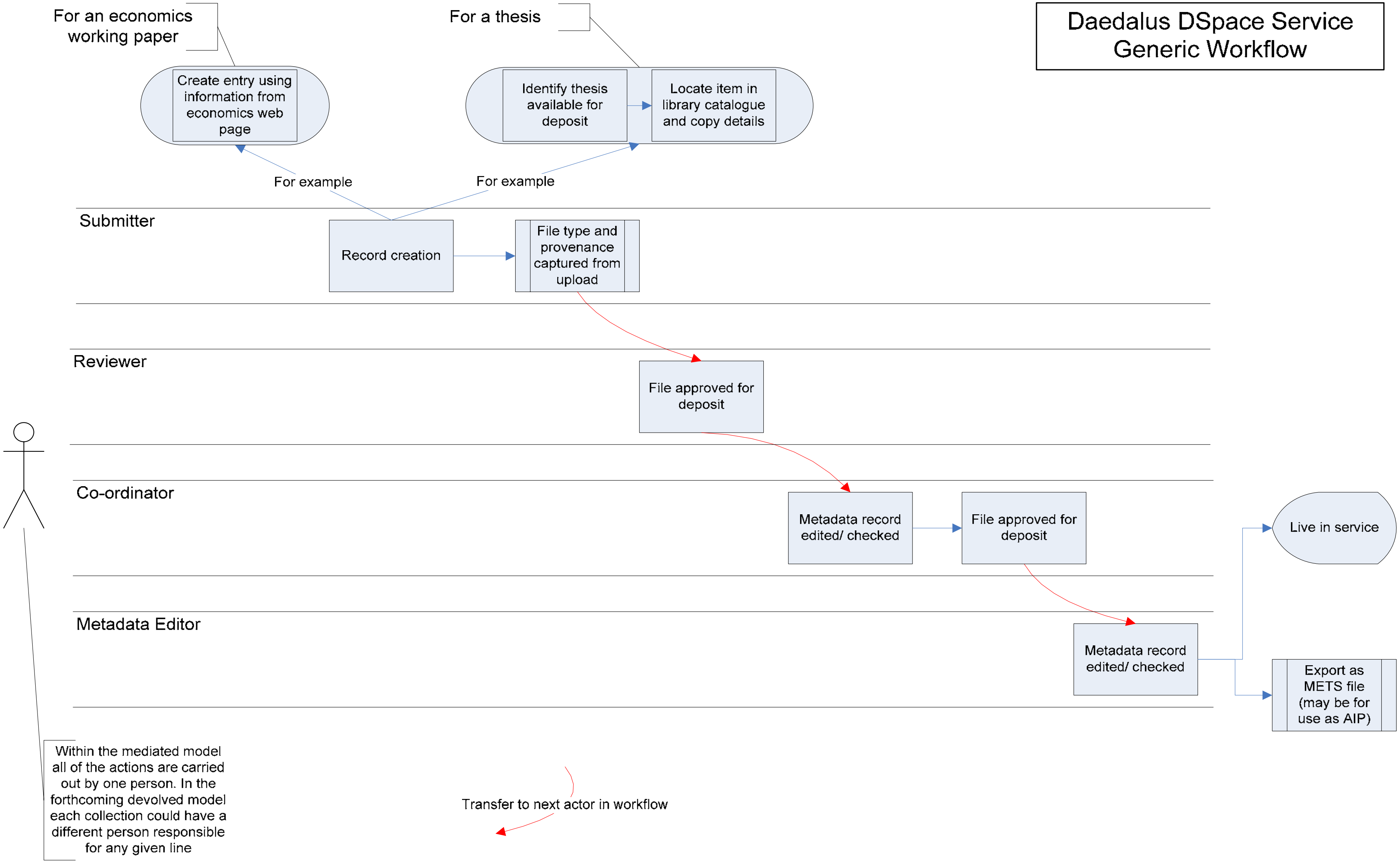

Daedalus DSpace Service Generic Workflow
For an economics working paper
Create entry using information from economics web page
For a thesis
Identify thesis available for deposit
Locate item in library catalogue and copy details
For example
For example
Submitter
Record creation
File type and provenance captured from upload
Reviewer
File approved for deposit
Co-ordinator
Metadata record edited/ checked
File approved for deposit
Live in service
Metadata Editor
Metadata record edited/ checked
Export as METS file (may be for use as AIP)
Within the mediated model all of the actions are carried out by one person. In the forthcoming devolved model each collection could have a different person responsible for any given line
Transfer to next actor in workflow

Figure 4 The DAEDALUS service DSpace workflow (generic version with examples)

6 Key findings from the current awareness review

6.1 Exploring the purpose of an institutional repository

The definition and purpose of institutional repositories is an issue of ongoing debate. One view sees an institutional repository as

> a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members. It is most essentially an organizational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organization and access or distribution (Lynch, 2003).

At the other end of the spectrum is a view which regards an institutional repository as solely concerned with the provision of open access versions of academics' published and peer-reviewed work. A repository's understanding of its purpose somewhere between these poles will influence which assets it chooses to store, what metadata it requires to support this, and may influence how it chooses to create that metadata.

6.2 Metadata Quality

The functionality a repository is able to support is not only dependent on the options provided by the software but also by the quality and nature of the metadata recorded about the objects it stores. Metadata quality – its fitness for the purpose of the repository – is not only difficult to define, but is often only noticed by its absence. Dushay and Hillmann (2003) suggest that poor metadata quality can be summarised as:

1. missing metadata (e.g. format and type fields empty);
2. wrong metadata (e.g. elements mixed up or meaningless entries);
3. confusing elements (e.g. multiple values in a single field, illegal characters, or bad markup copied into the field);
4. insufficient elements (e.g. sporadic element completion or variable use of qualifiers across the collection).

