7

Case study: The Edinburgh research archive

Introduction[1]

Theses alive: a project history

Until relatively recently, there has been minimal interest from the UK in e-theses, and a very select few institutions have been developing these capacities. To encourage the disclosure and sharing of content, the Joint Information Systems Committee (JISC) initiated the Focus on Access to Institutional Resources (FAIR) programme in late 2002. The purpose of this programme was to investigate the sharing of digital institutional assets, including e-theses, and to gather intelligence about and increase understanding of the technical, organisational and cultural challenges of these processes. Under this programme Edinburgh University Library (EUL) obtained funding for the Theses Alive project which began to work on a prototype for a national e-theses promotion and management concept at the end of 2002. This project worked alongside related projects Data-providers for Academic E-content and the Disclosure of Assets for Learning, Understanding and Scholarship (DAEDALUS), based at Glasgow University Library, and Electronic Theses, based at the Robert Gordon University Library. At the same time EUL was involved in the SHERPA project led by the University of Nottingham, which was primarily concerned with the creation, population and management of several e-print repositories in partner institutions in the UK. The synergy between these related projects has helped to reinforce and support each other through collaboration and shared experience, ultimately aiding the development of the Edinburgh Research Archive (ERA).

The drive for the proposal of the Theses Alive project came from the original e-theses investigations carried out at the Science and Engineering

Library, Learning and Information Centre (SELLIC) at the University of Edinburgh. The SELLIC team presented a report to the UK Theses Online Group (UTOG) in late 2001 on the results of a doctoral theses digitisation project. The report concluded that universities were moving into a digitally networked environment which had the potential to transform the current system for providing access to theses by making them open access online.

Under the Theses Alive remit to investigate the technological and cultural issues involved for UK higher education institutions wishing to attain e-theses capability, the following general objectives were proposed:

- to develop a digital theses submission system for use by interested universities;
- to develop a standards compliant digital infrastructure to enable e-theses to be published online (with a subobjective that 500 e-theses exist within the UK segment of the Networked Digital Library of Theses and Dissertations (NDLTD) within two years);
- to develop and support a metadata schema for the UK higher education e-theses environment;
- to test the value of a national support service for e-theses creation and management in the UK;
- to produce a 'checklist approach' institutional guide to adopting and managing e-theses;
- to work with other e-theses developments internationally, and in particular to assist the research aims of other JISC FAIR programme projects.

Throughout the course of the project a wealth of activities whose significance had not initially been fully realised were addressed. These included areas such as:

- advocacy; not only of the service, but of the concept of open access;
- licensing, copyright and other intellectual property issues;
- open source software development, maintenance and dissemination;
- post-production service administration and continued technical support.

In order to achieve these results a core team of three staff at EUL was formed, consisting of a project director, a project officer and an

information systems developer. The project investigated two main strands: technical development and advocacy/liaison. Each of these was primarily investigated by the information systems developer and the project officer respectively, under management from the project director. Each strand was, however, closely related to the other, making feedback essential to shape the development of the work packages in each strand.

Preliminary decision making

Beyond the project proposal's suggestions and recommendations there were some minor additional decisions to be made before work could start in earnest, concerning the software development process. The first was to take the route of open source software (OSS) to provide the basis for the resulting e-theses management system. This decision was influenced by two main points:

1. It is desirable, when following the ethos of open access, to endorse OSS, as both have highly related objectives.
2. The JISC recommend the use of OSS wherever possible in funded projects.

There are also general advantages in using OSS, including zero cost of acquisition, the ability to use and adapt to meet local requirements and the freedom to distribute modifications.

The subsequent decision from here was which packages to adopt for evaluation for the repository. As the appendices demonstrate, there are now many packages which may provide the functionality, and it was not feasible (or possible at the time) to evaluate all options. We therefore chose between two likely packages, knowing that the DAEDALUS project was evaluating two packages, with one package being common between projects. This would provide us with the opportunity to compare three packages before making a final decision.

Development

This section discusses the many issues encountered during the development process of a combined e-theses and e-print repository which ultimately became ERA. Much of what is described in the following subsections happened concurrently across the project, and there is a great deal of interaction between each of these areas.

Software evaluation and development

Initially the project carried out a broad review of current open source digital repository packages available, and an in-depth evaluation of two packages. It was felt that a formal evaluation of the most commonly used platforms would provide the most robust approach and eventually yield the most comprehensive and meaningful results. These results could then feed back into the design process for developing a system suitable for use in the UK context.

The comparison, carried out as per the evaluation guidelines outlined in Chapter 3, looked at some of the common elements between the packages and drew conclusions on which was best in each field. In addition, it looked at how difficult it would be to modify each of the packages to provide an e-theses service for the UK. This analysis was considered alongside the medium-term future of each of the packages as they are developed as well as the scope for expansion that each package had within the library and the university as a whole.

A direct comparison of the software was difficult because of the differing focal points of their functionality and design philosophies. The main part of the study considered elements particularly relevant to e-theses as well as essential requirements such as security and administration. For example, in the area of metadata collection we were particularly interested whether the data collected was sufficient and relevant, or more importantly extensible or flexible in any way. We compared the metadata handling features of the system particularly in light of the complications we were expecting to encounter during schema development. We also looked at the support for the OAI-PMH protocol, via which exposure of data was an essential requirement and part of the initial project proposal. Holding material in a digital repository confers on the host institution some responsibility with regard to long-term care, thus another factor to consider was the preservation focus of the package and the stability of its storage layer.

With the e-theses functionality evaluated, we then gave consideration to more general features of the software, such as its ease of customisation, the configuration options available to the system administrators, and the general design methodology employed. In addition we were interested in the community surrounding the software, as this can often be an indicator of the likely longevity of the package, especially in an open source arena.

We arrived at a situation where one package had the features we wanted in some form, but was not at a stage of development where we

would be happy deploying a service on it, while the other was a solid package with much of the groundwork for e-theses in place, but no specific functionality. Therefore we could ask the question defined in Chapter 3 in two ways:

1. How hard would it be to add the required functionality to Package A to make it support e-theses for the UK?
2. How hard would it be to add the additional support features to Package B to make it acceptable for institutional usage?

After considering the feedback from the DAEDALUS project regarding the third package and after several months of testing and evaluation we decided to build ERA using the DSpace software (see Appendix C for more details). The reasons for this were, at the time, as follows:

- metadata capture and storage techniques were relatively flexible;
- support for OAI-PMH was at the most recent version;
- the storage system was geared towards digital preservation, although at the time there were still no clear procedures;
- the underlying application design and implementation was of a reasonably high quality, supporting good internal authentication and authorisation procedures;
- the administrative interface was relatively mature, and provided many features;
- the community surrounding it was already strong and showing signs of growth which gave us confidence in its future;
- feedback from Glasgow suggested there was no specific way in which DSpace and their other evaluated package could be defined as better than the other.

