JISC

Project Document Cover Sheet

Before completing this template please note:

- *This template is for completion by JISC funded project managers*
- *Text in italics is explanatory and should be deleted in completed documents.*
- *Please check with your programme manager before completing this form whether they would like to use a specially adapted template specific to your project.*
- *Please see Project Management Guidelines for information about assigning version numbers.*

Project Information			
Project Acronym	IncReASe		
Project Title	Increasing repository content through automation and services		
Start Date	01/07/2007	**End Date**	28/02/2009
Lead Institution	University of Leeds		
Project Director	Brian Clifford then 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	*JISC Final Report*		
Reporting Period	*for progress reports only*		
Author(s) & project role			
Date		**Filename**	
URL	*if document is posted on project web site*		
Access	☐ Project and JISC internal	☐ General dissemination	

Document History		
Version	**Date**	**Comments**
0.1	26/02/2009	Steering group and project staff for comment
0.2	26/03/2009	Steering group
1.0	03/04/2009	JISC
1.1	01/05/2009	Minor amendment to appendix

JISC Final Report

Before completing this template please note:

- *The Project Management Guidelines explain the purpose of final reports.*
- *Fill in the information for the header, e.g. project acronym, version, and date.*
- *Prepare a cover sheet using the cover sheet template and attach to final report.*
- *This template is for completion by JISC funded project managers*
- *Text in italics is explanatory and should be deleted in completed documents.*

IncReASe (Increasing Repository Content through Automation and Services)

July 2007 – February 2009

Rachel Proudfoot, Archana Sharma-Oates, Beccy Shipman and Bo Middleton

Contact for project: Rachel Proudfoot

March 2009

Acknowledgements

The IncReASe project was funded by JISC under the Repository Start Up and Enhancement (SUE) strand of the Repositories and Preservation programme. We would like to acknowledge the cooperation and helpful contributions we have had from Cormac Connelly, Head of Information Environment and Dale Heenan, Web Project Manager, from the ESRC. We are also grateful to the administrators across the White Rose Consortium who have provided support for the project and the many researchers who completed our questionnaire and/or were interviewed by members of the team.

Project Team
Bo Middleton, Project Director
Rachel Proudfoot, Project Manager
John Salter, Computer Officer, University of Leeds Systems Team
Archana Sharma-Oates, Software Developer
Beccy Shipman, Project Officer

Project Steering Group
Bo Middleton, Head of e-Strategy and Development, University Library, University of Leeds
Elizabeth Harbord, Head of Content and Customer Services, University Library and Archives, University of York
Peter Stubley, Assistant Director for Academic Services, The University Library, The University of Sheffield
Rachel Proudfoot, White Rose Research Online Officer and IncReASe Project Manager

Table of Contents

Executive Summary

The IncReASe (Increasing Repository Content through Automation and Services) was an eighteen month project (subsequently extended to twenty months) to enhance White Rose Research Online (WRRO)[1]. WRRO is a shared repository of research outputs (primarily publications) from the Universities of Leeds, Sheffield and York; it runs on the EPrints open source repository platform. The repository was created in 2004 and had steady growth but, in common with many other similar repositories, had difficulty in achieving a "critical mass" of content and in becoming truly embedded within researchers' workflows.

The main aim of the IncReASe project was to assess ingestion routes into WRRO with a view to lowering barriers to deposit. We reviewed the feasibility of bulk import of pre-existing metadata and/or full-text research outputs, hoping this activity would have a positive knock-on effect on repository growth and embedding. Prior to the project, we had identified researchers' reluctance to duplicate effort in metadata creation as a significant barrier to WRRO uptake; we investigated how WRRO might share data with internal and external IT systems. This work included a review of how WRRO, as an institutional based repository, might interact with the subject repository of the Economic and Social Research Council (ESRC).

The project addressed four main areas:

(i) researcher behaviour: we investigated researcher awareness, motivation and workflow through a survey of archiving activity on the university web sites, a questionnaire and discussions with researchers

(ii) bulk import: we imported data from local systems, including York's submission data for the 2008 Research Assessment Exercise (RAE), and developed an import plug-in for use with the arXiv[2] repository

(iii) interoperability: we looked at how WRRO might interact with university and departmental publication databases and ESRC's repository.

(iv) metadata: we assessed metadata issues raised by importing publication data from a variety of sources

A number of outputs from the project have been made available from the IncReASe project web site http://eprints.whiterose.ac.uk/increase/.

The project highlighted the low levels of researcher awareness of WRRO - and of broader open access issues, including research funders' deposit requirements. We designed some new publicity materials to start to address this. Departmental publication databases provided a useful jumping off point for advocacy and liaison; this activity was helpful in promoting awareness of WRRO. Bulk import proved time consuming – both in terms of adjusting EPrints plug-ins to incorporate different datasets and in the staff time required to improve publication metadata.

