



## Project Document Cover Sheet

Project Information			
<b>Project Acronym</b>	IncReASe		
<b>Project Title</b>	Increasing repository content through automation and services		
<b>Start Date</b>	01/07/2007	<b>End Date</b>	31/12/2008
<b>Lead Institution</b>	University of Leeds		
<b>Project Director</b>	Bo Middleton		
<b>Project Manager &amp; contact details</b>	Rachel Proudfoot Edward Boyle Library University of Leeds LS2 9JT 0113 343 7067 r.e.proudfoot@leeds.ac.uk		
<b>Partner Institutions</b>	University of Leeds, University of Sheffield, University of York		
<b>Project Web URL</b>	<a href="http://EPrints.whiterose.ac.uk/increase/">http://EPrints.whiterose.ac.uk/increase/</a>		
<b>Programme Name (and number)</b>	Repositories and preservation programme (04/06)		
<b>Programme Manager</b>	Andrew McGregor		

Document Name			
<b>Document Title</b>	<i>Research publication metadata. Where is it? How useful is it? Can we import it?</i>		
<b>Reporting Period</b>	<i>for progress reports only</i>		
<b>Author(s) &amp; project role</b>			
<b>Date</b>		<b>Filename</b>	
<b>URL</b>	<i>if document is posted on project web site</i>		
<b>Access</b>	<input type="checkbox"/> Project and JISC internal		<input type="checkbox"/> General dissemination

Document History		
Version	Date	Comments
0.1	24/10/08	For internal circulation and comment
0.2	26/10/08	For internal circulation and comment
0.3	06/11/08	To Steering Group for comment
1	10/11/08	Public

# IncReASe: Increasing repository content through automation and services

---

## **Research publication metadata. Where is it? How useful is it? Can we import it?**

October 2008

*Rachel Proudfoot, Project Manager*

*Archana Sharma-Oates, Software Developer*

*Beccy Shipman, Project Officer*

### Key findings

- A number of metadata sources (external and internal to the institution) are likely to be available to repositories.
- There is potential for bulk population of repositories (at least with metadata) from these sources.
- Sources of metadata are often imperfect; repositories need to make a realistic assessment of the resource needed to improve said metadata and decide whether the added value justifies the cost of the resource. Repositories need to think about how the metadata will be re-used.
- Repositories will need to consider whether speed of dissemination trumps quality of metadata.
- Bottlenecking is a constant danger where metadata verification takes place before material is made openly accessible.
- There are ambiguities around what reuse of metadata is permitted if metadata has been sourced from a commercial database.
- Authority files for author names and journal titles would, of course, increase metadata consistency.
- Tools to work in conjunction with bulk metadata ingest (e.g. identifying empty fields or potentially anomalous metadata) could help improve metadata quality.
- It would be useful for harvesting services (e.g. Intute Repository Search) to issue analysis reports for harvested system against an agreed set of metadata quality criteria to help us measure, monitor and improve metadata quality. Ideally any analysis service would provide practical guidelines to achieve improvement (where utility warrants). The NZ KRIS service provides error alerts via RSS feed (Nichols et al 2008).
- A combination of manual and automated metadata creation is likely to be required.
- Clear good practice metadata guidelines would be helpful.

## 1. Introduction

Researchers across the White Rose consortium have requested that White Rose Research Online (WRRO) hold not just very recent research but their “back catalogue” of publications.

Some researchers have pointed to metadata sources which might be used to populate the repository in an automated way. Indeed, obtaining basic bibliographic data for research outputs is sometimes assumed to be a trivial activity – if only we could “crawl the web” or harvest metadata in an effective way, we could populate the repository with relative ease. It is not surprising researchers are reluctant to re-create metadata which is readily available – or perceived to be readily available – from other sources. Ready sources of metadata are also likely to be of great interest to modestly resourced institutional repositories looking for time saving measures and “quick wins” for repository population.