Nonetheless, it lacked some of the functionality we were interested in and immediate support for the metadata schema which we were in the process of developing. Therefore, taking more from the evaluation than we first anticipated, we used our other package to help us define the work that we needed to undertake. The feature list that we then defined was:

- support for multiple metadata capture processes (submission procedures);
- enable capture of UK e-theses metadata;

- allow for rapid identification of content types within the repository;
- apply multi-part licences to the e-thesis;
- apply 'physical' restrictions to e-theses where necessary;
- a collaborative workspace where supervisors and students could observe and work on a submission;
- an annotation tool, to allow supervisors to leave comments for students.

The next challenge we faced was the best way to implement these changes to DSpace, which required developing or adopting a methodology for third-party software developments. We chose to write and maintain our own 'add-on' to the system, which would require installation onto an existing repository. We chose this method over writing our changes directly into a local copy of the source code or committing our changes to the central source code repository for a number of reasons (Jones and Andrew, 2005):

- our developments were not necessarily of interest to the whole repository community;
- the development model for DSpace at the time was not easily compatible with simply writing our changes back to the main code-base;
- our developments were UK focused, and we did not anticipate them moving at the same speed or in the same direction as the main DSpace development process.

For these reasons we created our own online source code repository, and were free to choose our own development model. Naturally it was necessary for us to always work from the most recent ('bleeding-edge') version of the DSpace source, and we employed a lightweight and iterative development cycle, which is easily to implement for a small product within a small development environment. We broke up the software into components as defined by the requirements stated above, and began by developing what we considered to be the most useful functionality first, taking into account the current state of developments outside the software process, such as the metadata schema.

The result of this development work was named the Tapir (Theses Alive Plugin for Institutional Repositories), and was free to download as a self-contained add-on to the core DSpace code. Subsequent download and use of this software by institutions all over the globe resulted in

quality feedback which in turn was fed back into the iterative development cycle for further advancement of the software.

At the end of the project, Tapir offered many of the features originally specified (although some fell by the wayside due to other developments in the area or lack of interest in the functionality). Some of the features were found to have uses outside of the e-theses sphere of interest, and a subset of features have also now found their way back into the main source distribution of DSpace.

Metadata schema development

A primary aim of the project was to work with other e-theses developments both internationally and with the research teams of other e-theses projects in the UK. As part of this objective we participated in the creation of a recommended UK e-theses core metadata set. Led by the Robert Gordon University, working with representatives from the University of Edinburgh, the University of Glasgow and the British Library, this set was created in preliminary form and sent out to interested parties for comment. Feedback from this consultation then resulted in further refinements to the metadata set, which is now maintained by the Robert Gordon University (see Copeland et al., 2005).

As a guiding principle, we felt it was necessary to ensure that we coordinated activities with other initiatives and projects to produce meaningful outputs and results. Where international standards were already available and in common use (e.g. the OAI-PMH) we would try to adopt and support the concepts and implementations of these protocols. We examined potential metadata sets that may be able to support e-theses: the default DSpace Dublin Core registry configuration; the Electronic Theses Dissertations Metadata Schema (ETD-MS) from Virginia Tech and the NDLTD, and the Theses and Dissertations Markup Document Type Definition (TDM DTD). With the aim to 'genericise' metadata creation processes for UK e-theses we drew on the recommendations by these various schemas to produce the final set.

The recommended UK e-theses metadata set, therefore, supports the elements that are common to all UK theses, with suggested additional options for classification of holdings using various common classification schemes. As an advantage, it is easily represented in an *element.qualifier* style in which qualified Dublin Core can be expressed, and by obeying pure Dublin Core basic element definitions can be reduced from its full form to that which can be natively transported via OAI-PMH without major data loss.

As DSpace effectively supports any metadata in the *element.qualifier* form, and will compress this data into the standard Dublin Core elements for exposure via OAI-PMH, it was relatively straightforward to implement this schema as part of the Tapir software. Using the submission software the students insert their own metadata, which is subsequently quality-controlled by the library, and thus automatically compliant with the requirements defined at this stage of the project.

Simultaneously the University of Nottingham provided us with the core metadata set that the SHERPA project intended to work with across the institutional repositories with which it was involved. This could also be represented in an *element.qualifier* style, so it was straightforward to see that we would be able to support both metadata sets within DSpace and were able to provide a dual submission interface to deal with each set.

Policy and procedure development

Alongside the technology strand of the project, there were also many administrative and managerial policies and procedures to investigate, define and develop. The software for the research archive would allow for the collection of e-theses metadata and e-print metadata and the additional tools required to manage them. It was, therefore, also simultaneously necessary to investigate how the repository would be looked after and fit into day-to-day working of the institution so that feedback could be passed to the technical strand.

The first form of this feedback was to suggest that in addition to creating an e-thesis archive under the guidance of the Theses Alive project, it would be advantageous to support within the same environment e-prints, and potentially other types of research materials. We found that there was a strong endorsement from academic staff and students alike to support the inclusion of e-theses in an institutional repository with other research outputs. With a firm decision made to house the content together it was then possible to look at the implications for the service and how it would be managed. It was at this stage that a firm advocacy strategy was developed and put into action. More details of the planning, form and implementation of the advocacy campaign are discussed in Chapter 5.

As previously mentioned in Chapter 5 we felt it would be beneficial to perform a baseline survey of research material already held on departmental and personal web pages in the University of Edinburgh

domain. A systematic approach was taken, whereby each departmental and staff web page was visited and the content of self-archived material was noted. The survey looked at each college in turn, searching for content at each level of the hierarchy, down through the school and to individual levels. During the period of this survey over 2,500 staff web pages were visited.

Initially the survey began with documenting formal research material (post-prints, pre-prints and e-theses) within the science and engineering domain, but when other colleges within the institution were examined it became apparent that the type of material available online varied considerably between subjects. To represent these different research cultures other content such as book chapters, conference and working papers was also considered when compiling the data.