A number of deposit scenarios were developed in the context of our work with ESRC; we concentrated on investigating how a local deposit of a research paper and attendant metadata in WRRO might be used to populate ESRC's repository. This work improved our understanding of researcher workflows and of the SWORD protocol as a potential (if partial) solution to the single deposit, multiple destination model we wish to develop; we think the prospect of institutional repository / ESRC data sharing is now a step closer.

IncReASe experienced some staff recruitment difficulties. It was also necessary to adapt the project to the changing IT landscape at the three partner institutions – in particular, the introduction of a centralised publication management system at the University of Leeds. Although these factors had some impact on deliverables, the aims and objectives of the project were largely achieved.

1. Background

[1] White Rose Research Online http://eprints.whiterose.ac.uk/

[2] arXiv - http://arxiv.org/ - an e-print service from Cornell University in the fields of physics, mathematics, non-linear science, computer science, quantitative biology and statistics.

A number of previous studies have illustrated some of the barriers and potential drivers for repository population (for example Henty 2007; David & Connolly 2007; Mackie 2004 ; Hey 2004). White Rose Research Online (WRRO) is a well established open access repository. Like many other repositories, we have had slow – but steady – growth and wish to increase the proportion of institutional research outputs we capture and disseminate. Unusually, the repository is shared equally between three partners – the Universities of Leeds, Sheffield and York, known collectively as the White Rose Consortium. WRRO uses the open source EPrints software from University of Southampton.

The IncReASe project aimed to build on the work of previous projects including the University of Leeds based EVIE (Embedding a VRE in an Institutional Environment) project. We were keen to understand more about our potential users, what might persuade them to utilise the repository on a regular basis and how capture of research outputs might be more effectively embedded into their research workflows. We were interested in exploring the feasibility of bulk population of the repository with metadata and publications from a variety of sources - from within and outside the partner universities - and whether this critical mass of legacy data would play any role in persuading researchers to deposit their new research outputs.

Previous JISC projects have highlighted that although self-archiving by authors may be an initial aim for an institutional repository service, it is not unusual for services to find it difficult to persuade academics to deposit – at least initially. It can be beneficial for a repository service to offer a semi- or fully mediated upload service (e.g. TARDis (Simpson, 2005); DAEDALUS (Nixon and Greig, 2005)). Historically, WRRO has offered authors the opportunity to self-archive directly; they have also had the option to send files for mediated archiving by repository staff. Self-archiving has been sporadic; mediated archiving has proved somewhat more popular but has still not been widely adopted. There is serious concern that a fully mediated service will not be scalable – particularly as we hope to accommodate the research outputs from three research-intensive universities. Through IncReASe, as part of our general move from "project" to scalable service, we were interested in considering who might be involved in repository "ingest" beyond the authors of the deposited works.

At the time of writing, all seven UK Research Councils have introduced a requirement (or "strong encouragement" in the case of the Science and Technologies Facilities Council) for their grantees to deposit research outputs into a suitable open access repository. The wording of funder policies varies; some policies can be satisfied with local deposit into an institutional repository whereas other funders require deposit elsewhere – specifically, the Economic and Social Research Council (ESRC) requires deposit to its own "awards and outputs" repository[3] and the Medical Research Council (MRC) requires deposit into UK PubMed Central[4]. During the IncReASe project, we wanted to look at how our local repository could help researchers fulfil their grant-related open access obligations. The interaction of local repositories with funder and subject repositories is a key issue for all institutional repositories, and we hoped IncReASe would make a useful contribution to exploring this rather fragmented repository landscape.

2. Aims and Objectives

At the outset, the aims and objectives were as follows:

2.1 Aim

The project aims to increase content in White Rose Research Online, to automate aspects of the repository ingest process and to start to embed the repository within research workflows by lowering barriers to deposit and investigating repository based services which may be useful to researchers. The project aims to produce reports and scenarios which will be helpful to other institutional repositories working towards embedding a repository within their own institutional workflows.

[3] ESRC's open access guidance is available at http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Support/access/
[4] MRC's position statement in support of open access is available at http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Openaccesspublishing/Positionstatement/index.htm

The overall aim of the project remained unchanged. However, developments in local IT systems, particularly the introduction (or scoping) of central research management systems, required us to consider different ways of meeting the overall aim.

