Project Acronym: IncReASe
Version: 1.0
Contact: r.e.proudfoot@leeds.ac.uk.
Date: 06/11/2008

JISC

Project Document Cover Sheet

Project Information	
Project Acronym	IncReASe
Project Title	Increasing repository content through automation and services
Start Date	01/07/07

		End Date	28/02/09
Lead Institution	University of Leeds		
Project Director	Bo Middleton		
Project Manager & contact details	Rachel Proudfoot Edward Boyle Library University of Leeds LS2 9JT 0113 343 7067 r.e.proudfoot@leeds.ac.uk		
Partner Institutions	University of Leeds, University of Sheffield, University of York		
Project Web URL	http://eprints.whiterose.ac.uk/increase/		
Programme Name (and number)	Repositories and preservation programme (04/06)		
Programme Manager	Andrew McGregor		

Document Name			
Document Title	Progress Report		
Reporting Period	25/04/08 – 31/10/08		
Author(s) & project role	Rachel Proudfoot, Project Manager		
Date	13/10/2008	**Filename**	Increase_progress_report3.pdf
URL	*if document is posted on project web site*		
Access	☐ Project and JISC internal		■ General dissemination

Document History		
Version	**Date**	**Comments**
0.1	13/10/08	Draft progress report for internal circulation
1.0	07/11/08	Public

Document title: JISC Project Report 3 – October 2008
Last updated: October 2008

Project Acronym: IncReASe
Version: 1.0
Contact: r.e.proudfoot@leeds.ac.uk.
Date: 10/11/2008

JISC

Progress Report Template -

Text in italics is explanatory and should be deleted in completed documents.

Project Name	*Increasing repository content through automation and services, University of Leeds*
Project Website	http://eprints.whiterose.ac.uk/increase/
Report compiled by	Rachel Proudfoot
Reporting period	*26/04/08 – 31/10/08*

Section One: Summary

We have continued to investigate sources for bulk import to the repository and have trialled import via a number of routes including arXiv, importing using PubMedID, import via DOI, import from EndNote and BibTeX, import based on publication details from an individual's web page and import from a centralised publication database. Our experiences are summarised on the IncReASe project web site and in the report *Research publication metadata. Where is it? How useful is it? Can we import it?* Repositories investigating import from such sources need to be realistic about the quality and completeness of metadata from each source, about what resources may be required to validate metadata and to what extent validation and quality checking processes add real value to deposited research. Of course, the answer to this may well depend on the uses envisaged for the repository and its contents. There are undoubtedly sources which could be used to provide "quick wins" for repository population (either with metadata or with metadata and text) but they may require some compromise of data quality. There may be a role for one-off exercises of this type to "beef-up" repository content whilst a more sustainable growth pattern is established. The development of tools to enhance post-deposit metadata quality would be helpful.

The results from our online questionnaire have been summarised in the *Questionnaire Report* and a copy of the questionnaire is freely available online in case it is of interest to other repository services. A number of interviews have been conducted – more will take place before the end of the project. These are available as anonymous case studies from the project web site. The interviews so far illustrate different responses to open access across disciplines. On the whole, the idea of open access is received favourably but some individuals have concerns over versions of works. Most respondents/interviewees are reluctant to engage with copyright issues – even with support. A common theme is the need to sell the benefits of WRRO more persuasively and the need for WRRO to fit well with other university systems. Respondents have been neutral or positive about the creation of a consortial repository (rather than a single institution repository). There is no uniform point of deposit which will suit all disciplines; however, a recurrent suggestion is that "acceptance for publication" could act as the key prompter for deposit. This fits well with the version of the work we want to capture. Scalable mechanisms for supplementing deposited works with subsequent, complete, post-publication metadata will be crucial for repository content to be used for anything other than basic discovery and access.

Project Acronym: IncReASe
Version: 1.0
Contact: r.e.proudfoot@leeds.ac.uk.
Date: 10/11/2008

Section Two: Activities and Progress

WP1 Project Management

- The project website is being redesigned so that it will provide a more readily navigable platform for project outputs.

WP2 & WP3 Activity survey / Bulk import

- Field mapping and sample import from the old University of Leeds Publication database completed. However, this system has now been replaced by Symplectic.

- Discussion with relevant Leeds staff and Symplectic staff has taken place. There is agreement in principle that developing an effective workflow linking the two systems is highly desirable. There is no exact timescale or agreed work plan currently in place – however, there is some discussion of a deposit mandate for Leeds. Though at an early stage, if this comes to fruition it will make system rationalisation a much higher priority. The urgency of this work may mean that linkage is achieved in (at least) two phases – with the initial phase being a relatively low tech, interim solution which would reduce the need to advocate two separate systems to Leeds depositors.

- We are investigating importing the RAE submission database from University of York into WRRO. Although this database is metadata only, it could be a good way to boost the profile and credibility of WRRO. Of course, we would be seeking full text deposit for as many items as possible. Data is stored in an Access database so we're optimistic the metadata will be readily importable and of very good quality.