A synopsis of surveys and commentaries on what constitutes high quality metadata suggests the following collection of indicators: accuracy; timeliness; consistency (e.g. using terminology or notation in the same way across the metadata collection); intelligibility; completeness; currency; appropriateness; correctness; sufficiency; reliability; interoperability; persistence; verification (Greenberg et al. 2001; Moen, Stewart & McClure, 1997; Tozer, 1999; Rothenberg, 1996; Greenberg & Robertson, 2002; NISO Framework Advisory Group, 2004; Bruce & Hillmann, 2004).

When considering how such metadata should be created, elements which should be factored in include: scale, granularity, required functionality, required interoperability, staff information literacy (i.e. how staff search and how they create metadata), interface design, and metadata standard implementation choices (Barham, 2002; Crystal, 2003; Müller et al., 2003; National Information Standards Organization, 2004; McClelland et al., 2002).

6.3 Repository workflow models

An initial expectation, widely held within the FAIR Programme and beyond, was that self-archiving would be the predominant repository workflow model. There was also the assumption that unedited self-archiving would produce metadata of a reasonable quality and thus that populating ePrints archives would be relatively easy and inexpensive. Many of the early adopters of repositories, such as DAEDALUS, have found that even when possible within enthusiastic departments, it does not work across an institution.

Mediated models or self-archiving models enhanced with automated and/or manual checking are commonly accepted to be the norm[6]. Depending on the purpose of the repository there are a variety of ways in which a workflow can be formulated within such models. As articulated in the design framework for metadata workflow (figure 1, page 6), the creation and refinement of a workflow will be dependent on operational and strategic factors (for example, the purpose of the repository, resourcing available, software used, enthusiasm of staff). In most repositories, however, there will be:

- *A creator of the digital object*
- *A depositor of the digital object*
- *An owner of the digital object*
- *A specialist with information management knowledge*
- *A specialist with technical knowledge*
- *A repository manager/ editor*
- *A quality assurance specialist (with or without specialist knowledge)*
- *A preservation specialist*

[6] For example, even the institutional repository at Southampton has an editorial review of metadata before a record is made public. See Wendy White (2006) Embedding our Institutional Repository into the institutional research culture presentation at Institutional Repositories and Research Assessment (IRRA) Meeting 7th April 2006 Available: http://irra.eprints.org/bcsmeet/ ; Nixon and Greig (2005), also comment on this.

Within a given setting the appropriate division of labour for some of these roles will be self-evident; some of the roles will be quite limited (for example, the role of a preservation specialist may only be needed during the design of the repository) and more than one may be carried out by a given person (for example, the repository editor may also fill the roles of the specialists in information management and technical knowledge).

The role of resource creators in a metadata generation workflow is hotly debated (the tasks assigned to repository staff often depend on a prior decision about the role and relevant competencies of resource creators). Greenberg et al. (2001) suggest that, given training, creators can produce reasonable quality metadata and that they "may even be able to produce better quality metadata than professionals for selected elements" (for example, the date of creation and relation elements). Greenberg and Robertson (2002) take this further, concluding that trained authors are confident in creating most metadata but value input from information professionals when creating subject metadata. Their work complements, but is also in tension with, the findings of Barton, Currier and Hey (2003) who express reservations about the quality of solely author-generated metadata (also Bower, Sleasman & Almasy, 2004).

Projects that have been successful in implementing metadata workflows that include object creator contributions have tended to adopt either compulsion or a stealth approach in introducing the task (Barham, 2002). This is in part because current "academic measurement and reward systems are also not synchronised with the need for open access information environments” (Seaman, 2004 quoted in Campbell, Blinco & Mason, 2004). One possible incentive for academics to participate is in the use of the repository to promote their work, enhance collaboration, and generate researcher profile pages (Foster & Gibbons, 2005). When involving academics, it is important not to require highly-skilled individuals to perform basic tasks, and as a result workflows which can capitalise on the availability of a student or administrative workforce for some parts of the metadata task are strengthened (McClelland et al., 2002).

The role of the library in repository workflows has been discussed in further work by Greenberg (2004), which points out that although librarians are capable of creating or checking most of a record it is not efficient for them to do so. In a workflow they should be able to focus on evaluating subject headings and not spend time evaluating urls.

On the distributed nature of repositories and digital libraries another study notes that increasingly "libraries do not manage these digital resources, yet are being called upon to provide access to these collections", they now have a new potential role as trainers, consultants, and service providers, but will not have control of metadata creation [as they have with traditional cataloguing] (Mercer, 2003, p. 94). It is suggested that future digital library developments are likely to be collaborative efforts. It should be observed, however, that many libraries and library staff will have difficulty with the compromises inherent in producing metadata at a level of quality acceptable for a repository (El-Sherbini & Klim, 2004; Duval & Hodgins, 2004).

