



DSpace and ETD-db Comparative Evaluation

**Conducted by the Theses Alive! team at Edinburgh University Library
for the Joint Information Systems Committee under the remit of the
Theses Alive! Electronic Theses and Dissertations (ETD) project.**

August 2003

Richard Jones, Systems Developer, Theses Alive!

1. Introduction	3
2. Support	3
2.1 <i>ETD-db</i>	3
2.2 <i>DSpace</i>	4
2.3 <i>Conclusions</i>	4
3. Installation	5
3.1 <i>ETD-db</i>	5
3.2 <i>DSpace</i>	5
3.3 <i>Conclusion</i>	6
4. User Accounts	6
4.1 <i>ETD-db</i>	6
4.2 <i>DSpace</i>	7
4.3 <i>Conclusion</i>	8
5. Submission Procedures	8
5.1 <i>Comparison</i>	8
5.2 <i>Conclusion</i>	10
6. Archiving	10
6.1 <i>ETD-db</i>	10
6.2 <i>DSpace</i>	10
6.3 <i>Conclusion</i>	11
7. Browse and Search	11
7.1 <i>ETD-db</i>	12
7.2 <i>DSpace</i>	12
7.3 <i>Conclusion</i>	13
8. Administration and Security	13
8.1 <i>ETD-db</i>	13
8.2 <i>DSpace</i>	14
8.3 <i>Conclusion</i>	15
9. Sundries	15
10. Conclusion	15
11. References	16
Appendix A	18
Appendix B	21
<i>ETD-db</i>	22
<i>DSpace</i>	22
<i>Comparison</i>	24
Appendix C	27
<i>ETD-db</i>	28
<i>DSpace</i>	28
<i>Comparison</i>	29

1. Introduction

We evaluated two open source packages with a view to using one for the creation of an online submission and archive system for Electronic Theses and Dissertations (ETDs) in the UK. These packages were the ETD-db written by Virginia Tech, and DSpace written in partnership between Hewlett-Packard (HP) and the Massachusetts Institute of Technology (MIT). This evaluation is performed under the remit of the JISC funded Theses Alive! project.

A direct comparison of these packages is fairly difficult as they are driven by different motivations. ETD-db is specifically designed for ETDs, containing a workspace for continued authoring of documents, and thesis-specific metadata requirements. DSpace, on the other hand, has been developed to aid the creation of institutional archives, with the emphasis far more on flexible submission workflows and potential digital preservation.

This comparison, therefore, will look at the common elements between these packages and draw conclusions on which is the best in each field. In addition, it will look at how difficult it will be to modify each of the packages to do exactly what is required by the Theses Alive! project. This analysis will be 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 has within the library and also the university itself.

2. Support

It is important when attempting to provide a reliable, widespread service that the system used to provide it is constantly monitored and modified whenever problems are found. If we are using a package written by a third party, it is very helpful if that third party is available to aid in set-up, configuration, general running and development of the software. Without this support, it is necessary to provide our own, which can be challenging and time consuming.

2.1 ETD-db

ETD-db has been developed over the previous few years by one or two developers at Virginia Tech using Perl as the server-side language. As of February 2002 development of the official release of this package ceased at version 1.7c, although it is believed that some development still goes on within Virginia Tech and it is used as their ETD submission, archive and search tool. Currently this package is the most widespread ETD package in use, with institutions all over the globe either using or evaluating it. Despite this, there seems to be little directional development, with some institutions choosing to install the “vanilla” version, while others choose to make their own changes to the system, and these are not generally easily available.

The last release of ETD-db (v1.7c) is a relatively stable piece of software, although periodically there are bugs reported. Most of the bugs are of relatively minor significance although as a whole they represent a reasonably large body of work to be done.

The future of this package seems uncertain, and it looks unlikely that a central developmental body will emerge to guide further authoring of this open source software.

2.2 DSpace

DSpace has been developed over the past few years by a team comprised of both HP and MIT technical staff using Java as the server-side language. Development is still in progress, with the acknowledgement that the system is still in its infancy, and v1.1.1 was released in August 2003. As institutional archiving software, DSpace is slowly making its mark, with an increasing number of institutions around the globe installing the package.

Most development is undertaken by the original developers and a growing technical user base is drawing suggestions for future features which can be added to the development schedule. There is still a lot to do to DSpace before it could be considered a world-class package, and bug reports arrive relatively often. In addition, the DSpace development model may be due to change, as the creation of the DSpace Federation marks a transition into more open development.