This report summarises our investigation into several potential metadata sources and our use of various import “plug-ins” which are available to us as part of the EPrints (3.0) repository software. The report is written from a “lay” repository staff perspective: currently there are no “metadata specialists” involved in the White Rose Research Online (WRRO) ingest processes. In particular, we have been interested in addressing the following questions:

- (i) what sources of descriptive metadata are available to us?
- (ii) what is involved in importing or harvesting metadata from these sources?
- (iii) how satisfactory is the imported metadata and what can/should be done to improve it?
- (iv) will population with older metadata improve the deposit of older and, crucially, new research outputs?

The original ingest model for WRRO envisaged author self-deposit (of metadata and relevant files), with improvements to metadata made by repository staff. However this model has not yet been widely adopted by researchers. Thus, repository staff and some administrators within research centres have been the main creators of metadata within WRRO. Predominantly, single records have been created in response to authors’ requests to deposit individual items. This has led to a reasonably high standard of metadata, with most records having relatively full publication information. Can levels of metadata quality, consistency, accuracy and completeness be maintained or improved if we harvest metadata in bulk from a number of different sources? Where metadata is created by hand by authors or non-library administrators, what quality issues occur and how can these be addressed?<sup>1</sup>

## **2. Self-archiving / proxy archiving by departmental staff**

In our experience, the quality of self-archived records varies hugely. Some are excellent but, typically, manually created metadata from authors or proxy depositors in departments can be:

- Inconsistent: e.g. forms of author names, variations in journal title formats – especially abbreviations
- Incomplete: e.g. ISSN, issue number
- Inaccurate: e.g. incorrect pagination, incorrect file appended

Currently, all ingested items are deposited to the holding “buffer” where they are checked by repository staff and enhanced / corrected as far as possible. In effect, this means every item

---

<sup>1</sup> We are aware that we will need to address administrative and technical metadata - but have placed these outside the scope of this report, concentrating instead on descriptive metadata. WRRO describes research outputs using simple Dublin Core. The limitations of simple Dublin Core are well documented elsewhere (e.g. Allinson et al 2007).

must be reviewed in some detail; a time-consuming task. The potential benefits of complete, accurate metadata need to be weighed carefully and realistically against the staff resource required to achieve this and against any consequent delay in making research openly available. Perhaps the most fundamental consideration is what the repository is for and how the data within it may be reused.

EPrints software has the capability of generating drop-down, autocomplete lists. These are not yet deployed in WRRO. However, their introduction should help improve consistency. Certainly a controlled list of journal titles would be helpful. It is difficult to see how we can standardise author data so that each author is identified unambiguously. It would be helpful to be able to connect publications which are by the same author, whether or not the name form is standardised. It is likely this will not be solved until a system of author identifiers is adopted - for example, as a result of the Names<sup>2</sup> project.

Where items have a DOI, it could be helpful to develop a “check against CrossRef” tool to highlight any differences between manually entered metadata and that held in CrossRef in order to identify omissions or errors. However, metadata in CrossRef is not necessarily complete – in particular, scant author data is held. In addition, research may be deposited prior to publication and prior to DOI assignment. From release 3.1, EPrints will have better reporting capabilities including the identification of long-standing *In Press* items. This should help identify records likely to need updating with publication metadata and DOIs. Again, some use of CrossRef – perhaps feeding the *In Press* metadata to CrossRef’s “Article title” or “Automatic Parsing” search – may be worth investigating.

### 3. Reusing pre-existing metadata

#### 3.1 Importing via CrossRef using DOI

EPrints software ships with a DOI import capability. DOIs can be cut and pasted (individually or as a list) into the import box.

Assuming the DOI has been registered (not always the case – particularly for recent works), using the DOI will populate:

- Article Title
- Journal title
- Volume
- Number
- Start page (not complete range)
- Year (not full date)
- DOI (in EPrints “Identification Number” field)
- Official URL field (link to the published copy using <http://dx.doi.org/doi>)
- ISSN (no spaces or hyphen)

The major drawback with CrossRef is lack of author data.