Considering the wide-ranging self-archiving trends between academic colleges, and even within schools, there appeared to be a direct correlation between willingness to post-material online and the existence of subject-based repositories. In the small variations from this rule we would argue that some subject repositories (such as the Los Alamos ArXiv for high-energy physics) have become so successful at capturing and making persistently available a very high proportion of the output in their domains that academics trust it as their 'natural' repository for self-archived material. So it appears that where there is a pre-existing culture of self-archiving e-prints in subject repositories, scholars are more likely to post research material on their own web pages until such time as the subject repositories become trusted for their comprehensiveness and persistence. As personal web pages tend to be ephemeral, the long-term preservation of the research material held on them is extremely doubtful. We were, therefore, proposing to provide a more stable platform for effective collaboration, dissemination and preservation of research.

This scoping study (for more details see Andrew, 2003) proved to be extremely valuable and provided evidence that there was already a substantial corpus of research material available from personal and departmental web pages in the Edinburgh domain. It was extremely encouraging to see that such an unexpectedly high volume of research material (over 1,000 peer-reviewed journal articles) were available in this manner. Originally we planned to contact the pre-existing self-archiving authors to gather initial content for the repository (as described in Chapter 5). Unfortunately a high proportion of the material was published on the Internet with no consideration to intellectual property

rights. In practice this meant that we, as responsible repository owners, were not in a position to take all of this content.

It was also identified fairly early on that academics were interested in maintaining at least some distance between e-theses and research papers, suggesting that in some situations the former were 'research training' and not necessarily up to research standard. This then fed back into our repository design by introducing a requirement that all content types are rapidly identifiable.

We also successfully defined the requirements of the relationship between thesis authors and supervisors. The requirement was to allow supervisors to observe the work of students, to make changes, suggestions or comments prior to submission of the thesis. By proposing a collaborative workspace wherein items in the process of being authored could appear in both the supervisors' and student's private areas, we were able to define how an e-thesis repository and an e-print repository could be natural partners. As a unified workspace could contain both the supervisors' students work and also their own academic works, we could reduce the number of systems necessary for authors to use, lowering barriers to adoption. Allowing annotation of items in this workspace would also enable us to support online, recorded communication between students and supervisors, and increase the likelihood that academics may also wish to use the system for peer-to-peer collaboration.

One aspect of the survey demonstrated the lack of consistency in dealing with copyright in intellectual property issues. Some academics responded to these uncertainties by not self-archiving any material at all; others used general disclaimers which may or may not be effective; a minority posted material online which is arguably in breach of copyright agreements. Most, however, took the middle line of only posting papers from sympathetic publishers who allow some form of self-archiving. It is apparent that if institutional repositories are going to work, then this general confusion over copyright and intellectual property rights (IPR) issues must be addressed at the source.

It has, therefore, been necessary to investigate the effects of IPR and other legal implications (e.g. the Freedom of Information (Scotland) Act 2002) which arise when publishing research material online. These unforeseen problems have proved to be a significant barrier to the progress of the project and the development of repository programmes in general.

As previously mentioned in Chapter 6, there exist some genuine concerns about the premature release of research material in PhD theses,

which raises the need for some items in ERA to remain confidential. The e-theses solutions developed by the Theses Alive project (for example Andrew, 2004) have proved to be very valuable to the higher education community. In practical terms for e-theses, we considered two main issues:

- the range of parties involved: the submitter, the institution and the end-user,
- that the restrictions placed on an e-thesis are not necessarily absolute; they may have time or domain dependencies.

In order to address these points we defined six scenarios where restrictions could be applied to an e-thesis such that it could be stored within the repository:

1. *No restriction*: the item is not restricted from access in any way.
2. *Domain restricted for one year*: the item is restricted only to users within the institutional domain for one year.
3. *Domain restricted for two years*: the item is restricted only to users within the institutional domain for two years.
4. *Withheld for one year*: the item is restricted from all users including the author for one year.
5. *Withheld for two years*: the item is restricted from all users including the author for two years.
6. *Permanently withheld*: the item is restricted from all users including the author for all time.

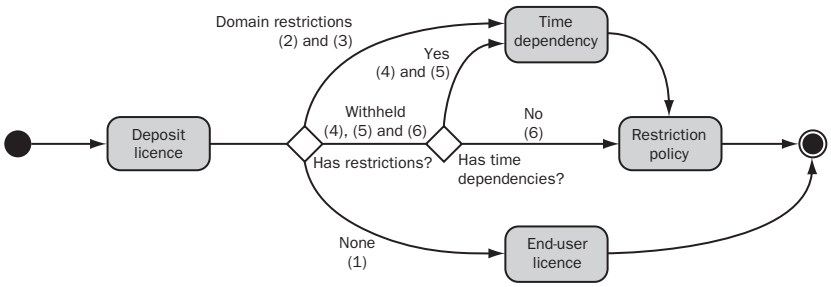
Thus, we defined a three-part licence which would allow for a comprehensive treatment of this problem. The licence is split into a deposit licence, a use licence, and a restriction policy. The deposit licence primarily gives the rights to the repository to hold the material in perpetuity, and to transform and migrate that work as and when necessary in order to meet the requirements of digital preservation without changing content wherever possible. We have also selected a creative commons (CC) licence under which the theses can be used; the authors are required to agree to this, as we felt this would make the material sufficiently open access to be of use, without compromising the author's rights. The version of this licence that is in use is an attribution, non-commercial, share-alike licence, which implies that any derived works must attribute the author of the thesis, and must also share that derived work under the same licence, with

no commercial use of the item allowed. Of course any of the terms and conditions can be renegotiated at any point with the author if they are not deemed suitable in the future. Finally, the submitter is prompted to select the desired scope of restriction and provide appropriate FOI exemption wording specified during the submission process. Figure 7.1 shows how this licence is constructed.

While it was necessary to investigate these separate issues for e-theses, IPR concerns for e-prints are primarily based around the publisher's policy regarding self-archiving. Later in this chapter we discuss the processes that must be followed when depositing an item into the repository to ensure the intellectual property rights are not breached. At this stage we note that there are generally no restrictions applied to e-prints in ERA because any items we are legally permitted to deposit are not affected by the Freedom of Information Act and prior publication issues in the same way that e-theses are. Instead we require the submitter to confirm to the repository that the author is the sole copyright holder, or that they have permission to archive the item in a public space. A more comprehensive discussion of the other deposit licence criteria that we considered vitally important can be found in Chapter 6.