It is expected that auto-generation and auto-population of some metadata fields will increasingly result in more complete records. Auto-generation already happens in DSpace for the file format, and auto-population is possible through adding default fields to a community's metadata entry template. The auto-generation of subject terms is, however, nowhere near to being a complete and deployable solution, and subject identification is likely to involve human effort for a long time to come. The role of automatic techniques in creating efficiencies in repository workflows is important since being able to complete obvious and less crucial fields more easily frees resources to implement subject headings and other elements which benefit from human intervention. A related issue is that batch processes of importing records or augmenting them is increasingly proving to be an efficient and effective way to manipulate records sets and improve their metadata quality (Currier et al., 2004; Hillman, Dushay, Phipps, 2004)

6.3.1 Examples of repository workflows:

The DiVA project highlights the benefits of workflows for reducing duplication of labour and for increasing the scalability and speed of depositing objects into a repository. The DiVA system adds templates to common word processors which allow not only the creation of a title page of the thesis, but also pull a unique id for the document from the Swedish National Library (the Royal Library in Stockholm), and create an XML metadata record which is attached to the document and then deposited with the electronic copy in the University library and the National Library. The metadata record is reviewed prior to 'publication' in each of the libraries. This division of labour facilitates a much more rapid deposit of materials into the libraries and, as it reduces the burden on each agent, is a more scalable model (Müller et al., 2003).

The Internet Scout project highlights a multi-agent workflow; the project originally used librarians to create all its records but this proved too expensive to maintain. Consequently it created a workflow using "a different staffing configuration: one professional librarian supported by one or more library and information science graduate students. The librarian defines policy and oversees quality assurance, while the students get hands-on experience and education” (Bower, Sleasman & Almasy, 2004, p. 168).

6.4 Interoperability /Interaction with other services

In the development of metadata workflows one other major consideration about the quality of metadata or about transformations is to think about where else the metadata will go or be used. Although fitness for a local purpose will be the baseline for metadata quality, some consideration of its fitness for external purposes should be made if the service intends for its metadata to be interoperable with, and be exposed for use by, external services.

This could affect a repositories metadata in two ways. It could mandate the use of content standards throughout the repository; for example, the use of a recognised subject scheme which would permit subject searching and browsing across distinct record sets. It could also address what changes might be made to the metadata before it is exposed for harvest (for example, adding a rights statement to all locally produced resources).

Connected to this is an awareness that the perceived value of external aggregator services may suffer if any of the harvested repositories have poor metadata. This would affect the service provided to the end user and, have significant implications for how seriously the open access movement or the Open Archives Initiative is taken, since the value of this movement and this protocol will be judged by the service an end user receives.

7 Analysis and discussion

7.1 Purpose, requirements and priorities

7.1.1 Changing contexts

In the growing diversity of opinion over what an institutional repository is and what it does, DAEDALUS has addressed the issue by developing a two repository approach. It has one repository which is an institutional open access provider of published peer-reviewed material and one repository of other materials. It has selected and implemented different software, each developed to suit one of the above purposes, and is in the process of developing distinctive workflows for each of the repositories. This division of labour represents one of the strengths of the project as it allows the service to make different decisions about metadata or workflow for different types of content. In the ePrints service, providing open access to Glasgow's academic papers, the library is creating a resource with long term potential and is investing in high-quality metadata. In the DSpace service providing access to a wider range of materials, the implementers of particular collections can make decisions about the quality of metadata required to fulfil their needs – thus, for example, not requiring transitory material to have such a significant cataloguing overhead. The split also gives both services the option of providing search features and supporting metadata specific to the types of material they hold.

At the end of the project phase it was concluded that:

> [the project's] experiences have led to an expectation that the future of the ePrints service lies in a centralised model whereby the repository is populated via regular imports either from a central internal publications database or from departmental or faculty databases. It is interesting to note that other institutions, e.g. the Universities of Southampton and Durham have also adopted a centralised approach. The adoption of a centralised model is also important for the relationship between repositories and the RAE 2008 [DAEDALUS final report].

As described in an section 5.2.1, it has subsequently emerged that ePrints is unlikely to be used in this way for the RAE, and that this sort of bulk upload will be removed from normal use as the ePrints service strives to move away from metadata-only records (it will still be available if an academic or department is able to provide the full-text article for those records). That the service has been able to clearly reformulate its intent, and reassess it workflows, in only a matter of months (in response to both internal and external changes) is an indicator of the degree to which the staff are aware of the implications of changes in context and purpose on the service's workflow. This staff awareness bodes well for the continuing development of both services, in particular the upcoming expansion and devolution of the DSpace service, and the development of targeted advocacy for each service.

7.1.2 Workflow design

The workflows developed during the project phase of ePrints and DSpace have largely been able to transfer into the service delivery phase. Both repository services are providing high quality metadata created with a distributed skill set utilising both project and, directly or indirectly, Bibliographic Services staff.[7] The changes that have occurred in the workflows, such as the removal of the bulk upload option, have been driven by, and reflect changes in, the purpose of the repository services rather than workflow refinements as such.

The securing of funding to underwrite the continued use of Bibliographic Services staff in repository services shows the degree to which the services, and by implication the workflows, are becoming embedded in the library and the University's thinking. This integration should support future development of the workflows as part of the core library services . The workflows' use of trained cataloguers improves efficiency at university level. They will be able to classify assets more quickly and effectively than other staff – cataloguers need less time and cost less per record than professors do.

7.1.3 Identifying points of paper capture for the workflow

With the shift in focus in the ePrints service to only creating records for papers when an open access copy is available, the service is considering how it can best support academics in supplying those copies. One of the objectives of this evaluation is to suggest the key workflow points at which academics might deposit their paper or create metadata.