Some of our test imports have pulled in very little data from CrossRef– e.g. the DOI field is populated but all others are empty. At the moment, EPrints records a “successful” import from DOI, even where most fields are blank. As far as we can tell from our testing, the EPrints DOI plug-in assumes the imported item is a journal paper; this may not always be the case (e.g. papers in Lectures Notes in Computer Science may be better described as book chapters).

---

<sup>2</sup> <http://names.mimas.ac.uk/>

There is definitely scope for time saving through bulk DOI import, but it should be realised that:

- there is sometimes a delay in DOI registration and, therefore, successful resolution (some publishers are worse offenders than others!)
- there is some inconsistency in which fields are imported from CrossRef
- author metadata will need to be added from another source
- abstracts will not be imported and,
- in the case of WRRO, we would need to assign each record to its appropriate Academic Unit(s).

Of course, the relevant full text item would also need to be attached where available.

### **3.2 Importing metadata from commercial bibliographic databases**

Potentially, commercial databases are a source of high quality metadata. However, any bulk and/or systematic import and reuse may not be possible under the rights holders' terms and conditions. Opinion seems to be divided on whether this is a fruitful avenue to pursue or not. Some investigation and clarification – perhaps coordinated by JISC - might be helpful. In terms of generating metadata for “back catalogue” publications, as well as new research, it would be useful to consider where the most valuable sources of publication metadata are located. It would be helpful to clarify whether non-commercial metadata reuse within an open access repository is possible for metadata drawn from a commercial database (e.g. Web of Knowledge, Scopus). Are service providers open to negotiation? As many repositories allow import from EndNote and other bibliographic software, they may already include metadata originally output from a commercial product. It may be argued that such metadata reuse is simply a fact of scholarly practice.

Whether or not reuse of metadata is possible, bibliographic databases can be a useful source of new publication alerts. A tool to compare *In Press* repository items against such alerts might be helpful in updating records with full publication details.

### **3.3 arXiv**

arXiv provides open access to papers in a number of disciplines, particularly physics, mathematics and computer science. It is routinely used by many academics in these fields and contains a significant number of records for staff across the White Rose institutions. These records could provide a source of metadata for WRRO. However, the records are often missing key pieces of publications information. For example:

- papers are often uploaded to arXiv before they have been published
- there may be several versions of a research paper within arXiv – linked through the arXiv versioning system - but authors do not always supplement their arXiv records with the final, full publication metadata
- author names are not always consistent; this may be due, in part, to co-authors depositing a paper on behalf of their colleagues
- the metadata is held in a free text format, which causes some interesting challenges in parsing the information unambiguously into its constituent parts for import into EPrints. The technical aspects of this work, making use of a parsing tool developed by Tim Brody in Southampton, are described on the IncReASe web site <http://eprints.whiterose.ac.uk/increase/arxiv.html> .

In addition, affiliation data is often absent so identifying records by institution is problematic.

Improving the quality and completeness of the metadata records would require considerable work from repository staff. This would include searching for the published version of the

paper and adding any relevant, missing information. It is worth considering the extent to which significant metadata improvement by the local repository is valuable. Whatever the shortcomings of arXiv's metadata might be, they do not seem to perturb arXiv users unduly. Indeed, the arXiv users we have talked to were underwhelmed by our offer to enhance metadata for their arXiv papers within WRRO! Imperfect metadata may suffice for this community. Of course, this sits uncomfortably with WRRO's aim to produce high quality metadata and could limit the extent to which the metadata and publications within the repository could be used for other purposes.

It would be useful to identify DOIs, where they exist, for imported arXiv records – perhaps using the Simple Text Query facility in CrossRef or something similar. The DOI could enable enhancement of skeleton arXiv metadata with additional metadata via CrossRef.

Import from arXiv also highlighted the difficulty of handling different versions of the same work within WRRO (and other repositories). In particular, where a work is known to be published, we create a metadata record of full publication data, including the title of the published item. However, it is quite possible for an earlier version of the work to have a different title i.e. there can be a logical inconsistency between the descriptive metadata and the appended file. Implementation of the Scholarly Works Application Profile could resolve this inconsistency (discussed on the IncReASe web site <http://eprints.whiterose.ac.uk/increase/swap.html> ).