The DSpace Federation is currently a collective of institutions interested in DSpace and supportive of institutional archiving, aiming to keep on the development and use of the package and its associated ethos after the initial development funding runs out. The DSpace Federation is relatively small at present but meetings in June/July 2003 were aimed at deciding how it should evolve, and it may expand in the future.

The future of DSpace seems stable in the medium-term, although it is difficult to predict what the outcome of the federated approach will be.

2.3 Conclusions

It is clear that DSpace has far more backing and future than ETD-db, and that a cooperative approach as proposed in the DSpace Federation seems far more likely to produce useful, powerful and interoperable software. It is possible that DSpace suffers from more bugs at present than ETD-db, but this is tied to the fact that DSpace offers more functionality than ETD-db.

There is certainly no problem with the choice of language for either of the packages. Both Perl and Java programmers are common, and both languages are well respected and powerful.

Due to the factors considered here, there is a choice between:

- 1) Having a relatively stable, but relatively basic, package which is designed specifically for ETDs, but which would require a large commitment to patching and supporting.
- 2) Having a powerful but under-development package which is not specifically for ETDs, but which looks like it will be part of a global community for some time to come.

3. Installation

Although not such an important consideration in the long-term, it is useful to know how difficult to install, and thus how difficult to maintain at a systems administration level, each of these packages is. The installation will be dependent not only on the package itself, but also the additional software that each package requires in order to operate.

Both packages are delivered via a web interface, and both make the assumption that this will be served by the Apache web server.

3.1 ETD-db

ETD-db requires that Perl and MySQL are installed on the server. Perl (a scripting language) is native to most Linux and Unix installs, and MySQL (an open source database package) is also very common. In addition to the standard Perl installation, it is also necessary to install additional “Perl Modules” which enhance the functionality of the language. The level of skill required to do the prerequisite installation is that of a reasonably experienced systems administrator.

Following the prerequisites installation, the install of the ETD-db is a relatively straightforward matter, although we did encounter a bug in the v1.7a code that needed to be fixed before the software worked. In addition, configuration of the web server is fairly extensively required, and it is necessary to create a number of new users as ETD-db uses the web server to provide the security to the administration areas of the system.

3.2 DSpace

DSpace requires a number of Java elements to be installed on the server before it will work. These include Tomcat, which is a Java Servlet Engine (it provides the interpreter for pages written in Java in the same way that Perl provides the interpreter in ETD-db), a number of Java code libraries that the software relies upon, a Java compiler (Ant) which is required to compile the code, and PostgreSQL (an open source database). It is recommended that

DSpace be installed on a Linux or a Unix machine; the installation is possible on Windows, but there is not much documentation.

It is well known that the installation and configuration of Tomcat is particularly difficult, although the rest of the prerequisites are relatively straightforward. The level of skill required to do the prerequisite installation is that of an experienced systems administrator.

Following the prerequisites installation, the install of DSpace can be a little confusing but is quite straightforward. Configuration of the web server is fairly extensively required, especially if you wish to employ a Secure Socket Layer (SSL). In addition, DSpace itself must be carefully configured before it will function correctly.

After DSpace has been installed there is also the requirement of installing the Handle Server, which allows for the use of persistent identification of all of the items that will eventually fill the database. The installation of this comes in two parts, and it is necessary to register your institution with a third-party service provider (for free).

3.3 Conclusion

There is little doubt that ETD-db is easier to install and configure than DSpace, but once again we must take into account the additional extras that come with DSpace that require this extra technology. If the extra functionality is worth the effort then the maintenance of this package may be desirable. There is much talk about creating a better installation method for DSpace and it seems likely that there will be one in the not-too-distant future. This does not solve the problem of the installation of Tomcat, which has been a major factor in the time taken to make DSpace work.

4. User Accounts

Both systems require that submitters have their own account before submitting any items. We are particularly interested here in how authentic and secure the sign-up procedures are, and how useful the user accounts are once they are active.

Data for this section is available in Appendix A.

4.1 ETD-db

ETD-db builds the user account at the same time as it builds the main submission, and hence there is no hard barrier between the user account and the submission. A username and password are requested on the registration page, and this is sufficient to open an account containing the user's thesis.

The user is then moved on directly to create a new “Main Record”, in which email address, name, and department are requested.

Note that the registration page itself does not validate the user’s identity and anyone who can see the registration page is capable of creating an account without any problems.