The next issue to be addressed was that of how to the brand the repository service. The initial plan was to integrate the service seamlessly into the university library web presence, and to provide smooth transitional navigation between the two systems. Throughout the course of the policy development, though, it became understood that branding ERA as a library service may discourage potential users or departments from endorsing the service, and for that reason the design coupling between the systems was weakened; a derived but unique branding for the service was, therefore, proposed and deployed. It was also decided early on to refrain from using potentially confusing nomenclature and to

Figure 7.1 **Three-part licence construction**



use generic terms that the academic community would feel comfortable with. Although the Library decided to use the DSpace platform as the basis for the repository, all references to DSpace were removed. DSpace as a software product is well known in the Information Science and Digital Library communities, however, in the wider academic community it is relatively obscure. A generic service name like the *Edinburgh Research Archive* has a more instantaneous recognition of function than any associations that DSpace or e-prints would confer. An additional rationale to adopt a neutral nomenclature is that any efforts to develop a strong brand would not be wasted if at a future date the underlying repository platform was changed. There were also issues concerned with how the contents of the repository would be surfaced within existing and future university systems. A great deal of work was in progress with other projects to provide portal-like access to many resources including the institutional repository, so the need for specific product branding was reduced further on the grounds that ultimately the service may be invisible to many users.

Other outputs of this process of policy development include best practice guidelines for institutions wishing to adopt electronic theses, and the authoring of extensive management and administrative procedures which will be discussed in more detail later in this chapter.

Deployment

One initial aim within the project plan was to work with a set number of additional higher education institutions to help test and develop the proposed e-theses management system; the project officer would arrange and liaise with a number of pilot institutions taking delivery of the project software, to gather feedback about the system and to help provide installation and end-user support. As the project progressed it became apparent that a national e-theses support service was not entirely appropriate at the time. Although it is necessary to help institutions build repositories and appropriate policies, it was felt that other types of support such as student support or mediated deposit would be best offered by the local institution where embedded staff would have detailed knowledge of current working practices. This was a common theme through all feedback from the initial partner institutions during site visits.

With the software side of the project approaching maturity we moved on to provide a pilot e-thesis service within the University of Edinburgh.

As a proof-of-concept we worked closely with two schools within the university: the School of Informatics and the School of Geosciences. During the pilot phase we hoped to refine our e-thesis service from the combined experiences of users and administrators alike, before expanding to cover the university as a whole. At the same time we hoped to assist our partner institutions in setting up similar e-theses repositories by providing technical and advocacy support.

The two pilot schools were chosen to represent as fully as possible a wide range of disciplines, which could have an impact on the types of material submitted. The School of Informatics, to some extent, already had a culture of producing e-theses, but lacked an effective method of online dissemination. The School of Geosciences, however, had no previous experience creating or publishing e-theses, but were willing to embark on the pilot. To encourage submission we felt that incentive was needed, particularly for the geoscience students; to meet these aims we arranged for the project to pay for one hard-bound copy of a thesis for every electronic version submitted during the pilot.

Typical theses from geosciences include features that could be problematic to represent electronically; for example, large fold-out inclusions, high diagrammatical content and large auxiliary data sets. By including these types of thesis, the pilot study hoped to directly asses the impact on students and the repository itself. A significant component of this part of the pilot was dedicated to providing end-user support for postgraduate students and supervisors via telephone and web-based technologies. During this time 20 students completed their doctorate theses and submitted an electronic version.

The School of Informatics study was more concentrated on investigating and developing a sustainable strategy for high-volume ingest; this included topics such as providing efficient workflow and format conversion. During the pilot phase the project gathered 136 theses retrospectively and obtained 11 theses submitted electronically.

Developing such a system in isolation is, of course, unwise, and throughout the lifetime of the project it was necessary and desirable to disseminate findings as well as to interact heavily with other researchers in the field. From these interchanges we found that many institutions had achieved successful e-theses programmes by mandating at a top-level the electronic submission of theses and dissertations, especially in the USA. This persuaded us to pursue a strategy of persistent lobbying for postgraduate degree regulation change at the highest level to mandate that students submit their theses in both electronic and print forms. The successful adoption of this policy has been a crucial moment in the

development of ERA, and mandatory submission of e-theses will start to take effect around 2008/9. Changing university regulations is a notoriously slow process, and plenty of time should be allocated if pursuing this course. In addition, the postgraduate studies committee has been encouraged to regard the electronic copy as the authoritative version ('golden copy'). Printed copies can then be derived from the electronic version and bound by the library. If successful, then electronic theses submission may become the default, even before electronic deposit is mandated by regulations. A decision was made to develop a mediated deposit service and provide e-thesis creation support. In practice this consisted of providing guidance for postgraduates and supervisors on suitable file formats, scanning resolutions, conversion and system administration. This user support service was successfully piloted, and mediated deposit has become a formalised method of obtaining repository content of both e-theses and other research types, as will be discussed later in this chapter.

With the pilot study complete, and a small collection of content in the form of e-theses, e-prints, technical reports, conference papers and related research material, ERA went live in October 2003. The next stage in the advocacy process was to raise general awareness through internal publicity. To raise the general awareness of repositories and other open access issues we decided that an appropriate action would be to hold a seminar. We arranged such a meeting and sent invitations to every single academic staff member in the university. The only practical way to do this was via e-mail, and distributed leaflets.

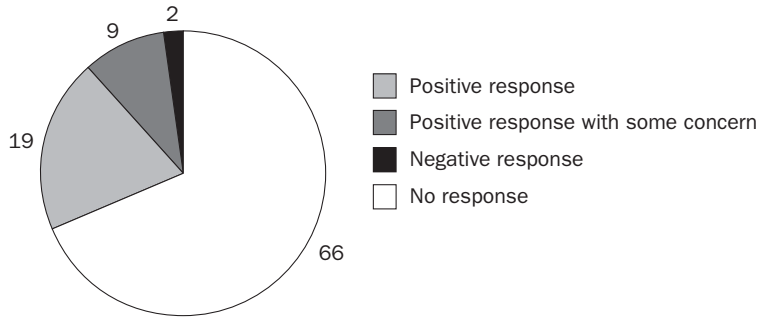
Careful consideration was given to the relative timing and the venue itself. To attract the maximum number of staff we held the seminar in late summer, when most faculty have no teaching obligations and were not likely to be on vacation. For ease of access the venue itself was situated in a central location. To widen the appeal, and to prevent our endeavours from appearing too parochial, a number of speakers from external organisations were invited to give presentations. Senior management were also invited to lend their support to the initiative. Despite our best intentions the event was only modestly attended by members of the academic community. We felt that this lower than expected turn out was in part due to the reluctance of faculty taking time out from their schedule to travel to a central venue to listen to presentations in which they may only be marginally interested. Learning from this experience we decided that any subsequent advocacy seminars would be better placed if we held the event in their own environment (Chapter 5).