### **3.4 Local metadata sources**

#### **3.4.1 Centralised publications database (ULPD) / RAE database**

University of Leeds created a centralised publication database in the wake of the 2001 RAE - partly to expedite data submission in subsequent assessment exercises. The database holds several thousand records, including publication details for journal papers, conference papers, book chapters and monographs; (the database was used to collect the University's RA2 submission data so has item types matching the RA2 categories). There are tools within the publication database to improve the consistency of metadata (for example, an authority list of journal titles; author list pulled from the staff database). Nevertheless, quality across the whole dataset is patchy. The main issue identified is incomplete metadata. For example, it is common to find "Conference presentations" with a presentation title but no conference details; or conference papers which indicate they are "published" with no details of the conference proceedings or other publication venue.

A sample set of records were identified in the Leeds publication database and imported to WRRO via EndNote. A breakdown of the steps involved is included as *Appendix A*. The database import highlighted that we have no ready way to identify duplicates / potential duplicates within WRRO. This is certainly an enhancement which would be of great benefit – particularly where metadata may be imported from multiple sources.

The subset of data submitted to the RAE will be of excellent quality. We are working towards obtaining the RAE submission data for each of the White Rose partners with a view to using this as an advocacy tool when the RAE results are released in late 2008. Assuming metadata quality is as high as we hope, it is possible that these records will be imported directly to the live archive.

#### **3.4.2 Departmental publication databases**

The web survey undertaken early in the IncReASe project<sup>3</sup> identified a small number of departmental based publication databases. We had hoped to find some rich sources of metadata and full text; but this proved over-optimistic. Assessment of three department based databases revealed the usual problems of inconsistent and incomplete metadata. At the time of writing, we have not directly imported any of the identified databases (in part due to the proposed creation of new, central publication management systems at each partner site). We would need to weigh carefully the extent to which such a database import would require metadata enhancement and how much resource this would require. Potentially, this route may be a way to capture interest in the repository from the departments involved and, to this extent, may be a good investment of time. However, if, as looks likely, similar data is going to be available from a centrally controlled, hopefully more consistent source, this would seem to be a more fruitful line of investigation in the longer run.

### **3.4.3 Personal publication databases e.g. EndNote import / BibTeX import**

EPrints can import EndNote libraries – see, for example, the short Philosophy Department case study in *Appendix B*. EndNote and BibTeX import has been available for some time in WRRO but, other than the University of Leeds Philosophy department, no departments or individuals have used this facility to date. Of course, this may well be due to low levels of awareness. Promoting EPrints' import capabilities could increase usage of these options – though the quality of resulting metadata would probably be very variable. As with arXiv data, we would need to decide how much central repository resource could/should be devoted to improving bulk imported metadata.

Import plugins can be unforgiving: the EndNote and BibTeX plugins are all or nothing - if the process falls over part way through, no records are imported. Substantial take up of import plugins by individuals and departments might increase demands for technical support from repository staff. In addition, where records from a database are imported to a work area within the repository, they must still be deposited one by one to EPrints rather than en masse. Bulk movement (of records from the user workarea to the repository buffer or to the live archive) is possible, but is not a standard EPrints feature offered to users from within their workarea. Perhaps we could aim for a more managed ingest process by offering central mediation for bulk uploads of this type; we could define a minimum required metadata set for different types of publication. It remains to be seen, though, whether there is significant demand from researchers to upload personal publication lists in these formats.