- We have started to build a bank of interviews with researchers and these will be made available from the project web site (Milestone 8, Self-archiver case studies). We are anticipating that more interviews will be conducted before the end of the project and these will be added to the project web site. An analysis will be included in the final "workflows" report on Workpackage 2.

- Perl script for web page analysis (based on specific publication list from University of Sheffield) written, assessed and made available from the project web site (Milestone 5).

- We are collating our experience of importing data into EPrints using plugins. We will make brief notes available on arXiv, PMID, ISI, BiBTeX (data provided by University of York researcher) and DOI.

WP4 Metadata

- The emphasis of Milestone 9, the "Metadata enhancement report" has been changed to reflect the practical work we have done in testing bulk import from a number of sources. This report is available from the project web site at:

- Additional metadata fields outlined in the workpackage will be reviewed before the end of the project (e.g. author manuscript, full text).

WP5 Funder Repositories

- We have provided a list of our fields to ESRC's technical developers and sample XML outputs from EPrints. We will be concentrating on using SWORD to put a locally deposited work (in WRRO) into ESRC's Awards and Outputs repository. Currently the developers are assessing the field match between WRRO and the ESRC repository. We will work together to agree a standard format for funder/ grant metadata.

WP6 Integration

- We have been investigating authentication across the partners (including discussion with staff from the White Rose grid). No firm conclusions have been reached as yet on how authentication should proceed. We may need to be clearer about what we are trying to achieve and why; this will be discussed with the repository steering group.

Document title: JISC Progress Report Oct 08
Last updated: November 2008

- A distributed input model is being trialled across one Faculty in Leeds: local administrators have received input training and an introduction to RoMEO and copyright issues. It is hoped the administrators will provide an effective link to researchers and be able to provide first line archiving advice. Repository staff will continue to monitor metadata quality and any potential copyright issues. Where there is doubt whether a work can be made openly available, we have encouraged deposit in any case. This will mean that some full texts are "hidden" and will increase the proportion of metadata only records in WRRO - but, hopefully, by lowering deposit complexity, we will gain more content overall.

- We have completed the Service Interviews report (Milestone 14). The emphasis is somewhat different from that originally envisaged. The report is based on responses from our researcher questionnaire and one to one interviews. (Assessing how the repository will interact with departmental databases is increasingly looking redundant).

Section Three: Institutional & Project Partner Issues

- The three White Rose partners have recently entered into a Memorandum of Understanding, putting our shared repository on a more formal footing. This is a significant step in moving away from WRRO as a "pilot" service and stepping up our commitment to WRRO as an ongoing, shared service.

- All three partner institutions are reviewing what local systems are required to collate data for the REF. At the moment, WRRO is not in the frame to provide this function for any of the three partners. However, WRRO has been included in relevant discussions at the three partner sites and there is an understanding that WRRO would add value to any research management systems which collate bibliographic data.

Section Four: Outputs and Deliverables

- Redesigned web site:
  http://eprints.whiterose.ac.uk/increase/

- PERL script:
  http://eprints.whiterose.ac.uk/increase/scraper.html

- Detailed report on questionnaire findings and the raw data from the questionnaire (Milestone 10c):
  http://eprints.whiterose.ac.uk/increase/questionnaire_report_public.pdf

- Researcher interview summaries:
  http://eprints.whiterose.ac.uk/increase/interview.html

- Report (Milestone 9) - Research publication metadata. Where is it? How useful is it? Can we import it?
  http://eprints.whiterose.ac.uk/increase/milestone9_metadata_report.pdf

- Report (Milestone 14) - Services interviews report (imminent)
  http://eprints.whiterose.ac.uk/increase/milestone14_services_report.pdf

Section Five: Outcomes and Lessons Learned

- Not a new lesson, but recent progress at one of the partner institutions has been heavily influenced by open access buy-in from a small number of senior academics. Library advocacy has played an important role in making senior staff aware of open access possibilities but support for open access from peers – including those from other academic institutions - carries particular weight. The high profile deposit mandate from Glasgow University certainly seems to be having an impact.

- Any bulk import can cause bottlenecking within the repository buffer. We will need to think carefully about the balance between quantity and quality and how metadata enhancement, post initial deposit, may be handled.

- There is a danger that investigation of the quite complex issues involved in populating repositories with older material distracts from efforts to acquire new research on an ongoing, sustainable basis.

- We are making progress but are still not achieving the growth rate we would like. A new standing item on the repository Steering Group agenda will be "growth drivers" where we will analyse our own and other repositories' experiences to inform repository strategy and management. This may well impact on advocacy strategy and, hopefully, help us to continue to develop scalable ingest models.

- Sustained publicity is key. We will be renewing our publicity and advocacy efforts during the current academic year.

- Although outwith the scope of the project, we have been developing a White Rose ethesis repository over the past few months. It is likely that this work will increase awareness of open access more generally and may well have a positive knock-on effect for WRRO. Training events for PhD students also gives a good opportunity to include more general information on IPR management, including publisher contracts, and the potential benefits of OA for scholarship.