Following the advocacy plan we developed, the next stage was targeted content recruitment (Chapter 5). Academics identified from the initial baseline survey (Chapter 5) with significant content (ten items or more) already online in personal or departmental web pages were invited to deposit their content into the repository. Due to the scale of content, the faculty members were initially approached via e-mail. Figure 7.2 shows the scale and range of responses from one targeted content recruitment project at the University of Edinburgh.

During this particular targeted content recruitment drive, made during the summer of 2003, 96 individual academics from the subject areas of science and engineering were contacted initially by e-mail. Subsequently we had a response from 30 individuals (a response rate of 31 per cent). Part of this lower than expected turn-out was due to the timing of the project – five academics were away on vacation or research. From the remaining respondents, 19 were happy to self-archive some of their material immediately, whereas nine were more cautious. After explaining the aims of the project and soothing concerns they were also happy to deposit material. Only two academics were strongly opposed to being involved in the study from the start. Interestingly, one of these academics later changed their opinion and was actively involved in a departmental pilot study (Chapter 5) after the involvement of an opinion leader.

The actual responses from academics made interesting reading and broadly fell into four categories. Examples 1–5 taken from real-life subsequent correspondence with academics illustrate these points:

Figure 7.2 **Pie chart showing the range of response from academics at the University of Edinburgh with content already online in personal web pages who were invited to deposit material in the institutional repository**



- Broadly welcoming:
  - 'This seems like a worthwhile endeavour and, yes, I am interested in having my research work in such a repository.'
- Concerned about extra workload involved:
  - 'You can include my papers as long as I don't have to do more than sign the permissions. Some of the departmental archives take a ridiculous amount of staff time to populate.'
  - 'My only reservation about using a centralised repository would be ease of use. Right now I send PS and PDF files to a public area with unix commands and I don't have to worry about passwords, formats or anything. I can change versions in an instant (I know this is horrifying to an archivist).'
- Concerned about copyright:
  - 'One thing though, I'm sure there are potential copyright issues … I think I would like more information on that side of things before I get involved with a more formal repository. I think one is unlikely to raise too much ire by having PDFs on a personal web page, but I could imagine journals being a bit more touchy about copyrighted material on a more official university website … This difference may seem trivial but sleeping dogs and all that!'
- Unwilling to participate:
  - 'No, not at present. There is already a world-wide archive of research papers in physics that is used extensively.'

These quotes seem to encapsulate a whole range of common reactions by academic staff towards institutional repositories. Familiarity with these points can help to formulate responses which will aid in content recruitment.

Administration

Providing ERA as a service is similar to providing any other institutional web resource, and the administration of these services can often be as challenging as the technical support requirements. When deciding how to administer ERA we had to determine how much effort would be put into areas such as metadata verification, administering policies for users,

setting up research collections, correcting post-submission errors, and defining the archive structure.

Various solutions have been found to these problems and detailed documentation has been produced to deal with almost every part of standard service maintenance. Tasks which fall outside the normal bounds of library and administrative work are dealt with by a group of individuals with the relevant knowledge and experience. These tasks include decision making with regard to the state and development of the system as well as liaising with academic departments. An informal ERA Management Group (EMG) has been set up in order to deal with these broader issues and will be discussed later in this chapter.

As discussed in Chapter 4, there are many areas of the system which need to be soundly administered in order to run a service which does not get out of control. This section introduces some of the actual administrative decisions and processes used to operate ERA with a reasonable degree of success.

An important goal for the repository was to define a relatively static collection structure and have this map onto the institutional organisation as easily as possible. We define a community as a collection which may only contain other communities and collections in a DSpace context, and, therefore, these are used to create a shallow hierarchy within which the university's research will be categorised. A community maps directly in most cases onto a single recognisable academic unit such as the School. The collection, then, maps onto any internal subsection of the community, including working groups, institutes or centres. For example, the informatics community contains the Centre for Intelligent Systems and their Applications, Department of Artificial Intelligence and the Institute for Adaptive and Neural Computation. Naturally, this relationship of communities and collections to academic units does not always hold, and we leave it to the administrator to use their own discretion in unusual cases.

In terms of how research types should be distributed among the collections we faced a number of decisions regarding how best to reflect this in the hierarchy. After trying a number of configurations, and taking into account comments from academics regarding the perceived necessity to separate e-theses from other forms of research output, we decided use separate community and collection structures to deal with the different types of content. Extending this idea to be more generic we allow the communities and collections to have special designations attached to them to define their function beyond that of being affiliated with a

particular academic unit. A particular case of this is that we define a community and its collections as being designated only for theses and dissertations.

To control all of the configurable system elements, we developed systematic naming conventions to which administrators must adhere. We identified two system elements to which this needed to apply: research collections/communities and user groups. The objective was to create name structures for each of these which allowed the purpose and likely content to be known quickly and easily, and for like elements to be easily found together in various browse contexts.

For communities the naming convention is defined (logically) in almost the same was as the community itself is defined; that is, by the academic unit to which it belongs, with an associated element which allows the administrator to define a special designation for the content. Thus the following general statement defines how they should be named: '<school to which community belongs> (<special designation>)'.

Therefore we would name the *theses and dissertations* designated *informatics* collection as simply: 'Informatics (Theses and Dissertations)'.

Similarly, the convention for collections is defined in the same way as the collection itself is defined, as being that of the subsection to which it belongs and the associated special designation, thus: '<group to which community belongs> (<special designation>)'.

Therefore we would name the *theses and dissertations* designated *Institute for Stem Cell Research* collection as: 'Institute for Stem Cell Research (ISCR) (Theses and Dissertations)'.

A similar methodology is used for naming user groups. We identified four primary user types: workflow administrators, theses supervisors, content submitters, and collection administrators. Each of these user types performs a specific role in the administration of ERA, and each will, therefore, have similar system policies associated with them. These policies can be effectively managed if applied to general groups of users, rather than on an individual basis (as is common in many computer systems), and we can make it easy to identify the relevant group at all stages by having sensible naming conventions. The general form for all these group names is: '<group prefix>: <associated system entity>'.

By having group prefixes associated with each group type, and a target entity of each group's policy, we can quickly identify who is working with what. We are simultaneously enforcing a very rigid 'one group, one purpose' model which can result in a large number of groups, but all of which are easy to manage. The prefixes we have chosen are:

- WF (<stage number>): A workflow group for the numbered stage in the process (there are three available stages in the workflow);
- SU: A supervisor group;
- IN: A submitter group;
- AD: A collection administrator group.