It is possible to set up the system to run under an SSL, and this makes the security of data being transferred to and from the submitter’s machine very secure. The user’s password is kept encrypted in the database using “crypt”, a Linux and Unix native encryption package, which is sufficient provided that the server itself is secure from attack.

Maintenance of the user’s information is done via Perl’s cookie-handling facilities, which require that the user allow cookies to be written onto their computer. We encountered a problem logging in to the system due to a bug in the way that the cookie-handling is done by the package, resulting in the necessity to log in more than once in a row before being granted access. In addition we discovered that the session maintained by the cookie runs out after 30 minutes, which may cause problems for users doing extensive authoring. There is also an additional security problem that arises from the way that sessions time-out, which allows other users to reactivate dead sessions.

There is no easy way to clean up dead user accounts in ETD-db, so it is possible that after a system has been in use for some time it will have a lot of excess data in the database.

In general, ETD-db’s user account system is not particularly good, and is too closely integrated with the user’s submission. It suffers from a number of potential security holes that need to be fixed before a live service could be provided.

4.2 DSpace

DSpace requires very minimal information about the user in order to register an account: only an email address and password are required. The user is then emailed after sign-up with an authentication certificate, which they must then present back to the system in order to have their account activated. On arrival at the site, it requests the user’s name, and contact telephone number. This reduces the chance of multiple accounts for one user, and also prevents people being signed up for an account in error. There is no specific validation, though, as to who can sign up for an account, and anyone who can see the registration page can register for an account.

It is recommended that you set up a Secure Socket Layer that makes the data being transferred to and from the user’s computer secure. The user’s password is kept encrypted in the database, which is sufficient provided that the server itself is secure from attack.

Maintenance of the user's information is done using a Java session management package, which requires that cookies be enabled on the user's computer. We have encountered problems logging out of the system using the web browser Opera (v7.x), the reason for which has not been determined – this could potentially cause a security breach.

The option is available to the administrator to remove or modify users' accounts, making maintenance possible but not particularly easy, and additional tools to speed up this job would be a useful addition.

In general, DSpace's user account system is quite traditional and intuitive, although it lacks the facility of a username, although this is not necessarily a requirement. There are no obvious security holes, although the problem with Opera ought to be investigated by the developing body (this bug has been reported).

4.3 Conclusion

The DSpace approach to user account is more common than the ETD-db approach. Nonetheless, in some circumstances it might be necessary to have more information about the user available in their account details than DSpace holds. The main shortcomings of the ETD-db method are in the security issues that exist, as it is feasible that only one submission per user is required in an ETD system. DSpace does not suffer from such severe issues, and the any bugs found in the system should be fixed in the development process.

5. Submission Procedures

Here we are not specifically interested in how well-organised the submission procedure is, although it will be valuable if we see a procedure that is logical and well laid out. Instead we are mainly concerned with what metadata is/can be collected by the system during submission. This comparison will also attempt to take into account the files that the user can upload containing the actual content of their thesis.

This comparison will draw heavily on the data provided in Appendix B.

5.1 Comparison

Examining the tabulated data, the first observation is that DSpace collects a lot more information at submission than ETD-db. This includes fields such as the uploaded file format and the user's telephone number. This works in the other direction too, though. For example, ETD-db collects fields such as Department and Defence Date (Viva). The question, then, is whether either,

both or neither collects enough information, and whether the data that is collected is extensible or flexible in any way.

The tabulated data explains which fields are analogous in each system, where discrepancies arise and some explanation as to why and how each system deals with that difference. For example, where ETD-db collects the user's department, DSpace includes that information implicitly by the item's location within a community and collection.

Here we will briefly look at the major differences between the metadata collected, before considering how metadata is stored and the pros and cons of each system.

ETD-db is designed specifically toward theses management, so it collects the Defence or Viva date for each thesis. No such analogue exists within DSpace and this may be something that needs to be addressed. Meanwhile, where ETD-db requests an availability level for the thesis, DSpace offers a far more sophisticated way using an authorisation policy system built into the administration area. ETD-db also collects information regarding committee members who will oversee the thesis as it is submitted and marked. Again, DSpace provides no analogue to this because it is not geared toward theses. In terms of document management, though, DSpace has a registry of file formats which it recognises and will store this as part of the metadata. This allows the administrator to be able to control the file formats that are deposited into the archive. DSpace also permits multiple authors per document, and takes users first and last names in separate fields (to aid browsing by author), whilst ETD-db takes the name in only one field.