There are a number of key points in the production of an article when an academic might logically provide the ePrints repository with a copy of their paper:

1. when an article is submitted to a journal;
2. when editorial or peer-review changes have been made;
3. when the journal is published electronically;
4. when they receive their paper copy of the journal (if applicable);

[7] This section only addresses current workflows; for issues in the development of workflows that are likely to take place as a result of the devolution of DSpace see section 7.2.3

5. when they enter the paper's details into their departmental bibliographic management system (if applicable).[8]

Given the local requirement that only post-prints are included in the ePrints service the first of these stages, submission to the service coincidental with submission to the publisher is not a possibility.[9] It is also unlikely that picking one of these other points (2-5) and encouraging deposit only at that point would be effective. Submitting the paper to the repository alongside resubmission to the journal, noticing the electronic publication, or receipt of paper copy all require the promotion and formation of habits. Promoting these as triggers will have to become part of the wider training and dissemination carried out by the repository. The case of paper submission alongside maintaining departmental bibliographic databases will also largely be dependent on habit, but it does, in theory, permit some degree of automated support for the process; for example, it may be possible to create a plugin or macro for the bibliographic software that supports emailing the repository a selected metadata record and opens a browse window to attach the postprint.

7.1.4 Issues in metadata quality

The metadata in both repository services is produced with care and using a combination of trained administrative and professional staff. Thus high metadata quality is maintained. This is particularly important for the ePrints service, as it is providing an ongoing resource of known value (i.e. published peer-reviewed journal articles are a resource that will be of value in the long term). As such, the investment in cataloguing and classifying these records is comparable to that for other library resources. The current level of metadata quality in the DSpace service is of similar quality to the ePrints service but this may need to be reassessed as the service develops and as control of the collections changes. As DSpace is used for a variety of purposes and a range of types of digital assets with different lifecycles, the level of metadata quality the service can justify is likely to diverge from the ePrints service's level of metadata quality.

Within the ePrints service there are some known issues in the quality of the metadata, but these reflect issues the entire repository community is attempting to resolve. The first relates to a conceptual question over what is actually being catalogued – the author's post-print or the published article. Other projects, such as SHERPA and VERSIONS

[8] Whichever of these options an academic or administrator uses they can do so with either the mediated or self-deposit workflow model.

[9] Although a copy could be submitted to the DSpace service at this point.

(http://www.lse.ac.uk/library/versions/), are addressing these concerns about how cataloguing and citation rules should be applied; for example, in assigning the place of publication or quoting a page number. One approach to resolving this problem has been to add cover sheets with basic information about each version of the article in the pdf that is created from the post print.

The second known issue relates to author browsing. In ePrints any author browse facility or name authority control within the repository is currently, by design, tied into the person who deposited the ePrint. As a result, the very strategy that led to the successful population of the service through a mediated model and via bulk uploads of metadata, created a difficulty. All item records created in such a manner are associated with the login of the original depositor who may not be one of the authors. With the re-emphasis on the self-deposit workflow this would have led to a tension between old and new records. This situation has been resolved at a community level by changes to the ePrints software which allows the creation of an author identifier in addition to the depositor identifier/ login.[10]

In the future development of the ePrints service there will be a corresponding increase in workload for the metadata editors as more authors use the self-deposit workflow. This increase in workload will be offset to some degree by improving users' information literacy skills through training.

As metadata creation and community management is devolved in DSpace there will inevitably be a corresponding change in the collection's metadata quality. The metadata will become more variable as a consequence of having both more metadata committers (i.e. those who can approve a record for the database) and of having metadata committers and creators who are not trained cataloguers/ information specialists. This is an inevitable development of the change in purpose. One significant implication of this is that as staff strive to ensure that the DSpace metadata is good (i.e. fit for purpose) there will be and should be a difference in the 'objective' metadata quality between ePrints and DSpace. The challenge for the DSpace repository is to ensure that the metadata for its purpose(s) is as good as it can be.

[10] This change to the EPrints.org software is a local development currently only at Southampton. It is, however, reasonable to suggest that, in time, the developers will include this in a EPrints.org release.

With the variety of purposes that are emerging for DSpace the metadata set it creates and holds is likely to have elements that are not completed across the entire set (e.g. thesis supervisor). The challenge then raised is ensuring completeness across the metadata-record subset for which any given element should be present. Conversely as different sub-communities use the system, consistency will become a challenge as they may seek to use the same metadata element in different ways. Sub-community administrators and trainers will need to remind users that their metadata is part of a larger collection, and that, as far as possible, the super community's use of an element should be borne in mind. An example of this may be in the use of the type element. Some communities may use this element at the same level of granularity as the super-collection and be able to use the same controlled vocabulary; whereas other sub-communities may only deal with a few of the super-collections types and require a more granular controlled vocabulary (e.g. super-collection element value: text; sub-collection element values: thesis, working paper, pre-print). One possible solution to this would be for the sub-collection to store two values for the given element – both one for local needs and one for the super-collections needs.

Another minor note in the consideration of metadata quality is that neither piece of software has yet integrated the sorts of tools to support cataloguing and classification that are found in library catalogues (e.g. automatic authority control and live indexing). It is likely that some of these tools may be introduced in longer-term future iterations of these products and that this will improve the overall quality of records produced.

7.1.5 Training and support

The repository services face challenges in training both academic and library staff. Training may involve classes or it may be the promotion of good habits and helpful concepts in the general dissemination of the services.

Academic staff require training in deposit and in the habit of retaining and managing their post-prints. For the ePrints service, the challenge remains in academics being able to identify the correct version of their work. For example, if they continue to use the mediated model this may involve repository staff promoting the idea of cc'ing the repository when they receive the printed copy of their article. If they use the self-deposit approach it may involve more detailed information-handling training (just as provided for bibliographic software). Support for academics is likely to also include information on the open access policies of key journals and guidance on general file management to alleviate version control difficulties.

Library staff in general will require training about the repository services so that in any face – to-face contact with users they understand the services and can answer basic queries. Bibliographic Services staff have thus far dealt primarily with metadata-only records. Training thus far appears to have been successful but in the day-to-day running of the services issues have arisen that are likely to require incorporation into future training including, for example, shortening session-specific urls to ensure that an alternate url element will continue to point to a meaningful location or understanding differences in metadata quality between traditional library services and repository services and between the standards they use.