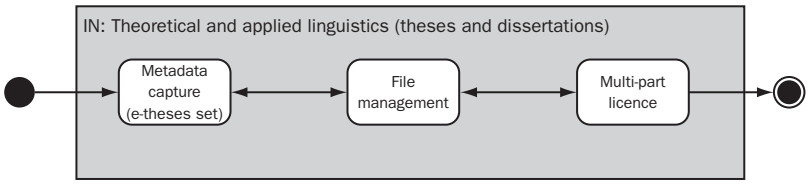
Therefore, the following group names would be allowable:

- *WF(1)*: *Institute for Cell and Molecular Biology (ICMB)* – The first workflow group for the Institute of Cell and Molecular Biology.
- *WF(3)*: *Accounting and Business Method (Theses and Dissertations)* – The third workflow group for the Theses and Dissertations designated Accounting and Business Method group.
- *SU*: *student@myu.ac.uk* – The supervisor group for the student with e-mail address *student@myu.ac.uk*
- *IN*: *Atmospheric and Environmental Science* – The submitter group for the contributors who can submit to the Atmospheric and Environmental Science collection.
- *AD*: *Celtic and Scottish Studies* – The collection administrator group for the Celtic and Scottish Studies group.

Each of these user groupings allows for a set of users with a defined purpose to be allocated the relevant system policies to permit their actions, or be referenced by other areas of the system to be allocated certain types of functionality. The workflow system, for example, is integrally linked to the workflow groups, in more than just pure policy (although this is also required to be correctly configured). Each workflow group has a set of defined actions associated with it such that it can be presented with the relevant options at the relevant stage of a submission's passing through the system. The first stage contains options to merely accept or reject the submission; the second has the additional option to edit the metadata and file content of the item; the third stage implies that the item is destined for the repository and permits only metadata and file management and ultimate acceptance for archiving.

The ERA is specifically aimed at handling research split into two broad categories: e-theses and all other research output. For this purpose there are several abstract pre- and post-submission workflow models defined which are implemented on a case-by-case basis for material as it is submitted into a collection. Each collection is associated with an implementation of one of these workflows, based upon the special

Figure 7.3 E-theses submission workflow



designation given to it, or the route via which it will be placed into the archive. The workflow diagrams in this chapter use examples of possible naming conventions for further clarity.

Figure 7.3 shows the submission workflow for an e-thesis. First, the metadata fulfilling the recommended UK e-theses requirements is collected from the submitter. These data include some information which is pre-populated by ERA, and unchangeable by the submitter, such as the host institution and department under which the work has been produced. Second, the files for the thesis are collected. Finally, the student must agree to a three-part licence which covers the rights of ERA to hold a copy of the thesis, the rights that the user gives to the users of the online version, and the Freedom of Information (Scotland) Act 2002 disclaimer associated with the restriction type (if any) that they define at this point. This multi-part licence is then constructed as explained earlier in this chapter.

Restriction of theses is only acceptable provided that one of the FOI exemptions is met, and the licensing stage also allows the submitter to choose which restriction option they require and also to provide a reason for this. The system then builds a large multi-part licence which is stored in the archive alongside the rest of the item (see Figure 7.1).

Figure 7.4 shows the submission workflow for all other types of research output. First, the metadata compliant with the recommendations from the SHERPA project is collected. Second, the files are uploaded. Third, the submitter needs only to agree to a deposit licence to allow ERA the rights

Figure 7.4 Other research material submission workflow

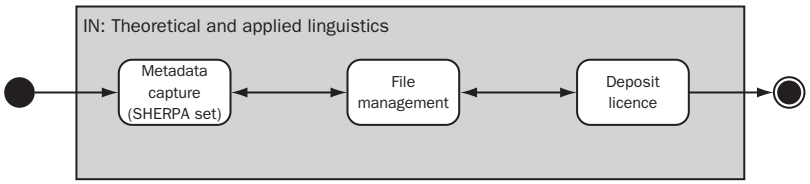
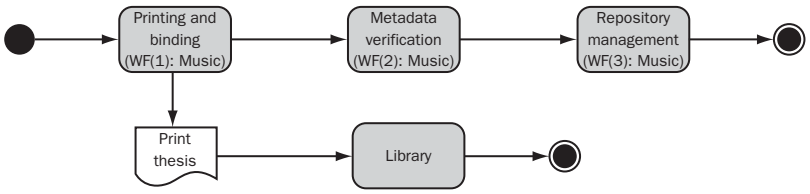


Figure 7.5 E-theses post-submission workflow 1

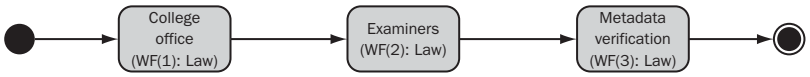


it needs to hold the material in perpetuity, and perform migrations and transformation as with the e-thesis. This is because the copyright situation is often more clear-cut at this stage for the material, insomuch as it is usually controlled by the journal publisher (as much of the contents are e-prints). There are also no FOI issues as the material is published and available anyway.

Figure 7.5 shows one workflow configuration for an e-theses collection in ERA. Once the post-submission stage of ingest has begun the thesis can go straight to the bindery, from where copies of the thesis can be produced and bound. In liaison with the student, this department can produce the requisite number of print theses as required by regulations as well as guarantee that the 'golden copy' (i.e. the electronic version), is identical to the print versions (a common issue with e-theses) by the very fact that the print is derived in a controlled way from the electronic. Once this has been done and the library has taken delivery of the print versions, the e-thesis moves on to the second stage, which contains collection librarians who will be responsible for ensuring the quality of the metadata and performing the final checks before the thesis is allowed to irrevocably reach the archive. Once value-added metadata has been inserted (e.g. the application of standard classification schemes such as Library of Congress Subject Headers where appropriate) and the quality of the submitter-authored metadata has been verified then the thesis can move on to the final stage. Here, the repository administrators get a final opportunity to ensure that any necessary restrictions have been applied and that the copy is in a fit state to be archived, performing any necessary additions to the contained files along the way (e.g. inserting copies in standardised formats).

Figure 7.6 shows an alternative workflow configuration which is currently not in everyday use, but exists should changes in the way people rely on ERA as a support service require it. Here, once the post-submission workflow begins, the first point of contact is the college office, which ensures that the thesis is intended and ready for submission

Figure 7.6 E-theses post-submission workflow 2



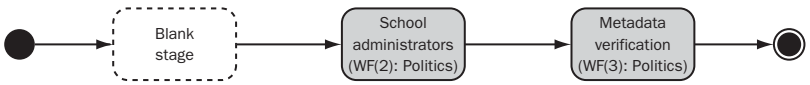
and that all relevant paperwork has been done in advance. If the thesis is ready to go ahead it can then be presented to the examiners of the work, to ensure that no corrections are outstanding or necessary and that the thesis has indeed been accepted for award. With these checks complete it is then finally up to the collections librarians to ensure the quality of the metadata, add additional catalogue information where necessary, and confirm that the thesis is in a fit state to archive.