2.2 Objectives

The IncReASe project will:

1. *survey sources of metadata and full text across the White Rose Consortium*
2. *test mechanisms for bulk ingest to the repository*
3. *enhance repository metadata using DROID and JHOVE*
4. *investigate whether it is possible to achieve economies of scale by organising the repository buffer by publisher*
5. *investigate enhancing metadata using CrossRef*
6. *identify strategies for scaling the repository from a pilot service based on central mediated deposit to a hybrid deposit model repository capable of ingesting and making available research outputs from the Consortium*
7. *review the relationship between institutional repositories and national and subject repositories, and explore the workflow implications for the population of Research Council repositories in particular*
8. *explore issues for academic and research staff around the research and publication lifecycle, and make recommendations for the optimal point at which research outputs should be deposited in both subject and institutional repositories*
9. *investigate what services could be offered back to depositing researchers in order to increase the utility of the repository and a feeling of greater ownership by the depositing community*
10. *produce reports, workflows and case studies of general interest to the repository community.*

White Rose Research Online will:

11. *double in size over the course of the project.*
12. *be capturing 20% of research outputs across the consortium by the end of the project.*
13. *maintain a high proportion of full-text outputs, with at least 80% full text content*
14. *offer services back to White Rose depositors: these could include tailored statistics, feeds for local databases and personal page generation*
15. *The three partner institutions will adopt and promote a formal open access policy.*

We revised some objectives during the course of the project. We introduced more user awareness / attitudes analysis through an online questionnaire. Some objectives became less pressing because of external developments; for example, objective 4 was not tackled because of overall improvements in buffer management introduced into the EPrints software. Because of staffing issues, there was some shift in emphasis from technical to non-technical aspects of the IncReASe project and this lead us to drop objective 3, which we felt was better tackled by other projects – or could be implemented as an enhancement by developers of repository platform software.

3. Methodology

3.1 Evolving strategy

The original strategy for the project assumed four main phases:

(i) Investigation of metadata sources across the Consortium and identification of pilot departments for workflow analysis and bulk data upload
(ii) Metadata enhancement as part of the repository ingest process
(iii) Building repository deposit into the standard research workflow
(iv) Offering services back to departments

It was planned that the four phases would overlap and feed into each other, but would be broadly sequential. However, we were not able to stick to this neat approach. Recruitment issues impacted on our capacity to liaise with pilot departments and import bulk data; this activity occurred later in the project than envisaged. There were also significant changes in the publication management

landscape at the partner institutions which altered our approach to the project – in particular, the implementation of phase (iv). There was an increased emphasis on assessing how the White Rose repository could complement other central services, such as publication management systems; the rationale for the services the repository might offer directly to individual departments became less clear.

The adapted approach to the project can be segmented into 4 main themes:

(i) researcher behaviour: investigation of researcher awareness, motivation and workflow
(ii) bulk import
(iii) interoperability, including fit with local systems and with the ESRC repository
(iv) metadata

3.2 Researcher behaviour

3.2.1 Web site survey

We undertook a survey of the three University web sites in part to identify patterns of self-archiving, but primarily to discover sources of full text and metadata (see 3.3 below). A more detailed methodology for this activity is available in our *Database Prevalence Report*.[5] The findings were used to create a map for each partner (using MindGenius[6] software), outlining where and how research outputs are currently presented online.

3.2.2 Online questionnaire

An online questionnaire was created, using Bristol Online Survey[7] software. The questionnaire is included as Appendix 1. A more detailed breakdown of the questionnaire methodology is available in the *Questionnaire Report.*[8] The questionnaire was distributed during Feb-March 2008 and the main topics covered by the questionnaire were:

- Research publications online (including current archiving behaviours)
- Knowledge of open access and repositories
- Drivers for deposit
- Funding bodies / awareness of funder open access policies
- Attitudes towards institutional and funder open access mandates

3.2.3 Interviews

Six detailed interviews (with seven individuals) were undertaken during the project, as well as a range of informal communications with academic staff and administrators. The interviewees were chosen to represent a spread of subject areas; some individuals had already deposited work in the White Rose repository, most had not. It is envisaged that more interviews / user needs analysis will occur post-project as an ongoing part of service development.

The in-depth interviews were semi-structured and covered the following topics:

- how researchers access research outputs and where they would wish their own outputs to appear
- awareness of open access journals and archives
- attitudes towards copyright agreements
- types of research outputs researchers wish to make available
- how research outputs could be captured
- role for repositories in funder mandate compliance
- other repositories/databases the researchers currently populate, or would wish to
- how WRRO might be of most assistance to the researchers
- attitudes towards WRRO as a consortial repository.

[5] IncReASe *Database Prevalence Report* http://eprints.whiterose.ac.uk/increase/web_survey.html
[6] MindGenius web site http://www.mindgenius.com/
[7] Bristol Online Surveys http://www.survey.bris.ac.uk/
[8] IncReASe *Questionnaire Report*
http://eprints.whiterose.ac.uk/increase/questionnaire_report_public.pdf

3.3 Bulk import

3.3.1 Web site survey

A survey of the institutional web sites was undertaken. It was hoped that we would find a number of potentially importable databases and some good sources of full text content. Although our emphasis is on acquiring new publications, it was hoped that bulk import could help create the "critical mass" of content we have been hoping to achieve since the inception of the White Rose repository. It was hoped that any bulk import processes would help the repository become a more credible and attractive system for depositors, possibly replacing legacy systems and obviating the need for departments to maintain separate lists of their own publications.