A further area of training concerns the upcoming devolution of DSpace community and collection management and administration. This will require not only the identification of departments or units to be key early implementers, but is likely to become the dominant focus of the repository's work for a period. Training will require the extensive articulation of implicit assumptions and working practices built up in the project's experience of the past three years. As some administration is gradually devolved there will be an ongoing need for guidance and discussion with departments as they think about collection-specific metadata creation workflows – especially if they are expecting to interact with repository staff or Bibliographic Services.

7.2 Issues for the future development of the repository services

7.2.1 Sustainability

The future development and sustainability of the repository services has to some degree been secured by their integration with existing workflows and structures and their incorporation into the University's public persona. The degree to which this sustainability can be translated into long-term sustainability is dependent not only on continued advocacy and service provision, but is also likely to be influenced by the degree to which the project is able to capitalise on possible efficiencies (through integration with other library or external services) and maintain the high visibility of the repositories content in a diverse range of external services (such as Google) – are the repository services value for money and do they provide good access to Glasgow's assets? Efficiencies may arise through workflow integration with internal services (for example, the network's directory services) or through workflow interaction with external services (for example, registries or access services).[11]

7.2.2 Changes in software

As noted, one of the strengths of the project is its ability to separate out different repository functions and apply appropriate software to support them. As software options are refined and improved and as new products emerge, two issues will have to be managed. The first is that as new software products emerge for particular purposes their use for those purposes should be considered (for example, separating out learning objects from the DSpace installation, or using Open Journal System for e-publications). The second is that, in order to maintain the efficient generation of high quality metadata, function creep should be avoided –functions added to new iterations of the software in use should only be implemented in, and supported by the metadata and the workflow if they are central to the service's purpose.

7.2.3 DSpace Workflows

With the proposed partial devolution of the administration of the DSpace service the workflows of each collection are likely to begin to vary considerably. Any given collection may involve any combination of: academics, administrators, repository staff, and library staff – all utilising varying skill sets in any given instantiation – as it attempts to produce metadata to suit its particular purpose. Not only will the design of these workflows require a significant degree of input from repository staff, but they will also produce a range of metadata quality (reflective of the requirements of the range of purposes and lifespan of some of the assets). This devolution poses a degree of risk to the DSpace service as its success will rely on the successful integration of metadata workflows into departmental practices and priorities. Its success will also require a degree of integration between departmental and library workflows. The project will have to accommodate the support of these workflows and address the potential impact of divergence in metadata practices within the repository on the services it offers (for example, the use of different metadata elements or different vocabularies). It should be noted that the project already does this successfully for theses records.

7.2.4 Technical and Administrative metadata

It is likely that in the foreseeable future the repository services will want to ensure that they store (or can generate on demand) technical and administrative metadata for their digital assets. This already happens to a limited degree, but as the curation and archiving of digital

[11] Sustainability has also been discussed in the previous sections on Workflow design (7.1.2) and on Metadata quality (7.1.4).

assets begins to become a community-wide focus and as more assets are made available to externally provided web services, the need for such metadata will become more important. The specifics of such technical and administrative metadata are as yet unknown at both local and community levels. It is currently unclear precisely what such metadata would be used for, how it should be captured or represented, and to what degree such information might be supported by the repository community. Possible requirements arise, for example, from supporting curation and preservation (which require, amongst other things, administrative information to support tracing the provenance or integrity of an item and technical information to support future usability) or from supporting Web services (which might use technical data to choose what assets to offer to a given user).

The positive side of this increased demand for administrative and technical metadata is that some of it should, in theory, be able to be created automatically. The metadata likely to be required to meet technical and administrative requirements is often derivative of known values and existing practice. For example, from the user login and system values, repositories should be able to tell and record who is doing what to a particular asset and when. From the asset format (mime type) repositories should be able to deduce the system requirements for presenting the asset to users. It is expected that such features will become a routine part of future evolutions or developments of repository software.[12]

Before such developments are complete and commonplace, however, steps can be taken to support the addition of some types of technical metadata. As local repositories may not require the storage of such metadata for their purposes, it is suggested that such metadata should only be added into a record when it is being exported (whether for external or archival purposes). As indicated some technical metadata is dependent on key existing metadata, so it can be added to records *en masse* using a script. It should be noted, however, that some types of administrative and technical metadata can *only* be captured when the object is created (for example, the camera angle or exposure at which an image has been taken). It is not known if such information will, as a rule, be necessary, but it should be noted that capturing metadata of this type probably requires a substantive increase in effort on the part of the asset creator.

[12] Some of this functionality is already present in DSpace.

In the context of DAEDALUS' services technical and administrative metadata requirements- beyond those that occur as part of the software in use - are, at this stage, largely unknown and consequently little can be said about workflows to create them. As community-wide services begin to emerge, however, it is possible to consider how some such metadata could be created. *Dependent* technical metadata (e.g. deriving minimum system requirements from the asset's mime type) could fit into existing workflows. How this might fit in with the workflow is illustrated in figure 5; after the record is edited and made live, a copy is transformed to be stored as part of an OAIS Archival Information Package or an on-demand export package. Additional administrative metadata could also be added manually or via a template at this stage (for example, rights metadata (if absent)).[13]

It is important to stress, however, that without a full analysis of administrative and technical metadata requirements in the context of known and stated policies on issues such as preservation, it is unsafe to make too many specific recommendations as to how best to discuss associated metadata workflows.