DSpace and ETD-db take file uploads via the web interface, but only ETD-db provides the administrators with the option to give FTP login to users to upload files that way. The reason that DSpace does not use FTP is due to the way that it links the files and the metadata together. Both systems allow multiple files, although DSpace adds the option to attach descriptive text to each file that is uploaded.

Finally, the additional fields that DSpace collects allows for a significant array of possible submission types that may not necessarily be relevant for theses, such as an identifier for the item (e.g. ISSN). Additionally, DSpace archives the copyright licence at the time of writing in with the thesis so that it may be preserved. ETD-db's analogue is to store the copyright notice in the database with the metadata. The subtle difference in these two approaches may be important.

DSpace employs qualified Dublin Core to identify the stored data. This is a well-established basic metadata standard, and using qualifiers it is possible to extend the basic information to be relevant to many types of digital object. This method overcomes a number of extensibility and flexibility problems that can arise when storing defined data. ETD-db, conversely has a number of pre-determined database fields (defined by the ETD-ML standard created by Virginia Tech) which require programmer intervention to alter.

5.2 Conclusion

It is clear that DSpace has a more comprehensive metadata collection process and that it stores this metadata in a more flexible manner. Due to the customisable nature of the Dublin Core registry within DSpace and the option to modify the submission interface (although this is a job for a programmer), DSpace will take any data that can be represented within the qualified Dublin Core. ETD-db has no such flexibility and future changes in metadata schemas could cause significant problems.

Overall, the DSpace approach is more flexible, and the result is a package that may be customisable for theses as well as many other types of digital object, although ETD-db is geared specifically to ETDs with its standard metadata set.

6. Archiving

Specifically, both of the packages are, at heart, designed to make the archiving of digital resources quicker and easier, but these are not the only requirements. We wish to make the archive available via the OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting), which allows the archive to be searched from outside the institution by special “Harvesting” servers. Additionally, we would like to see an archive that is preservable, stable and secure. In this section, we will be looking to see how each package addresses the archiving issue.

6.1 ETD-db

ETD-db undertakes archiving in a straightforward manner. All files are stored in a basic directory structure; the area that they are in determines the security level applied to the item. If the security settings are changed, then the item is physically moved to another directory.

The metadata associated with the item is maintained within the database for as long as the item remains in the archive.

ETD-db comes with the facility to expose the archive via the OAI-PMH, but in v1.7a only v0.9 of the protocol has been included. Since the current version is 2.0 it would be necessary to make a major upgrade to this area of the system.

6.2 DSpace

DSpace has made an effort to include elements of digital preservation into the archive. In fact, advice was sought from Edinburgh University Library regarding digital preservation while the system was under initial development. To this end, all items have a kind of “wrapper” in which parts of the relevant data are stored. This includes all the files, and the copyright licence. The metadata is maintained in qualified Dublin Core format in the database for as long as the item remains in the archive.

Security settings for the archive are dealt with via the authorisation policy tool, and the security of the archive then depends upon the way that the DSpace Administrator configures the policies for each community, collection, and item.

DSpace also comes with OAI-PMH v2.0 built in, allowing for immediate compatibility with the more advanced features of this standard. We would expect that future versions of DSpace will have the most up-to-date version of this protocol.

6.3 Conclusion

The DSpace archive is perhaps more geared toward digital preservation, although this issue is still very much in debate and it may be discovered that their method is not necessarily the right way to go about solving the problem.

Moving files around may be a weak spot within ETD-db, i.e. the more you move files, the more chance there is of them being lost or corrupted. This method of providing security is not the best way, although it is much simpler to use and implement than the DSpace approach.

Storing the files in a standard directory structure, as advocated by ETD-db, makes the files far easier to access without using the web interface. DSpace requires you to use an “export” facility in order to remove files from its internal archiving structure.

7. Browse and Search

A defining feature of any digital object archive is how easy it is to find what you want. Since each of the systems we have evaluated are designed to stand alone, it is necessary that their native search and browse interfaces be sufficiently good to deal with what could be many thousands of records. Over time, the total number of searchable records could expand without bound.

Browsing and searching are the two main methods one would expect to use to interact with an archive. Browsing relies exclusively on the hierarchical way in which the data is structured, whilst searching can transcend these boundaries and consider every item in the archive for potential return in any

result set. As such, they are both important and fundamentally different ways of finding information.

7.1 ETD-db

ETD-db has no concept of internal directory structure. It is possible to browse the entire archive and to sort the results in a number of ways, but it is not organised such that it is possible to view all items within one department or discipline.