### **3.4.4 Author Web page scraping**

Many authors maintain a publication list on their web page(s). From time to time authors have suggested that we simply “scrape” data from their pages and use this to populate the repository. This activity is potentially attractive – particularly where authors have:

- (i) a large number of publications
- (ii) links to appropriate copies of their work from their own web pages (e.g. local copies of journal papers (accepted manuscripts), copies of conference papers)

A perl program was written to go to an html page (the personal website of the researcher), extract the list of publications and use this to populate the EPrints repository. The perl removes all the html tags such that only the text remains. The text is then parsed to extract

---

<sup>3</sup> See the Database Prevalence Report at [http://EPrints.whiterose.ac.uk/increase/milestone10\\_database\\_prevalence\\_report\\_v1.0.pdf](http://EPrints.whiterose.ac.uk/increase/milestone10_database_prevalence_report_v1.0.pdf)

the authors' names, year of publication, the title of article and the journal reference. This was quite difficult to program because the structure of the text was not in a consistent format and therefore some manual intervention was required to overcome inconsistencies in the text. The program outputs a (text) file in an EndNote compatible format which is then imported into EPrints using the EPrints EndNote plug-in.

There were a number of metadata issues with this approach, the most difficult to overcome was the inconsistency in text format. Other issues include journal titles often prefixed with "to appear in" or "in". Therefore "to appear in Pattern Recognition Letters" would appear as a title and needed manual editing. The use of the prefix before the journal title was inconsistent and not limited to just the two phrases. In addition checking was required to determine if the "to appear in" publication status was still correct.

Sometimes the year would not appear in the publication citation and occasionally the publication citation would be incorrect. A couple of times it was difficult to decipher the type of publication. There were cases where a work was presented at conference and subsequently published as a journal article. These would appear as a single item in the publication list but were in fact two different types of publications.

The names of authors did not appear in a consistent format, e.g. some with initials and some with full names.

The perl script was written to process the publications page for a particular researcher; even based on one relatively accurate and consistent publication list, this was a challenging activity. To scale this approach across departmental or institutional web sites would be very difficult. To cater for variation in citation formats and various inconsistencies which can occur in publication lists, the perl script would need modification for each new list and the need for some manual intervention would always be likely. For very long publication lists, perl scraping may save some time – but the difficulties are significant and we do not propose to spend more time on this approach. The perl code will be made available from the IncReAsE web site should others wish to use it. Other projects such as AIR (Automated Archiving for an Institutional Repository)<sup>4</sup> are working to develop a more sophisticated algorithm to extract publications metadata from web sites; this may be a more realistic approach.

#### **4. Problem of bottlenecking**

Bulk import could rapidly increase the number of records in WRRO. If the metadata is of variable quality, as has been suggested above, the tidying up of metadata could require considerable amounts of staff time. Of course, this is in addition to time spent on liaison with depositors to obtain the correct version of a work plus any copyright advice or checking required. There is a danger of bottlenecking, with metadata and publications languishing offline whilst metadata quality is improved. This is a potential problem whether we deal with bulk uploads or simply experience an increase in self-archiving.

One approach would be to make imported records live automatically, despite inconsistent metadata. Repository staff could then work at improving the quality of those records as time permits. This would provide users with the quickest access to records within WRRO. An important risk-management decision would be the handling of any uploaded research texts. Should copyright clearance always be checked before works are made live? Where there is ambiguity (e.g. a journal paper where we do not know the publisher's self-archiving policy),

---

<sup>4</sup> AIR Project <http://clg.wlv.ac.uk/projects/AIR/>

should we make just the metadata live pending checking with the publisher? Or do we live with the risk and apply our Take Down Policy if required?

The other approach would be for repository staff to improve the quality of the metadata records before they are made live. Delays could occur, but metadata should be of a higher level of consistency and quality.

Essentially, we need to think about where we are located on the “quality-speed continuum” described by Nichols et al (2008) and decide what best serves the short and longer term needs of:

- the depositors and users of the deposited research,
- the White Rose institutions
- the libraries as the providers of the service
- research funders.

Repository records could be seen as having different short term and longer term requirements. Skeleton metadata – but sufficient to allow discovery – may suffice in the short term. However, it seems desirable to enhance such records with complete descriptive metadata to improve both discovery and potential reuse. This suggests the need for tools to identify incomplete records as well as tools to facilitate metadata enhancement.