3.3.2 Departmental databases

During the course of the project, a number of potentially importable databases were identified and the relevant departments approached with a view to securing their cooperation. The delay in recruiting to our technical post meant that this work occurred later in the project than originally planned and thus shortened our window of opportunity to work with the identified departments. Gaining local cooperation and ensuring the availability of both departmental and repository staff at the same time, proved difficult. There was also the complicating factor of the move by each of the three partners to scope/institute a centralised research information system. As it was unclear how WRRO would fit into this evolving landscape, there were concerns that effort spent gathering, improving, importing data to WRRO and any ongoing links between local systems and WRRO would need to be duplicated in the not too distant future should a new, central system emerge. Nevertheless, investigating local databases led to productive liaison with departments and was valuable in informing discussion about how WRRO could be useful to depositing researchers. The departmental databases we have imported or are close to importing are:

(i) EndNote database, Philosophy Department, University of Leeds
(ii) RefBase database, Computing Science Department, University of Sheffield
(iii) Access database, Department of Information Studies, University of Sheffield

The technical details of the database imports are expanded on the IncReASe project web site at http://eprints.whiterose.ac.uk/increase/local_databases.html.

Unsurprisingly, the departmental databases identified during the course of the project varied in the quality and consistency of their metadata. Any bulk import of this kind requires careful weighing of the staff time required to improve the metadata against the potential benefits to depositors, the repository service and its users. (See Section 4.2).

3.3.3 University publication database

The University of Leeds Publication Database (ULPD) was a well established institutional system at the beginning of the IncReASe project, holding metadata for a good proportion of published outputs from the University of Leeds and providing the collection platform for Leeds' RAE2008 data. We aimed to create a robust link between ULPD and WRRO and consider the optimal workflow to populate both systems. We tested bulk import from ULPD. However, during the course of the project, ULPD was replaced with the Symplectic publication management system[9] so we changed our focus to investigate what would be involved in linking Symplectic and EPrints.

3.3.4 RAE data

We investigated the feasibility of importing metadata for publications included in the RAE2008 submission. The rationale for this was:
(i) to reuse readily available, high quality metadata
(ii) to use the high profile of the RAE to publicise the repository, with a view to establishing the repository's relevance for the forthcoming REF (Research Excellence Framework)
(iii) to encourage population of metadata records with full text
(iv) to contribute to overall repository growth.

[9] http://www.symplectic.co.uk/products/publications.html

For Leeds, importing RAE data was felt to be of less value, with effort best directed at the Symplectic linkage outlined in 3.3.3. For Sheffield, efforts were concentrated on departmental databases (see Sections 4.2.1) though we plan to revisit RAE data import should this prove effective in York. York's RAE data was imported in Jan 09.

Technical details of the import can be found on the project web site at http://eprints.whiterose.ac.uk/increase/ir_local.html.

3.3.5 Web page scraper perl script

Researchers often maintain their own publication lists on personal web sites and it is not usual for researchers to suggest that the repository "takes what it wants" from their web page. This is not ideal in that the researcher is not directly engaging with the repository; nonetheless, there is an understandable reluctance from researcher to re-create metadata which is readily available elsewhere. Therefore, we investigated the feasibility of "scraping" metadata from researcher pages using a perl script.

3.4 Interoperability

3.4.1 Authentication

WRRO exists as a standalone system at each of the three institutions. Depositors are required to create individual accounts on the system and this is a barrier to self-archiving. The isolation of WRRO is also a barrier to surfacing the repository through other applications – for example, offering deposit and display via an institutional portal. The project investigated authentication options for WRRO.

3.4.2 Publication management systems

Neither the University of Sheffield nor the University of York had a centralised publication database at the start of the project. However, both institutions have been actively investigating their options, driven in part by the anticipated demands of the Research Excellence Framework (REF)[10]. We have kept up to date with local developments and promoted WRRO as a potential solution to central publication management and/or a service which can significantly enhance central publication systems.

As outlined in Section 3.3.3 above, University of Leeds implemented Symplectic during the IncReASe project (Autumn 2008). The new publication system is known as TULIP (The University of Leeds Inventory of Publications). We are working with Symplectic to investigate a suitable workflow to populate both TULIP and WRRO and plan to synchronise TULIP and WRRO metadata via SWORD/ Atom Publishing Protocol. The Symplectic work has created both new possibilities and new challenges for WRRO – expanded in Sections 4.1 and 4.3.2.