The search within ETD-db is essentially a basic keyword search with Boolean operators and specified search fields (such as “author” or “title”). This builds a basic search query which returns the results ordered by one of the fields, but not by relevance. It has the option of having an effectively infinite number of search boxes that can be used in any one search, allowing for high precision searches, which is a distinct advantage.

Also, the OAI-PMH is supported as a remote search facility, but v0.9, which comes packaged with ETD-db v1.7c, cannot take advantage of the most recent developments within this standard.

7.2 DSpace

DSpace has a highly restrictive directory structure, split into Community and Collection levels. All items reside at the Collection level and it is possible to browse the complete list of items within any collection, ordered by title, author (surname) or submission date.

The DSpace “Advanced Search” feature is a basic keyword search with Boolean operators and specified search fields, in much the same way as ETD-db. However, it is limited to only three search fields and changing this requires programmer intervention. DSpace also permits searches to be confined to specific Communities in the advanced search region. Irritatingly, it is necessary to browse to the collection level if you wish to perform a search there. The benefit of browsing the collection home page is that an additional set of browse and search options for that collection and its parent community become available.

It is worth noting that DSpace are discussing the possibility of integrating the Google search engine into the product. This would be a valuable addition to the system. Additionally, the University of Kentucky are looking into combining DSpace with Endeavour’s EnCompass system, which is currently also being deployed at Edinburgh University Library. This would allow DSpace to be cross-searched alongside other institutional resources from one interface.

7.3 Conclusion

Although DSpace's data hierarchy could do with being considerably more flexible, it is superior to that offered by ETD-db. The search features are not radically different, although the DSpace search is more embedded in the system as a whole (although this is not quite as smooth as it might be), but the ETD-db search is expandable. Neither system return results according to the relevance.

Overall, the search is adequate in both systems for a small quantity of data, but DSpace may be working toward building a search system which can cope with large quantities just as well as with small quantities.

8. Administration and Security

Aside from the features that each package provides to front-end users, each package also provides administrative features for service providers and administrative staff. These include some workflow facilities that allow certain users to perform tasks on submitters' items, as well user administration tools. The full spectra of available administrative options will be discussed in the sections below.

In addition we will see how the security in each package functions at this level, and consider the best way of addressing the security issues that arise.

This section draws on data in Appendix C.

8.1 ETD-db

In order to access the three ETD-db administrative functions (**Review Submitted ETDs**, **Manage Available ETDs** and **Manage Withheld ETDs**), it is necessary to have the specific login details for each area. In this system there is only one username and one password needed to access each area, which means that it is impossible to give an administrator access to a small subset of the available options. This may be important, as we will later see.

Review Submitted ETDs gives the administrator the option to browse the list of all ETDs currently in the submission process and to perform all of the actions that the submitter can perform on the item. Effectively this provides a "workspace" where a student and a supervisor can collaborate and communicate on the thesis. From here, the administrator may then also approve the thesis for inclusion into the archive in either "available" or "withheld" status.

Manage Available ETDs provides the facilities to administer the ETDs that actually appear in the archive and which are exposed via the web interface in browse and search options. Primarily, at this stage this allows the

administrator to remove the item or move it into the “withheld” section of the system.

Manage Withheld ETDs provides similar functionality to that of **Manage Available ETDs**, but with the option to move items into a status of “available”.

The main drawbacks of this method are that there is no way of providing a single supervisor to a single submission, and that all supervisors with the permission to access the **Review Submitted ETDs** section can see the theses of all the students who are currently submitting. The list of theses in progress may also be quite large, so the supervisor will be presented with a list of potentially hundreds of theses upon login. It is also assumed that it will be the supervisor who will eventually agree that the metadata for the item is correctly written and that the thesis is complete, marked and ready to enter the archive. In general this will not be the case.

The advantage of this system is it applies Apache (the web server) security to the directories that are restricted. This method of securing directories is well developed and known to be reliable, ensuring that all content is genuinely secure. The basic structure of a sensible administration system is here, but a major security overhaul would need to be performed before a live service could be provided, and this may include the addition of a policy system for within the **Review Submitted ETDs** section.

8.2 DSpace

It is clear from looking at the tabulated data in Appendix C that DSpace has many administrative options. DSpace also splits its administrative facilities into two parts: **Workflow users** and **DSpace administrators**. The fundamental difference between these two sections is that workflow users may only perform their actions within the constraints of the workflow system in DSpace. These duties include reviewing submissions after, and only after, the author has submitted them for consideration. In reality, DSpace has three well defined “workflow steps” which groups of individual users of the system can be assigned to in order to perform reviewing and administrative actions on submitted items. DSpace administrators, on the other hand, have access to a large set of tools located in a different area of the system, allowing them to administer user accounts and user groupings, create and modify system policies on items, collections, communities and users, and various other system maintenance tools.