## 5. Metadata analysis tools

The metadata analysis tool from the University of Waikato (NZ)<sup>5</sup> highlighted some weaknesses in metadata completeness within WRRO. For example, some dc elements are not populated e.g. dc.language and dc.rights. The analysis tool also offers handy visualisation of the results.

DRIVER<sup>6</sup> generated metadata compliance reports for various European repositories at the end of 2007. Again, this has been useful in identifying shortcomings in metadata and also provided information to help deal with any identified errors – though a reasonable level of technical knowledge is required to understand and act upon the information included in the report (examples included as Appendix C). The report was useful in highlighting a problem in the way we were exposing dc.creator metadata (subsequently resolved).

It would be helpful for repositories in the UK to have available a suite of easily deployable tools to support the identification of any metadata shortcomings across the repository dataset and to facilitate rectification of these shortcomings where possible. The tools could be set against particular standards (e.g. compliance with REF data format; compliance with DRIVER guidelines) or could be customised by repositories to analyse against their own criteria. It would be useful to be able to run analysis as a periodic batch process; regular feedback on material as it is added, e.g. from harvesting services, could also be very helpful. To rectify errors, it would be helpful if repository software allowed administrators to change multiple records at a time.

## 6. Duplication

As discussed elsewhere in the report, duplication can become an issue whether records are being imported in bulk or created manually – for example, a single paper may have multiple authors within the same institution (or, for White Rose, across three institutions) and more

---

<sup>5</sup> Waikato Metadata Analysis Tool <http://www.nzdl.org/greenstone3/mat>

<sup>6</sup> Digital Repository Infrastructure Vision for European Research <http://www.driver-support.eu/>

than one author – or their proxy - may archive the same paper. Ideally duplication would be detected before data is added to the live archive. Where duplicates are detected in the live archive, repositories need to be able to retire duplicate records gracefully and enable redirection to the appropriate copy of the work.

## 7. Conclusion

Potentially, bulk import – of metadata and possibly associated texts – is feasible. The development of SWORD facilitates the putting of material from one system into another and it can be seen that the opportunities for data sharing and the technology to support this are likely to improve over time. Most institutions will have a potential mountain of older research which could be added to an institutional repository. Although the acquisition of newer research may be prioritised, it is likely some institutions will wish to – or be encouraged by depositors to – flesh out their back catalogue – at the very least with items within scope for the Research Excellence Framework. The scale of this task suggests only automated population methods will be feasible. This may require some compromises and pragmatic decisions by repository managers.

Of course, as important as metadata is, the primary *raison d'être* for many open access repositories is the open dissemination of the research outputs themselves. Although bulk and proxy depositing is attractive – albeit often imperfect in terms of data quality – there are certain key pieces of information which can *only* come from the authors of the work themselves including:

- the author's own copy of the research (where a published copy may not be used)
- the refereed status of the supplied work
- the status of an unpublished work (i.e. whether it is In Press)

Whatever ingest models are used, it is important the creators of the research do not become too divorced from the deposition process.

We have identified a variety of potential metadata sources; unsurprisingly, they vary in quality. There is scope for the development of more effective ingest/post-ingest metadata analysis tools to help repository staff identify missing or suspect metadata. Even with analysis tools, however, the resource required to check a large collection of bibliographic data for accuracy (and check attached files for copyright clearance) should not be underestimated. Unfortunately we do not have sufficient data, as yet, to know whether significant effort directed towards bulk ingest of, primarily, metadata will make the repository a more attractive proposition for the deposition of new research.