3.4.3 arXiv

How institutional and central/subject repositories can work together effectively continues to be a significant issue. We were interested in whether there was any way to make the institutional repository more relevant to physicists, mathematicians, computer scientists and others who already deposit their work in the Cornell University based e-print service arXiv[11]. Our own conversations with arXiv users, plus investigations by others (e.g. Davis and Connelly (2007); Xia (2008)) suggest arXiv users will be reluctant to switch to local deposit, their needs already being well met by arXiv. Working on the assumption that arXiv users would continue to deposit directly to arXiv, and given that the deposited material was already openly accessible we investigated (i) what the rationale for developing an arXiv import facility would be and (ii) how import might be achieved using an EPrints plug-in.

[10] Research Excellence Framework information from HEFCE http://www.hefce.ac.uk/Research/ref/
[11] arXiv http://arxiv.org/

3.4.4 ESRC

As we began the IncReASe project, six out of the seven UK Research Councils had adopted policies encouraging or requiring the deposit of research outputs into a suitable open access repository. In most cases, local deposit into an institutional repository will fulfil this requirement. However, the Medical Research Council requires deposit into the UK PubMedCentral (UKPMC) repository[12] and the Economic and Social Research Council (ESRC) requires deposit into the ESRC Awards and Outputs Repository. Initial investigation into the UKPMC deposit revealed a two stage process where, post deposit, the principle investigator is asked to

(i) approve the UKPMC created PDF and, subsequently,
(ii) approve and, if necessary, correct the XML version of the work.

It was unclear how WRRO would fit into this deposit process. We felt the interaction between local repositories and UKPMC was likely to be complex and might well require liaison with UKPMC at a strategic level. We stepped back from investigating the UKPMC linkage and concentrated on the link with ESRC's repository.

We worked with ESRC to investigate different deposit scenarios and establish how ESRC might work effectively with WRRO and other institutional repositories, so that funded researchers would need to deposit their outputs in only one location which would populate both their local system and ESRC's repository. We considered both local deposit by the researcher (into WRRO) and remote deposit by the researcher into ESRC's repository; we considered OAI-PMH and SWORD as data transfer mechanisms.

3.4.5 RePEc

The York Management School was interested in adding its Working Papers series to the repository and was keen for them also appear in the economics repository RePEc[13]. We initially thought that RePEc would simply harvest papers if we exposed them over OAI-PMH but contact with RePEc revealed that this was not the case. A local RePEc compliant archive was created to house metadata in a form suitable for harvesting by RePEc.

3.5 Metadata

We investigated what would be involved in implementing the Scholarly Works Application Profile in our EPrints repository – in particular, we wanted to ensure we could capture funder and grant information to facilitate data exchange with the ESRC.

As bulk imports were an important part of the project, we were interested in how we could maintain metadata quality and consistency in the repository and whether there were authority sources we could use for different metadata fields.

4. Implementation and Results

This section is divided according to the four main themes of the project.

4.1 Researcher behaviour: investigation of researcher awareness, motivation and workflow

4.1.1 Survey, questionnaire, interview

We have been in contact with researchers throughout the project on a formal and informal basis. We introduced an additional element to the project – the questionnaire – as a useful accompaniment to the data from our web survey and as a prelude to more detailed interviews with researchers. Initial feedback was sought from the WRRO Steering Group and a small number of academics; this helped to inform the content and wording of the questionnaire (included as Appendix 1). We also contacted

[12] UK PubMed Central http://ukpmc.ac.uk/
[13] RePEc Research Papers in Economics http://www.repec.org/

the EMBED project[14] whom we knew had worked with the consultants Key Perspectives to undertake a Community Requirements Study, including a number of in-depth interviews with researchers. It was helpful to compare the EMBED findings with our own. This contact also led to the suggestion that we include a question on attitudes to open access mandates in our questionnaire; this was certainly one issue which prompted interest and enquiry from the repository community. Unlike the well known Swan and Brown (2005) study, we asked respondents separately about institutional and funder mandates. The timing of the questionnaire, during the Spring term, worked reasonably well. We had a good initial response - though, overall, lower than we hoped given that the email publicising the questionnaire was sent from the Pro Vice Chancellor for Research at two of the three partner institutions. Subsequent reminders about the questionnaire did not significantly boost the response rate. Unfortunately, timing in York was not ideal coming close on the heels of two other institution-wide questionnaires. Questionnaire fatigue is probably always going to be a factor in this type of research. Nonetheless, we felt 325 responses was reasonable and yielded a useful snapshot of researchers' views. (A summary of findings plus a more detailed Questionnaire Report are available from the project web site at http://eprints.whiterose.ac.uk/increase/quest_summary.html).