Login to the administrative area is provided through precisely the same system as users log in to the system, and the differences in the behaviour of the accounts is purely down to the policies applied to the user account (so DSpace can have multiple administrator accounts for example).

This method has no real drawbacks and is a more consistent method of system design. Its advantages lie in the fact that there is only one type of user and that each user’s behaviour can be modified, even over time if

necessary. It is true to say, however, that the policy administration within DSpace is confusing and could be a potentially large consumer of administrators' time. It should be noted that in order to withhold items from public access, whilst still maintaining them in the database, it is necessary to perform some fairly intricate administration.

8.3 Conclusion

The level of customisation available within the DSpace administrative area puts it far ahead of ETD-db in this respect. Although the customisation is not quite as sophisticated as we might want, it is only necessary in a few cases to delve into the code itself to make changes. It is a limitation that the well-defined workflow could stand in the way of creating the steps which institutions the world over could fit into their current working methods.

ETD-db is designed to allow the easy authoring and supervision of ETDs, and the tools that it provides for this purpose are straightforward and relatively effective. DSpace provides none of this functionality and would need to have it added before a true live service could be provided.

We have also seen that when withholding items, ETD-db provides far simpler and more effective (although potentially flawed) functionality than DSpace, which is another shortcoming that would need to be addressed prior to the launch of a live service.

Overall the methodology employed by DSpace is superior to that of ETD-db, and many of the shortcomings of the DSpace system can be reasonably solved. Conversely, the work required to bring the ETD-db up to the same standard in all other respects is fairly extensive and may require rewriting of much of the software.

9. Sundries

This section is dedicated to briefly listing a number of other options/facilities available within the systems, which are of passing interest but not of any particular sway in this comparative evaluation.

- DSpace provides an option for the user to subscribe to a particular collection, with notification by email when new items are submitted.
- Both systems have embedded "Help" files covering most necessary areas.
- DSpace lists a selection of the latest submissions into each community or collection when you visit the home page.

10. Conclusion

In the majority of comparative areas that we have investigated we see that DSpace is a clearly ahead of ETD-db. It is a well-supported package with a

future that is being planned now, while ETD-db is very much at the end of its development cycle. DSpace is far more functional with regards to essential features such as security and administration and this sort of infrastructure is important for any piece of software, no matter what additional features are available.

There are also areas where there is no great distinction between the packages. Both have similar browse and search facilities, and with the uncertainty of the evolution of digital preservation, their archiving methods could be difficult to choose between. Likewise, their submission procedures are adequate, and similarly difficult to modify. It is worth noting, though, that at each stage in which these elements have been considered, DSpace has the edge over ETD-db, often because of the solidity of its infrastructure and its potential to be developed to fix any shortcomings.

It is worth considering that ETD-db is designed specifically for ETDs, whilst DSpace's support in this regard is fairly generic. The questions that must then be answered are as follows:

- 1) How hard would it be to add thesis support, as we require it to DSpace?
- 2) How hard would it be to bring ETD-db up to the standard that we would require for a live service?

During product evaluation both questions have been considered. The results indicate that bringing ETD-db up to standard would require extensive bug fixing as well as major feature upgrades to improve data structuring, security, and overall behaviour of much of the system. Creating ETD functionality in DSpace, however, is not only one of the possible features in the developers' plan, but mainly requires minor modifications to the system, with some additional software written to provide more functionality. The estimate for not only the ease of doing this, but also the long-term support of our modifications, suggest that DSpace would provide a better core system for Theses Alive!

The future of ETDs and of archiving and searching in general depends on institutions being able to deliver top quality services, with a high degree of interoperability. This means, among other things, that systems must continue to be developed and they must be able to handle all sorts of different types of digital object. We believe that DSpace will fulfil these requirements to a higher degree than ETD-db, and will continue to improve in this way in the future.