7.2.5 Intellectual Property Rights (IPR)

Although the project has extensively addressed rights issues in the ePrints service, they have not yet significantly occurred in the DSpace service. When the service begins to accommodate some types of digital assets such as learning objects, rights clearance will occur as an ongoing issue and will have to be addressed in some form. Managing rights and educating about the appropriate re-use of materials might naturally be added the responsibilities of the information management specialist in the workflow (as occurred for such issues in the ePrints service).

7.2.6 External requirements

Another issue which may effect the repository services is the upcoming change in research funding within the United Kingdom. Such changes may include funding council requirements to deposit papers or data in a designated external repository or to include details of the funding award in a local repository. The implication of this is that university repositories should interoperate with funder's repositories. In practice this implies that local repositories may have to import from or export to external repositories – a feature supported by repository software. The challenge, however, when dealing with a known exchange of records is that

[13] The underlying database software is likely to record such information at some level. However, if the repository does not explicitly utilise this metadata accessing this information may require technical support.

local and external metadata practices may diverge. To support the best possible interoperability of metadata, records harvested or exported should automatically undergo appropriate simple metadata transformations to fit in with the practice of the importing repository. Possible examples of transformations include changing from 2 to 3 letter language codes, mapping to simpler asset types (thesis to text), or adding institutional or rights data.

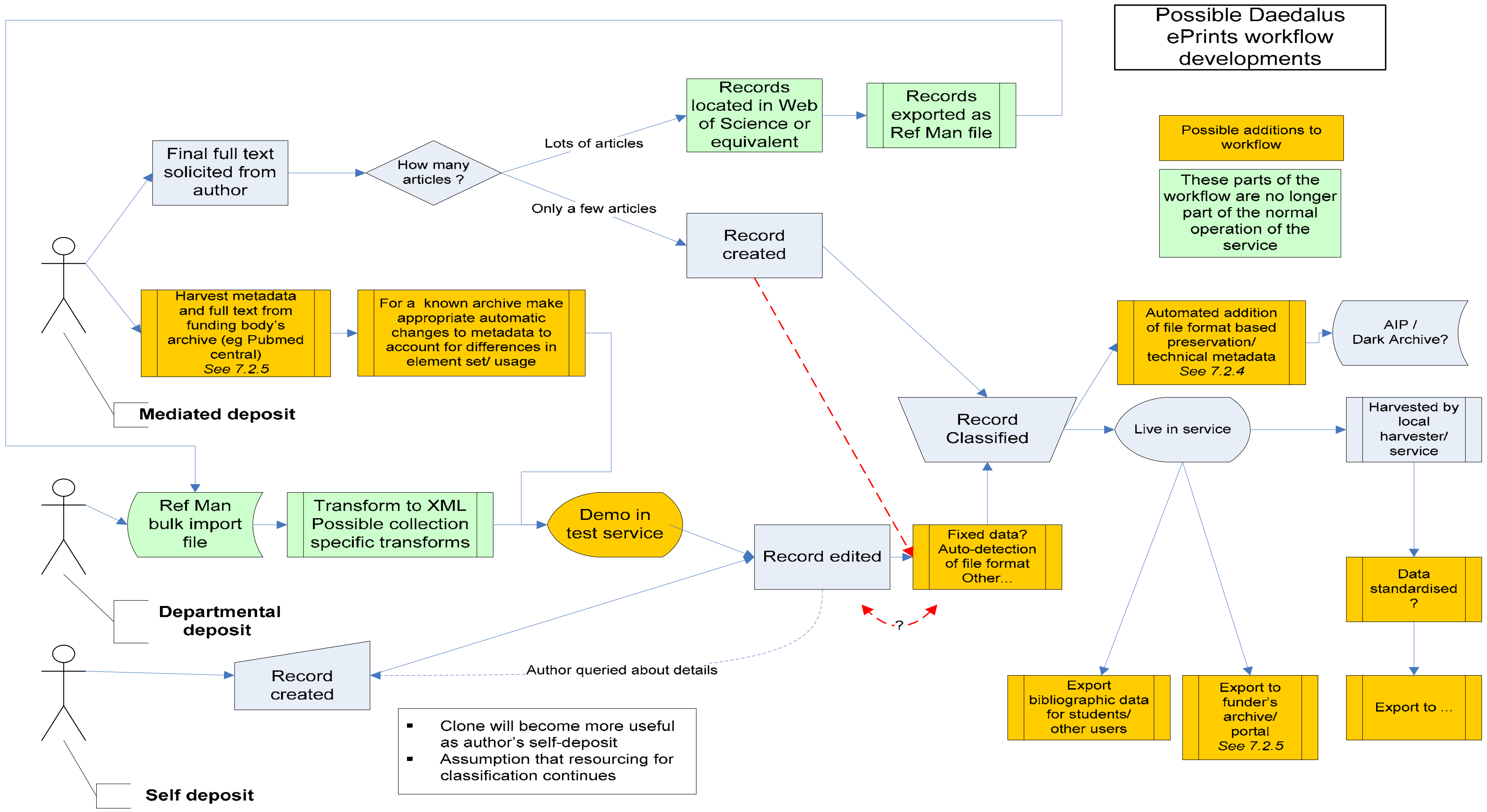

Possible Daedalus ePrints workflow developments
Possible additions to workflow
These parts of the workflow are no longer part of the normal operation of the service
Final full text solicited from author
How many articles ?
Lots of articles
Only a few articles
Records located in Web of Science or equivalent
Records exported as Ref Man file
Record created
Harvest metadata and full text from funding body's archive (eg Pubmed central) See 7.2.5
For a known archive make appropriate automatic changes to metadata to account for differences in element set/ usage
Mediated deposit
Automated addition of file format based preservation/ technical metadata See 7.2.4
AIP / Dark Archive?
Record Classified
Live in service
Harvested by local harvester/ service
Ref Man bulk import file
Transform to XML Possible collection specific transforms
Demo in test service
Record edited
Fixed data? Auto-detection of file format Other…
?
Departmental deposit
Data standardised ?
Record created
Author queried about details
Self deposit
Clone will become more useful as author's self-deposit
Assumption that resourcing for classification continues
Export bibliographic data for students/ other users
Export to funder's archive/ portal See 7.2.5
Export to …