The face to face interviews took about an hour – though some rather longer than this. One of the interviewees had archived in WRRO, another had had work archived by repository staff. The other five interviewees had not archived in WRRO; two were regular arXiv users. The interviews elicited some useful comments and suggestions and were helpful in understanding individual work processes in more detail. The selection of whom to interview was tricky – we did not want a self-selected sample of those who were already familiar with open access and the repository but we found that asking individuals to comment on a system with which they had little or no familiarity was perhaps not the most effective approach to gathering user needs. In future, we anticipate continued dialogue with researchers to understand in more detail the different requirements of varied subject disciplines, and how this may impact on both our advocacy and the design of the repository. We should probably also work more closely with known self-archivers and actively seek their feedback and suggestions.

The results of the researcher focussed aspects of the project highlighted low levels of self-archiving (including on personal web pages) and poor awareness of both the availability of White Rose Research Online and of the existence of research funder mandates. The researcher constituency across the three White Rose partners is very diverse in terms of academic discipline and large in number; a broad-brush advocacy approach has not been sufficient to reach this wide audience effectively. The three University library services are currently planning a renewed advocacy campaign: each partner will agree a new advocacy plan (Summer 09), mapping out which departments we will work most closely with and who will be involved in the advocacy work. Professionally designed publicity leaflets and posters have been created with input from the WRRO Steering Group and subject librarians from the White Rose universities[15].

4.1.2 Self-archiving rates

It is often stated that, worldwide, the spontaneous level of self-archiving is around 10-15%[16] (i.e. about 15% of published articles are made openly available by their authors). We found similar levels of archiving: 16% of questionnaire respondents link to local, open copies of their work; 19% link to external copies – though often these are not openly accessible. Having said this, much of the self-archived content on web sites is working papers, reports and conference papers; the % of published journal papers spontaneously self-archived (on personal web sites or in any repository) by White Rose authors is likely to be lower than 15%. Of course, there is considerable variation between subject disciplines. This highlights the immediate potential value of open access repositories but also, perhaps, underlines the scale of the cultural change required – even after several years of institutional repository development - to engage researchers in active dissemination of their outputs.

4.1.3 Facilitating self-archiving

In our project bid we stated, rather ambitiously, that we would "..make recommendations for the optimal point at which research outputs should be deposited in both subject and institutional

[14] EMBED project http://cclibweb-1.dmz.cranfield.ac.uk/embed/index.php/Embed_Wiki
[15] The leaflet and poster can be viewed at http://eprints.whiterose.ac.uk/increase/publicity.html
[16] E.g. Harnard (2006), Björk, B-C., Roosr, A. & Lauri, M. (2008)

repositories." There was no overall consensus from researchers regarding the optimal point of deposit, though, for published works, the two most popular points for deposit were at the point of acceptance for publication and after publication. We know that it is the author's own version of a work, as accepted for publication, which is most often the appropriate version for archiving (according to publisher standard publication agreements). We are also aware that authors dislike the uncertainty surrounding which version of a work can be archived. We have concluded that asking authors to investigate appropriate versioning for each item is a barrier to self-archiving, even with the availability of SHERPA RoMEO[17]. Our observations suggest that conditions likely to improve self-deposit are:

(i) keeping things as simple as possible from the author's perspective

(ii) always asking for the author's final version of a work (we have incorporated the definition of "Accepted Version" suggested by The VERSIONS[18] project into our publicity)

(iii) facilitating capture of the work at the point of acceptance for publication. There are a number of potential approaches here – in the longer run, effective "desktop" capture may be developed, or liaison with publishers may lead to the provision of an "archive friendly" version of a work (either to the author or possibly directly to a repository service). In the absence of these developments, a regular, simple, advocacy message to associate acceptance for publication with a deposit action (be it personal self-archiving, archiving by proxy or emailing the copy to the repository service) is probably the most realistic approach

(iv) providing central support to monitor uploaded files and seek copyright clearance where required

(v) reminding authors to deposit: this could be a periodic reminder, or could be linked to a publication "event" such as a publication being indexed in a bibliographic database

(vi) highlighting the impact of deposit through the regular provision of usage data

Thinking specifically about published research outputs, there may be some benefit in seeing "self-archiving" as a process with two components with slightly different requirements (i) capture of a research output (ii) creation of metadata about a published work. The Symplectic system which we will be working with at the University of Leeds imports metadata for publications from Web of Knowledge (WoK) and PubMedCentral (PMC). Authors are emailed and prompted to accept the metadata record or, if it is a "false hit", reject it. Authors can change the metadata for the publication within the Symplectic system should they wish to do so. Once the link between Symplectic and EPrints is achieved, authors will also be prompted to upload a copy of the published work. From a self-archiving perspective, this activity is "a bit late in the day" as there will be a gap – potentially quite significant, depending upon subject area – between acceptance for publication and appearance as an indexed work in WoK/PMC. However, the provision of full publication metadata is a very attractive feature of the system. This suggests it will be helpful to investigate a self-archiving system which allows:

(i) capture of the research at the point of publication with basic metadata (the full publication details may not be known at this stage)

(ii) addition of full publication metadata, possibly from external sources, through supplementing the existing repository record for a given work or, if more appropriate, through the creation of a new version of the record

A repository is unlikely to work effectively in isolation but must become embedded within local IT infrastructure and within relevant local research environments. The association of grant data with research outputs will become increasingly important if repositories are to work to their full potential in meeting, monitoring and demonstrating research funder mandate compliance. In part, funder/grant data will highlight records where metadata/text should be pushed to central or subject repositories. Ideally, depositors should be able to indicate whether a work should be deposited elsewhere – perhaps from a drop-down list of well established repository services. But this is a complex area at relatively early stage of development. We have made some progress in exploring subject repository deposit through our liaison with ESRC (see Section 4.3.4 and Appendix 3).

[17] RoMEO Publisher copyright policies & self-archiving database http://www.sherpa.ac.uk/romeo/
[18] Versions Project http://www.lse.ac.uk/library/versions/

4.1.4 Archiving by proxy: working with departmentally based administrators

We trialled the use of nominated proxy archivers; administrators working on behalf of a specific school who create basic repository records and provide the main point of contact between their local authors and central repository staff. Authors email their research papers directly to their nominated contact. Departmental administrators vary considerably in the amount of ongoing training and support they require. Nonetheless, we have good examples of departmental staff who have built up confidence and knowledge in this area and have been key in sharing the repository management workload. Clearly, there are pros and cons to this approach. It remains to be seen whether an administrative infrastructure which moves authors further from direct self-archiving is counterproductive; it may not engender sufficient "cultural change" amongst authors to maximise sustainable self-archiving. Our experience to date, though, suggests authors will make the most of administrative support and that a helpful administrative framework results in higher levels of self-archiving overall. In particular, authors are responsive to well-known individuals in their departments: for example, local administrators have good success rates in persuading authors to re-send appropriate versions of their work where a non-archivable version (generally the published PDF) has been sent initially. Local administrators are well placed to "champion" and support the repository in ways that more "remote" central repository staff are not; this advantage needs to be balanced against the need to provide training and support for departmentally based administrators.

4.2 Bulk import

4.2.1 Publication databases and other local collection systems

Early in the project, we worked with the University of Leeds Department of Philosophy to import their EndNote database of research outputs. The database was created by two postgraduate students, employed over the summer, who, after receiving training from repository staff, checked copyright permission for papers, including writing to publishers for deposit permissions, and obtained research outputs from Philosophy staff. It would have been possible for the students to simply upload materials directly to WRRO on the authors' behalf - and this was the method suggested by repository staff. It is interesting to note that the department preferred on balance to create their own local database and upload material en masse at the end of the summer. Similar suggestions have been made from time to time by other departments even though creating an additional collection system involves more work at the local level. For example, we have been asked to provide an Excel template to allow data to be collected ready for periodic bulk import into the repository. Though this approach may seem counterintuitive, local academics and administrators have suggested that, for some departments, this may be a more sustainable method of data collection. Such solutions may be worth considering, perhaps as an interim measure, where sustained self-archiving activity is proving particularly elusive - though could prove counterproductive overall. Clearly, researchers and administrators favour well known software products over the unknown, alien "repository" platform, despite reassurances and demonstration of ease of use. In the longer run, capture methods which move repository deposit more firmly into the researcher's workflow – such as capturing research, as it is created, from the researcher's desktop – may well improve data capture from departments / individuals who have not been persuaded to deposit directly into the repository.

The Computer Studies department at the University of Sheffield maintain their own publication database using the open source web-based bibliographic management software RefBase[19]. One export format offered by the Computer Science database is that used by Thomson Scientific's ISI Web of Knowledge database. As there was an EPrints plug-in which was compatible with this format, this was used as the main mechanisms for export/import of the test data. Similarly, the Department of Information Studies maintains its own publication database: in this case, an Access database . Test data was output in Excel and imported into EPrints using the Multiline Excel plug-in modified to import author data. This work has come at the very end of the IncReASe project; the work to establish how the systems will inter-link is still underway. Essentially, this is a one-off import exercise to (i) reuse metadata (ii) create publicity within the departments with a view retrospectively populating the metadata records and, more importantly, start capturing more research outputs from the department

[19] RefBase software http://sourceforge.net/projects/refbase/

as they are created. The ongoing consideration of a new central research management system for Sheffield is pertinent to our next steps with these departments. Whatever systems arise, the key factor is the timely capture and exposure of full-text research outputs which might otherwise be lost; this is the message which underpins our liaison with these, and other, departments.