Figure 7.7 shows the much simpler workflow required to support the post-submission phases of all other research output from the university. Only two stages are required for this sort of material. First, a school administrator has the opportunity to confirm the validity and authenticity of submissions; that is, that they are appropriate for the collection to which they have been submitted. Second, collections librarians will be responsible for ensuring the quality of the metadata, as before, and verifying that the submission is in a fit state to enter the archive. Note that the group names for the two existent workflow stages correspond to parts two and three. This is because our three workflow stage types in DSpace support different activities, and these groups are associated to those specific roles: accept/edit/reject and edit/reject respectively.

All these workflows and naming conventions along with how-to and troubleshooting guides and full administration procedures have all been gathered together into a single ERA Administration Guide. This documentation acts as a single reference point for all administrators, ensuring that there is consistency throughout and that long-term maintenance is possible. As an added bonus, the documentation acts as a set of extremely detailed use-cases against which new versions can be tested for functionality and suitability for purpose.

In the early stages of ERA's life we also offered a mediated submission service; this service has brought with it its own workflow procedures

Figure 7.7 Other research material post-submission workflow



which override the previously discussed ones. The reasons for this, and the resulting related workflow issues are discussed in the next section.

Management

Beyond the procedural administrative requirements of the whole ERA system (which includes the people involved in its running and maintenance) there are data management issues which need to be looked at more closely and defined carefully. As previously mentioned, tasks which fall outside the normal sphere of library and administrative work are dealt with by the ERA Management Group (EMG). The tasks addressed here include the state of the development of the system as well as liaising with academic departments and handling mediated submission. It administers which institutional units are represented in the repository, obtains content, influences university regulations and implements functional requirements.

To manage the requests of various natures to the EMG we employ the university call management system (CMS) which is used for logging, tracking and reporting on the interactions involved in resolving support requests. The policy is that any task which takes more than a few minutes to administer in ERA should be entered into the CMS (note that this excludes requests purely regarding the underlying software packages, which are addressed directly to the development community). Other requests may not necessarily be directed to the CMS initially, but may lead to a call being opened. For example, if an e-mail is received asking '*Why is ERA not registered with harvester X?*' then a call would be opened in the CMS: '*Register ERA with harvester X*'.

We have defined a set of protocol tags which are attached to each call that is opened in the ERA CMS. These are placed in the short description of the request to ensure that efficient searching and querying of the open calls can take place. Effectively this is further use of good naming conventions for improved usability. These tags are as follows:

- *[FEATURE]*: high-level, non-technical feature request or suggestion for ERA. These should be requests which are not directly about the nature of the underlying software, and will be followed up by a member of the EMG.
- *[DEPOSIT]*: anything relating to the submission of items to the repository, including copyright issues. The aim is to answer each of these requests within seven days of receipt.

- *[ADMIN]*: any administrative task to be performed on the system. This includes user authorisations, group management, community or collection configuration and so forth. This will primarily be addressed by the ERA system administrator.
- *[FOI]*: Freedom of Information (Scotland) Act 2002 invocations. By law requests for information should be completed within 20 days. During this time advice is sought from the local FOI adviser.
- *[ENQUIRY]*: any information request which is not FOI related, such as requests for advice regarding issues including as copyright or best practice. The aim is to answer each of these requests within seven days of receipt
- *[!]*: the request is urgent and should be dealt with immediately. This enhances other requests on this list.

So, for example, you may find the following in the CMS: '*[ENQUIRY][!] Legal question over copyright content*'.

We aimed to provide a general information and user support service for submissions. This partly took the form of the mediated deposit service, which in practice consisted of providing guidance for postgraduate students, supervisors and academics on suitable file format types, scanning resolutions, and format conversion. This service has been warmly welcomed by students and academics alike. The submitter passes to EUL an electronic copy of the item to be placed in the repository and a member of the EMG checks the copyright status of the work, converts file formats as necessary and submits the item to the relevant collection with the relevant metadata.

This sort of service comes with quite a large administrative overhead, and the long-term sustainability is a question that should be considered by repository managers before implementing. For our situation, ERA is considered a core library service and features prominently in the University of Edinburgh's knowledge management strategy (Hayes, 2004). Given that the task of actually submitting on behalf of another is not too complex, documentation describing the process has been developed, with the design to delegate the work to sectors of the library, which although not necessarily specialists in institutional repositories, already have complementary working practices, for example, cataloguing staff.

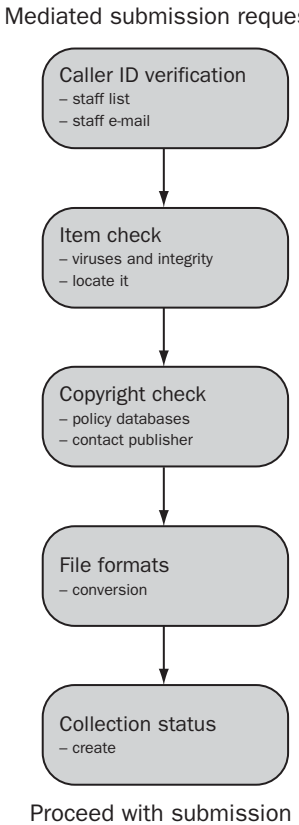
Once a call for mediated deposit comes in there are two possible strands that should be followed: *journal articles* and *theses and dissertations*. The following sections cover each of these strands in some detail.

Journal articles and other research material

Figure 7.8 shows the process for mediating the deposit of a journal article.

First, the item submitter needs to be sufficiently satisfied that the caller is a genuine member of the University of Edinburgh. For brevity, the recommended way of doing this is to check the university web pages and staff lists for the caller and to ensure they are using a recognised internal e-mail address. Next it is necessary to perform a check on the item being submitted. If the item is sent by e-mail then care should be taken to ensure that it is virus free and uncorrupted, and that any description of the item in the e-mail corresponds to the attached file. In some cases no file is sent, and only a reference to the electronic object is provided (often

Figure 7.8 **Mediated submission of journal articles**



in the form of a link to an online journal), and in these cases some effort needs to go into obtaining the item. In other circumstances it may be necessary to assist in the digitisation of a print-only resource.

In terms of whether a particular offer of submission is valid, the general policy is that if an academic thinks the material is worth putting online then it will be accepted. The rationale behind this ethos is to disseminate the university's research as widely as possible. As a member of academic staff has already been subject to a form of peer review during the job interview and selection process we automatically assign them a trusted contributor status.