Figure 5 Possible DAEDALUS ePrints workflow developments

8 Conclusions and recommendations

The DAEDALUS repository services at the University of Glasgow have, thus far, successfully navigated the transition from project to operational service. The repositories' purposes and metadata workflows have adjusted to external and internal changes and continue to support a high level of metadata quality. The workflows are likely to undergo further significant changes as the DSpace service devolves some community management responsibilities, but those issues which are likely to arise will have already been encountered during the evolution of the ePrints service.

Section 7 explored the issues that the repository services at the University of Glasgow are likely to face in their ongoing development; this concluding section will build on that analysis to make recommendations on metadata workflow issues in the continuing service

8.1 Recommendations

As it continues to develop, DAEDALUS should maintain its clarity of purpose, monitor its metadata quality in individual and aggregated services, support metadata enhancement, and continue to educate and support users.

8.1.1 Clarity of Purpose

The repository workflows have successfully adjusted to changes in their operating environment because of the clear division of labour they have established. Even when there have been shifts in the purpose of a repository, they have been clearly understood and articulated in the context of the service as a whole. This clarity enables efficient and effective metadata workflows and it is important it be maintained. Significant *ad hoc* changes to system function and focus should be avoided. Major changes should be agreed, planned, and introduced in a controlled and phased fashion. With this in mind, it is recommended that DAEDALUS should:

- monitor user needs and review institutional requirements on a regular basis
- monitor changes in software functionality and any effect implementing these may have on the use of the system;
- consider expanding either the portfolio of deployed software or increasing instantiations of particular systems to support any emerging distinct purposes;

- clarify the relationship of the ePrints open access service with any emerging services such as a co-ordination of bibliographic software or University of Glasgow research showcase.

8.1.2 Metadata Quality

The services have developed workflows that create high quality metadata. As author self-deposit increases in both the ePrints and DSpace services, it will become important to review and monitor metadata quality. This can be done by putting in place procedures to:

- carry out random checking of records;
- periodically check repository indices;
- export metadata for analysis in other software (e.g. import it into a spreadsheet to provide a graphical overview of completeness, or an indication of term frequency).

8.1.3 Metadata Enhancement

DAEDALUS should consider adding additional workflow elements to support metadata enhancement. In particular, the project should aim to :

- ensure that metadata recording IPR is entered manually or automatically for locally created resources (especially in DSpace collections controlled at the sub-community level);
- monitor developments in the use of and support for administrative and technical metadata – in particular the outputs of the Digital Curation Centre (DCC) and PREMIS working group;
- identify and utilise reference sources to provide authoritative representation information for the addition of technical metadata to records (for example, the registry of representation information under development by the DCC);
- put in place scripts to carry out appropriate metadata enhancements between repositories and aggregator services (for example, the transformation of a local use of the type element to a more generic label (e.g. thesis to text));
- develop scripts to extract information about assets (in particular for DSpace content) provided by a particular academic to support the generation of a researcher profile for them;
- assess the necessity for, and possibility of, such transformations between sub-collections in DSpace (if different sub-communities develop necessarily distinct

metadata practices higher-level repository services may need to adjust this metadata on the fly).

8.1.4 Education and User Support

In the context of the ongoing promotion of the repository services and expected devolution of sub-community control in DSpace to departmental-level administrators it is anticipated that staff in repository services and other parts of the library will offer various forms of training in support of existing and emerging workflows. Such training should include:

- a brief introduction to metadata (covering its role in providing points of access and greater precision as well as how term selection is affected by context);
- an introduction to issues and implications of IPR and Open Access;
- the promotion of 'trigger points' – to help academics remember to deposit their publications and other materials in the repositories.

9 Bibliography

9.1 DAEDALUS documentation and publications

Drysdale, Lesley, (2004) *Importing Records from Reference Manager into GNU Eprints*. Available: http://hdl.handle.net/1905/175

Greig, Morag and William J. Nixon, (2005) *Daedalus Final Report*. Available: **http://hdl.handle.net/1905/510**

Nixon, William J. and Morag Greig, (2005) *Populating the Glasgow ePrints Service: A Mediated Model and Workflow*. Available: http://hdl.handle.net/1905/387

9.2 Other resources

Barham, S. (2002). New Zealand Government Implementation of a DC-based Standard - Lessons Learned, Future Issues. In: *Proceedings of International Conference on Dublin Core and Metadata for e-Communities*. Florence, Italy: Firenze University Press, p171-176. Available from: http://www.bncf.net/dc2002/program/ft/paper20.pdf

Barton, J., Currier, S. and Hey, J.M.N. (2003). Building quality assurance into metadata creation: an analysis based on the learning objects and e-prints communities of practice. In: *DC-2003: 2003 Dublin Core Conference*, Seattle. Available from: http://www.siderean.com/dc2003/201_paper60.pdf

Bower, R., Sleasman, D. and Almasy, E. (2004). The Internet Scout Project's Metadata Management Experience: Research, Solutions, and Knowledge. In: Hillmann, D.I. and Westbrooks, E.L. (eds). *Metadata in Practice*. Chicago: American Library Association, p158-173.