## References

Allinson, J., Johnston, P. and Powell, A. (2007) A Dublin Core Application Profile for Scholarly Works. *Ariadne* 50, January 2007  
<http://www.ariadne.ac.uk/issue50/allinson-et-al/>

Nichols, D.M., Paynter, G.W., Chan, C-H., Bainbridge, D., McKay, D. Twidale, M.B. and Blandford, A. (2008) Metadata tools for institutional repositories. Working Paper: 10/2008  
[http://EPrints.rclis.org/archive/00014732/01/PDF\\_\(18\\_pages\).pdf](http://EPrints.rclis.org/archive/00014732/01/PDF_(18_pages).pdf)

## Appendix A

### Importing records from the University of Leeds Publication Database

Importing records from ULPD for an identified author in a particular department is quite straightforward. The steps involved require initially to setup an Access database locally and then retrieve all records in ULPD by using the “Get External Data” option in Access followed by “Link Tables” option choosing the File type option “ODBC databases” and then “Machine Data Source”. The next step requires the identification of the specific author in a particular department and their unique “person\_id” in ULPD. Using this unique “person\_id” all publication records associated with that person can be searched and exported in the tab delimited text format. This text file is then manipulated so that it is in a suitable format for import into EndNote. This is a manual process as all that is required is to search (for ;) and replace (with //) and also to copy and paste two lines at the top of the document from an EndNote text import template file. There are two ways to import metadata into EPrints although both require manual intervention. The first option being to import all items as a journal article type and then manually editing them into books or book sections etc. in EndNote. The other is to identify books and journal articles and save in different files and then import one file at a time into one library in EndNote. We followed the former as there were fewer steps and could potentially save a bit of time (although in reality there would not be much difference in time). The file is then exported in the EndNote format (similar to BibTex format). This EndNote exported file is then imported into EPrints via the EndNote plug-in.

In addition publications associated with a particular department can also be imported into EPrints using the same approach. However this could lead to duplications as it is likely that some records are already in EPrints and currently there is no quick method for identifying duplicated files.

Furthermore populating EPrints as a bulk process has the issue of incomplete metadata. For example there is no abstract or keywords, the academic unit field does not get filled in and the full text file would need attaching separately.

## **Appendix B**

### **Importing an EndNote library supplied by the UoL Philosophy Department**

The Philosophy department at the University of Leeds collated an EndNote library for import into WRRO. The database was created and compiled by two students who worked with the Philosophy academics to obtain publication information and the relevant files. The students received training in the use of RoMEO by repository staff and used standard templates to write to publishers to seek archiving permission where necessary. Because the EndNote library was created with input from repository staff, the collected metadata matched that required by WRRO; it was of a high standard of accuracy and consistency. The students could have simply archived directly into WRRO but as their collection of publications was a “summer project” they found it easier to collate the information over a number of weeks. On the whole, the import worked well. We experienced some technical problems – certain records caused the import process to fall over. We were not able to identify any differences between the importable and the “rogue” records. Also, we imported only metadata via EndNote and did not explore the feasibility of importing metadata plus attached full text. Texts were attached as an additional process.

## Appendix C

### Sample DRIVER report output

Rule name	Field - Rule	Error description	# errors	Rule Explanation
Matching datestamp granularities	<granularity> <datestamp>Pattern: {similar granularity}	When an error occurs your repository has different date patterns for the <datestamp> in Records and the <granularity> at the Identification page.	1	<p>The DRIVER guidelines recommend to use the same datestamp granularity in the OAI record as is provided in the Identify verb.</p> <p>Implications:The (DRIVER) harvesters rely on the correct information. Incremental harvesting is not possible when date granularity cannot be processed.</p> <p>Operation:?verb=Identify ?verb=ListRecords</p> <p>Check:Identify granularity pattern matches the Date pattern in record field.</p>
Incremental record delivery	<datestamp>Pattern: first and last datestamp within 'from' and 'until' boundaries	When an error occurs your repository does not have the capability of incremental harvesting.	1	<p>DRIVER guidelines recommend to have your repository setup with incremental harvesting capabilities.</p> <p>Implications:When the 'from' and 'until' commands are not working, DRIVER has to harvest your complete repository over and over again. The process is consuming a lot of you repository bandwidth, and the DRIVER harvester is consuming unnecessary resources.</p> <p>Operation:?verb=ListRecords ?verb=ListRecords&amp;from={date}</p> <p>Check:The validator checks if the expected values appear within the 'from' an 'until' dates.</p>