The copyright check is perhaps one of the most important stages of the deposit process. By placing content online EUL is acting in the eyes of British law as a publisher, and thus can be found liable if the content disseminated is defamatory, libellous or breaching copyright or licensing terms. While some materials, such as e-theses or unpublished manuscripts do not carry such risks for the repository, the situation is more complex when we want to archive previously published materials. Often authors pass their copyright over to the publishers, or assign a non-exclusive licence, which prohibits further distribution by other parties.

For journal articles we use the following sites to find a summary of permissions that are normally given as part of each publisher's copyright agreement:

- SHERPA Romeo database: *http://www.sherpa.ac.uk/romeo.php*;
- EPrints Romeo database: *http://romeo.eprints.org/*.

Searching through these databases should quickly give you the copyright information required. Commonly, we have found that journal articles are subject to the following archive conditions set out by publishers:

- self-archiving not formally supported;
- self-archiving of pre-print permitted;
- self-archiving of post-print permitted;
  - author's own version of accepted paper;
  - publisher's version;
- self-archiving of pre-print and post-print permitted.

Additional terms and conditions for self-archiving are defined by publishers, particularly with regards to where e-prints are permitted to

be deposited. To makes things clear for depositing authors we regard ERA as:

- a non-profit, non-commercial, institutional, open-access e-print server;
- *not* an author's personal website, departmental web page or password-protected site.

Sometimes we find that publishers are not listed in either of the databases cited, or we may be dealing with other types of content such as book chapters or conference proceedings. If this is the case, then direct contact with the publisher is required. We have devised a standard form letter which can be sent to the publisher by post and e-mail, requesting permission to use the item. If the response is positive we can go on to archive the item.

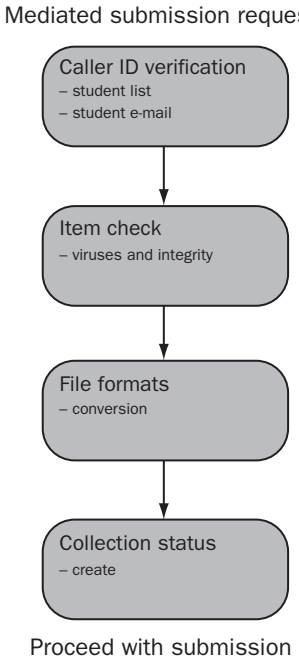
Regarding file formats, we prefer to archive PDF files above other proprietary formats, partly due to longevity and prolonged ease of access. If a file is received that is not PDF then it is converted using the appropriate tools. Most local PCs are installed with software to do this, however, sometimes more unusual file formats such as PostScript, LaTeX (or related device independent files) are supplied. If this happens then there are additional tools available for the repository staff to convert these files to PDF. When the item is archived, the source files and the PDF are archived together to provide the most options in future digital preservation efforts.

Finally, before an item is submitted it is necessary to check to which community and collection the item belongs. This is determined primarily by which academic unit the author is part of, and secondarily by item type. Theses and dissertations have specially designated collections, as discussed earlier, and need to be placed in the corresponding collection. If no community or collection exists for the item it must be created as per the ERA Administration Guide, otherwise the mediated submitter can proceed with inserting the item into the relevant collection and have it made available within ERA within a very short time.

Theses and dissertations

Figure 7.9 shows how the mediated submission process currently occurs for an e-thesis. This is effectively a subset of the steps required to archive journal articles, which we will recap here.

Figure 7.9 **Mediated submission of e-theses**



Again, it is necessary to verify the authenticity of the caller, and student lists (especially those documenting postgraduates in their final year) can be used in conjunction with verifying that the e-mail address is internally recognised. The items in these cases are usually delivered via e-mail or directly on CD to the EMG and the normal integrity checks are performed on the files. It is extremely unlikely that a print copy will need to be digitised. Note that there is no copyright verification stage, because in general the thesis is original and unpublished work. In unusual circumstances, such as thesis by research publication some action may need to be taken. As usual, conversion to PDF is ideal, although given the nature of some e-theses this is not always possible, but steps are taken to ensure that formats which meet basic preservation requirements are used. Finally, similar checks to before are carried out to ensure that the relevant collection exists for the item (taking care to adhere to the special designation requirements of the archive structure), and then the item can be submitted.

Conclusion

Through involvement with other JISC FAIR-programme funded projects we were able to develop and launch ERA within a year and a half. The repository now contains full-text e-theses, book chapters, journal pre-prints and post-prints as well as a small number of technical reports and conference papers. We have investigated and implemented revised thesis rules and regulations for the University of Edinburgh to permit and encourage e-theses. Similarly we have updated the thesis submission and management workflow to take advantage of the benefits that adopting e-theses creates. We have also delivered a report on IPR and e-theses commissioned by the JISC Legal Service to advise on the legal implications of this sort of work. Template use and deposit licences have been developed, along with advice on the FOI implications. At the same time a huge amount of community support for theses sorts of activities has been achieved via the dissemination of project findings.

For institutions worldwide one of the most recognisable outputs of the project was the development of the Tapir, which is now partly included in the general DSpace release. Meanwhile the creation of the UK e-theses core metadata set, along with our collaborating institutions has formed a good basis for further e-theses classification, storage and access. In addition, a major impact has been the provision of open access status to selected research and thesis literature; this toll-free access to students and academics is available constantly without the physical lending restrictions that are traditionally associated with published literature. In addition to the core project aims we have also addressed a number of critical side issues. The resolution of these issues, in particular IPR, proved to be of paramount importance, not just for project completion but also for the wider community.

The knock-on effects of this work confer dynamic impacts on the teaching, learning and research communities. There is an opportunity for enhanced teaching and learning in that source material such as book chapters and research articles are increasingly being made public through this repository and others like it.

The technical and cultural expertise garnered through developing and implementing ERA has been invaluable, and has been disseminated in various forms to the higher education information and library services community. This book has been one of our contributions to the community in the hope that the hard-won lessons we have learned will make this process for other institutions a much more enriched and enlightened one.

Note

1. This chapter is comprised primarily of the findings of the Theses Alive and SHERPA projects. Cited here are many of the articles and reports that were produced during this time. In addition, many project documents and presentations were consulted that were written during the project. Most of these resources can be found in one form or another on the Theses Alive project website: *http://www.thesesalive.ac.uk/*. Of particular importance are Andrew (2004a,b), Jones (2004a–f) and MacColl (2002a,b) and the contribution by Glasgow University Library (Nixon, 2003).