Section Six: Evaluation
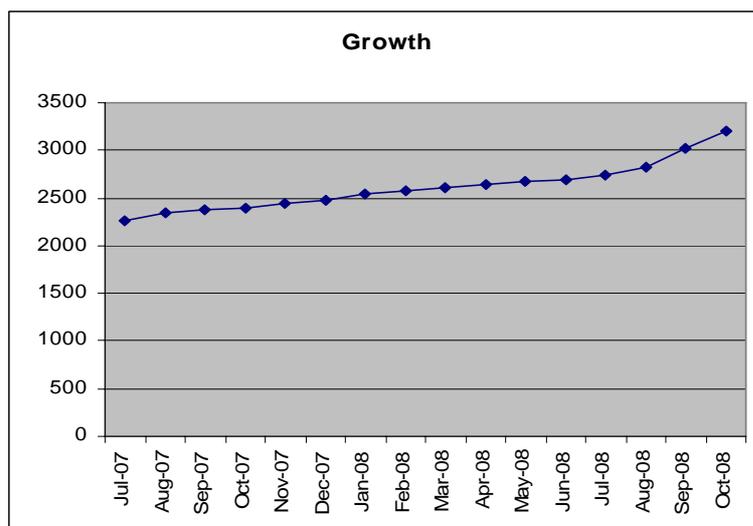
Monthly growth is monitored – as can be seen, there has been an incline recently but we are still a long way from achieving a significant increase in capturing a higher proportion of research outputs as they are produced. Advocacy activity needs to be widespread and regular. We are producing a more professional advocacy brochure which can be easily deployed across the consortium. Proactive work on content acquisition is probably still too diffuse – a bit here, a bit there. Regular follow up and reporting could help maintain momentum and encourage continued deposit. We are still struggling without a good statistical reporting tool. Ideally a regular, automatic/semi-automatic report should be generated and emailed – probably to the relevant subject librarian.

Newly added content continues to be made live in a timely way. Self-archived material is prioritised and is usually made live within one to two working days. Depositors are informed that their work is live and encouraged to submit further research outputs. Sometimes work is held somewhat longer in the buffer pending copyright permissions - but it could be better to make the metadata record live straightaway and "open" access to the file if permission if forthcoming. The

potential downside of this could occur where interested parties have an RSS feed for new work in a particular subject area – it could look as if a new entry was metadata only when, a few days later, it would be fully open access. A "contact the author" request button could be helpful here – though we have had technical problems implementing this (which we hope will be solved by upgrading to Eprints 3.1).

Section Seven: Dissemination

- Article pending in the White Rose Newsletter (online and paper copy).
- White Rose blog included in RSP blogs directory http://www.rsp.ac.uk/blogs/
- White Rose Exchange of Experience Event *http://wrro.blogspot.com/2008/08/repositories-exchange-of-experience.html*. A follow up meeting next year will include findings from the IncReASe project.
- Findings from IncReASe will be included in a presentation at the Bradford organised event "Beyond Advocacy" in November 08.
- Videoconference with representatives (Leeds, Sheffield, York) of the White Rose grid.

Section Eight: Risks, Issues and Challenges increase

As reported previously, we have been reliant on cooperation from departments to investigate how the repository fits with their local research management systems. The landscape continues to evolve, making it unclear where the best investment of time (by departmental and repository staff) may lie. There is shift towards centralised management of research publication information at Sheffield and York where, previously, much of this activity took place at the departmental level. There is a risk that work to link WRRO with locally managed systems may become redundant relatively quickly. Our effort may be better concentrated on, for example, acquiring the RAE submission data from the partners.

Section Nine: Collaboration and Support

What areas of work would you like to discuss with other projects?
Reuse of commercially sourced metadata.

Is there anything that you would like advice and support on? Do you have any specific training needs, requests or suggestions for supporting workshops for the programme?
Repository staff – and library staff – may limited experience of academic publishing processes. The more knowledge we have about this, the better i.e. what's typically involved in peer review, the various iterations of a research paper, how publishers deal with authors, what information authors are given before signing publication agreements and so forth. Could publishers become more involved in relevant workshops?

One area raised in the context of etheses, is the interaction between thesis deposit and subsequent publication and also the inclusion of previously published work within a thesis. Although this is addressed in the EThOS toolkit, more concrete information (from publishers) might be helpful.

There are also questions around the use of plagiarism detection software - by institutions but also by publishers. Again, more information might be useful to the repository community.

Section Ten: Financial Statement

Section Eleven: Next Steps

- If possible, acquisition of RA2 data (RAE) from York.

- Final shot at acquiring Comp Sci database in Sheffield.

- Upgrade to Eprints 3.1

- Conclude one to one interviews with researchers.

- Monitor impact of departmental administrator assisted self-archiving.

- Workflows report.

- Continue to work with ESRC re deposit via SWORD.

- Final project reports.

- Refreshed advocacy campaign for winter/spring.

Checklist:

Before you return this report:

☐ Ensure that your project webpage on the JISC site is up to date and contains the correct information. Attach details of any required amendments to this report. Project web pages can be found from: http://www.jisc.ac.uk/whatwedo/projects.aspx

☐ If there have been any changes to the original project plan and/or work packages, ensure that amended copies of the relevant sections of your project plan are attached to this report.