11. References

- Ant Java Compiler from Apache: <http://ant.apache.org/>
- Apache Web Server: <http://www.apache.org/>
- DSpace: <http://www.dspace.org/>

- Dublin Core Metadata standard: <http://dublincore.org/>
- Edinburgh University Library: <http://www.lib.ed.ac.uk/>
- Endeavour EnCompass: <http://encompass.endinfosys.com/>
- ETD-ML Document Type Definition: <http://etd.vt.edu/etd-ml/>
- Handle Server, CNRI: <http://www.handle.net/>
- Java at Sun Microsystems: <http://java.sun.com/>
- Joint Information Systems Committee (JISC): <http://www.jisc.ac.uk/>
- MySQL Database: <http://www.mysql.com/>
- Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH): <http://www.openarchives.org/>
- Perl scripting language: <http://www.perl.org/>
- PostgreSQL Database: <http://www.postgresql.com/>
- Tomcat, Apache Jakarta project: <http://jakarta.apache.org/>
- Theses Alive! project at Edinburgh University Library: <http://www.thesesalive.ac.uk/>
- University of Edinburgh: <http://www.ed.ac.uk/>
- Virginia Tech's ETD: <http://etd.vt.edu/>

Appendix A

Comparison of the information that ETD-db and DSpace collect in order to create a user account.

ETD-db	DSPACE	Comments
Username		DSPACE does not implement usernames in user accounts. Instead it relies on email addresses only.
Password	Password	
Email	Email	
Name	FirstName LastName	DSPACE collects the users first and last names separately
Display Email		
Department		This is not part of the DSPACE sign up, but is implicit in the collection to which the user is submitting.
	Contact Telephone	

Appendix B

Documentation and comparison of the metadata records collected by default in ETD-db and DSpace during item submission.

ETD-db

Stage	Heading	Field	Key for Comparison	Description
1	Add New Main Record	Email	1.Email	User's email address
		Name	1.Name	User's name
		Display Email	1.Display Email	Should email address be published in public documents?
		Department	1.Department	The department in which the user is studying
		Degree	1.Degree	The degree for which the user is studying
		Document Type	1.Document Type	Type of ETD (eg PhD, Dissertation)
		Defense Date	1.Defense Date	Date of Defense/Viva date
		Title	1.Title	Title of ETD
		Keywords	1.Keywords	Keywords for searching for ETD
		Abstract	1.Abstract	Abstract of ETD
		Availability	1.Availability	Desired availability level for submitted work (eg unrestricted).
2	Add Committee Information	Name	2.Name	Committee member's name
		Title	2.Title	Committee member's title
		Email	2.Email	Committee member's email address
3	Add File Information	Upload Files	3.Upload Files	Upload any number of files containing ETD data.

DSpace

Stage	Heading	Field	Key for Comparison	Description
1(a)	Describe your Item	More than one title	1.MultipleTitles	Check if your submission has titles in, for

				example, more than one language.
		Distributed before	1.Published	Check if the submission has been published or publicly distributed prior to submission.
		More than one file	1.MultipleFiles	Check if there is more than one file to be submitted.
1(b)	Describe your Item	Authors Last Name	1.AuthorLastName	The Author's surname
		Authors First Name	1.AuthorsFirstName	The Author's first name
		Title	1.Title	The title
		Series Number	1.Series	The series the article belongs in
		Report Number	1.Report	Report number
		Identifier	1.Identifier	pre-existing unique identifier
		Identity Numbers	1.Identity	Identity numbers
		Type	1.Type	The type of submission (e.g. thesis).
		Language	1.Language	The language of the main content of the submission.
1(c)	Describe your Item	Keywords	1.Keywords	The keywords for your submission
		Abstract	1.Abstract	The abstract in plain text for your submission
		Sponsors	1.Sponsors	The names of sponsors and/or funding codes associated with the submission.
		Description	1.Description	Additional descriptions or comments to be associated with the submission.
2(a)	Upload a File	Document File	2.File	The file to be uploaded
2(b)	File Uploaded Successfully	Wrong Format	2.WrongFormat	Takes you to section 3(c)
		Wrong File	2.WrongFile	Takes you to section 3(a)
		Show Checksums	2.Checksums	Displays the checksum information used by DSpace to verify file integrity.
2(c)	Select File Format	File Format	2.Format	Allows you to have DSpace auto-assign a format, to pick one from a list, or to enter one

				manually.
3	Verify Submission		3.Verify	Displays current information and allows submitter to change any part of the metadata.
4	Grant Distribution Licence	Grant Licence	4.Licence	User must choose to grant or not the licence