Bruce, T.R. and Hillmann, D.I. (2004). The Continuum of Metadata Quality: Defining, Expressing, Exploiting. In: Hillmann, D.I. and Westbrooks, E.L. (eds). *Metadata in Practice*. Chicago: American Library Association, p238-256.

Campbell, L.M., Blinco, K. and Mason, J. (2004). *Repository Management and Implementation: A White Paper for alt-i-lab 2004. Prepared on behalf of DEST (Australia) and JISC-CETIS (UK)*. Available from: http://www.jisc.ac.uk/uploaded_documents/Altilab04-repositories.pdf .

Crystal, A. (2003). Interface Design for Metadata Creation. In: *Extended Abstracts of the Conference on Human Factors in Computing Systems (CHI 2003).* Fort Lauderdale, Florida: ACM Press, p1038-1039. Available from: http://doi.acm.org/10.1145/765891.766136.

Currier, S., Barton, J., O'Beirne, R. and Ryan, B. (2004). Quality Assurance for Digital Learning Object Repositories: Issues for the Metadata Creation Process. *Alt-J*. 12 (1), p5-20.

Dushay, N. & Hillmann, D.I. (2003). 'Analyzing Metadata for Effective Use and Re-Use'. DC-2003: 2003 Dublin Core Conference, Seattle Available: http://www.siderean.com/dc2003/501_Paper24.pdf

Duval, E. and Hodgins, W. (2004). Making Metadata go away: "Hiding everything but the benefits". In: *DC-2004: International Conference on Dublin Core and Metadata Applications*, Shanghai, China. Available from: http://students.washington.edu/jtennis/dcconf/Paper_15.pdf.

El-Sherbini, M. and Klim, G. (2004). Metadata and cataloging practices. *The Electronic Library*. 22 (3). p238-248.

Foster N. F. & Gibbons, S. (2005), 'Understanding Faculty to Improve Content Recruitment for Institutional Repositories' D-Lib Magazine Vol. 1 no. 11. http://www.dlib.org/dlib/january05/foster/01foster.html

Greenberg, J. (2004). *ALCTS CCS- Cataloging & Classification Research Discussion Group: Optimizing Metadata Generation Practices*. American Library Association. Available from: http://ils.unc.edu/mrc/amega_ccrd.htm

Greenberg, J., Pattuelli, M.C., Parsia, B. and Robertson, D.W. (2001). Author-generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organization. Journal of Digital Information. 2(2). Available from: http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Greenberg/

Greenberg, J. & Robertson, D.W. (2002). Semantic Web Construction: An Inquiry of Author's Views on Collaborative Metadata Generation. In: *Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002*. Florence, Italy: Firenze University Press, p45-52. Available from: http://www.bncf.net/dc2002/program/ft/paper5.pdf.

Hillmann, D.I., Dushay, N. & Phipps, J. (2004). 'Improving Metadata Quality: Augmentation and Recombination'. DC-2004: International Conference on Dublin Core and Metadata Applications, Shanghai, China. Available: http://students.washington.edu/jtennis/dcconf/Paper_21.pdf

Lynch, C.A. (2003). 'Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age'. ARL Bimonthly Report, no. 226, Available: http://www.arl.org/newsltr/226/ir.html

McClelland, M., McArthur, D., Giersch, S. and Geisler, G. (2002). Challenges for Service Providers When Importing Metadata in Digital Libraries. *d-Lib Magazine*. 8(4). Available from: http://www.dlib.org/dlib/april02/mcclelland/04mcclelland.html.

Mercer, H. (2003). Striking a Balance: Metadata Creation in Digital Library Projects. In: Ury, C.J. and Baudino, F. (eds). *Brick and Click Libraries: The Shape of Tomorrow, Proceedings of a Regional Library Symposium.* Maryville, Missouri: Northwest Missouri State University, p94-99. Available from: https://kuscholarworks.ku.edu/dspace/bitstream/1808/140/1/mercerBCL03.pdf

Moen, W.E., Stewart, E.L. and McClure, C.R. (1997). *The Role of Content Analysis in Evaluating Metadata for the US Government Information Locator Service (GILS): results from an exploratory study*. Available from: http://www.unt.edu/wmoen/publications/GILSMDContentAnalysis.htm

Müller, E., Andersson, S., Klossa, U. and Hansson, P. (2003). Metadata Workflow based on Reuse of Original Data. In: *Proceedings of the Sixth International Symposium on Electronic Theses and Dissertations*, Berlin. p53-57. Available from: http://edoc.hu-berlin.de/etd2003/andersson-stefan/PDF/Andersson.pdf.

National Information Standards Organization. (2004). *Understanding Metadata*. NISO Press. Available from: http://www.niso.org/standards/resources/UnderstandingMetadata.pdf.

NISO Framework Advisory Group. (2004). *A Framework of Guidance for Building Good Digital Collections*. 2nd ed. Bethesda, MD: National Information Standards Organization. Available from: http://www.niso.org/framework/framework2.html.

Rothenberg, J. (1996). Metadata to Support Data Quality and Longevity. In: *1st IEEE Metadata Conference*, Silver Spring, Maryland. Available from: http://www.computer.org/conferences/meta96/rothenberg_Paper/ieee.data-quality.html.

Seaman, D. (2004). *Institutional Repositories, presentation to JISC Joint Programmes Meeting*. Available from: http://www.ukoln.ac.uk/events/jisc-jpm/programme.html

Tozer, G. (1999). *Metadata Management for Information Control and Business Success*. Boston: Artech House.