Comparison

VT-ETD	DSpace	Comments
1.Degree	1.Type	Dspace does not specifically look for degree type.
1.Document Type	1.Type	
1.Defense Date		Required for Viva date, and missing in DSpace
1.Title	1.Title	
1.Keywords	1.Keywords	
1.Abstract	1.Abstract	
1.Availability		Provides the security in ETD-db, which is done using authorisation policies in DSpace
2.Name		ETD-db allows for committee members.
2.Title		ETD-db allows for committee members.
2.Email		ETD-db allows for committee members.
3.Upload Files	2.File	Methods of handling multiple files are subtly different but both acceptable.
3.Name	1.AuthorFirstName, 1.AuthorLastName	Dspace collects first and last names separately, and allows for multiple authors
	1.Published	Generally not relevant for ETDs, although some parts of theses may have been previously published.
	1.Description	

<i>Implicit</i>	1.MultipleFiles	Dspace needs to know whether there will be multiple files, ETD-db assumes that there will be an arbitrary number.
	1.MultipleTitles	Dspace permits for multiple titles on submissions
	1.Series	
	1.Report	
	1.Identifier	
	1.Identity	
	1.Language	
	2.WrongFormat	
	2.WrongFile	
	2.Checksums	Dspace provides a quick way of verifying document integrity over time.
	2.Format	Document Type/Format may be determined automatically in DSpace.
<i>Implicit</i>	3.Verify	
	4.Licence	

Appendix C

Documentation and comparison of the administrative facilities available in the secure areas of ETD-db and DSpace.

ETD-db

Option	Name	Key	Description
1	Awaiting Approval List	1.Awaiting Approval	List of all files currently awaiting approval.
2	View Record	2.View Record	Allows the administrator to view all the details for the record, and provides access to the following features: Add Files, Change Availability, Add Advisors, Add Notice.
3	Modify Record	3.Modify Record	Allows the administrator to modify all of the details of the user.
4	Change Availability	4.Change Availability	Allows the administrator to change the security settings of the document
5	Add Committee Member	5.Add Committee	Allows the administrator to modify committee members for the submission
6	Add/Update Files	6.Update Files	Allows the administrator to add or remove files.
7	Send Notice to User	7.Send Notice	Sends a note to the user's email address and ETD-db account.
8	Remove	8.Remove	Allows the administrator to remove the record completely.
9	Approve	9.Approve	Allows the administrator to approve the submission

DSpace

Option	Name	Key	Description
1	Edit collections and communities	1.Collections	Allows the DSpace administrator to add/edit/remove communities and collections within the system.
2	EPeople	2.EPeople	Allows the addition of new accounts to the system,

			without sign-up being required.
3	Group Editor	3.Group	Allows the addition and configuration of user groups for different policy sets.
4	Edit/Delete Item	4.Item	Allows items in the archive to be deleted, or to have their metadata altered.
5	Dublin Core Registry	5.DC	Edit the elements of the qualified Dublin Core
6	Bitstream Format Registry	6.Bitstream	Manage the known/supported “bitstreams”
7	Workflow	7.Workflow	View a list of all currently active submissions, with the option to remove them.
8	Authorisation	8.Authorisation	Manage all system policies
9	Tasks In Pool	9.Pool	View all submissions that are in the workflow pool for the workflow group you belong to.
10	Owned Task	9.Task	View all tasks that you as an administrator are responsible for (taken from the Task Pool).
	Approve	9.Approve	Approve submission to move on to next workflow position.
	Reject	9.Reject	Return item to the submitter for alteration
	Edit Metadata	9.Edit	Edit the metadata for the item (see Workflow Operation for more details).
	Commit to Archive	9.Commit	Commit the submission to the archive (see Workflow Operation for more details).

Comparison

VT-ETD	DSpace	Comments
1.Awaiting Approval	9.Pool	
2.View Record	10.Task	

3.Modify Record	10.Edit	
4.Change Availability	8.Authorisation	DSpace uses system policies to do site wide administration of what users may see.
5.Add Committee		DSpace does not support committees
6.Update Files		DSpace administrators may not add files to the submission.
7.Send Notice	10.Reject	These are not quite similar, as Send Notice is more of a communication tool, but some of the elements are the same.
8.Approve	10.Approve, 10.Commit	
9.Remove	4.Item	
	1.Collections	DSpace allows for collections of papers in specific archives.
	2.EPeople	DSpace employs EPeople as a concept to allow computers access to the archive.
	3.Group	DSpace allows users to be grouped for policy admin
	5.DC	DSpace allows for administration of a qualified Dublin Core
	6.Bitstream	DSpace allows for admin of Bitstream formats supported/known.
	7.Workflow	DSpace has a workflow that goes beyond the standard authoring procedure in ETD-db