4.2.2 Individual web pages

Analysis of individual researcher publication pages revealed a good deal of inconsistency of formatting, including within individual publication lists. The idea of "scraping" publication metadata from researcher pages is attractive, but the reality is quite challenging. A perl script was produced which removes html tags, parses the references and imports to EPrints via the EndNote plug-in. Further details, including the code, are available online at http://eprints.whiterose.ac.uk/increase/scraper.html

The perl code written for one author could not be reused with another and would need tweaking every time. This was not considered an efficient method for automatically extracting publication data from the authors' personal web sites. A more sophisticated algorithm (possibly a machine learning algorithm trained on some real examples) would be needed to automatically scrape of websites and be able to determine what was likely to be a manuscript title, journal title, volume number, page number etc from the bibliographic information. The AIR, Automated Archiving for an Institutional Repository[20], project aims to do precisely this.

4.2.3 Use of EPrints plug-ins

We were able to utilise pre-existing EPrints import plug-ins and to customise import plug-ins to our own requirement; many EPrints plug-ins are available as standard with the software or available from http://files.eprints.org/view/type/plug-in.html . This is a valuable feature of EPrints though there is often scant documentation about the plug-ins and the plug-ins can be a bit unforgiving; one rogue record can cause the whole import to abort. Conversely, when importing using DOI, the import may be "successful" but, on inspection, some records may be blank bar the DOI itself. During the project we utilised import plug-ins for DOI, EndNote, BibTex, Multiline Excel and PubMed ID. Further details are available on the project web site[21] .

A note on the PubMed import plug-in

It's worth being aware of the distinction between ids used in PubMed and ids used in PubMed Central (PMC)/ UK PubMed Central (UKPMC); they are not the same. E.g. these two ids refer to the same paper:
PMID: 17210079
PMCID: PMC1774569
(It's possible to find out the relevant PMCID from PMID and vice versa from the PMID : PMCID Converter site at http://www.ncbi.nlm.nih.gov/sites/pmctopmid).

The current EPrints plug-in is set up to import metadata from PubMed using PMID. A PMC/UKPMC plug could be useful; particularly as PMC holds open access content whereas PubMed holds metadata.

4.3 Interoperability, including fit with local systems and with the ESRC repository

4.3.1 Authentication

Using local authentication removes the requirement for depositors to create a separate EPrints account and thus removes one deposit barrier. Essentially, we want users from any of the three partners to be able to login to WRRO using their institutional identity and password. Repositories commonly use LDAP or Shibboleth for this purpose. During the project, we discussed authentication issues with staff from the White Rose Grid e-Science Centre, which is also investigating implementing

[20] AIR Project http://clg.wlv.ac.uk/projects/AIR/
[21] http://eprints.whiterose.ac.uk/increase/plug-ins.html

authentication across the three White Rose partners[22] and with relevant technical staff from the University of Leeds. Unfortunately, we have not implemented an authentication solution; however, development resource has been earmarked to address authentication and author identification and we are hopefully of progress in this area during 2009.

4.3.2 Publication management systems

The evolution of publication or research management systems at each partner has been an important factor for the repository during the course of the project.

University of Sheffield

At Sheffield, we took part in an initiative, the *University Research Visibility Improvement Project (URVIP)*, which aimed to address research dissemination, including linkage with WRRO. However, the findings of the project were put "on ice" pending further information about the requirements of the Research Excellence Framework. In particular, it was felt the University would benefit from a research management system with broader functionality than that outlined by *URVIP.* Investigation into potentially suitable systems continues.

University of York

University of York is going through a similar process of assessing its research management needs and is in the process of scoping a Research Information System. The repository is represented on the working group designing the system; it is envisaged WRRO will be a key facet of whatever system is designed/ procured.

University of Leeds - Symplectic

University of Leeds has implemented the Symplectic system and we are in the midst of developing a link between Symplectic and WRRO. It is likely that, for Leeds, Symplectic will become the primary ingest route for both metadata and full text. This changes the way we work in some significant ways. Potentially, we have a source of high quality metadata for publications. But we also lose control of metadata quality as the Symplectic installation becomes our metadata authority source for any records that co-occur in Symplectic and WRRO. Although Symplectic harvests metadata from quality controlled sources, because of the wide subject spread at University of Leeds, a significant proportion (as yet unquantified) of additions to Symplectic will be via manual metadata creation. It remains to be seen whether repository staff – or library staff more generally – will have a role in ensuring metadata consistency, quality and completeness within the Symplectic system. Such proactive improvement is likely to be of long term benefit – not just for metadata quality within WRRO – but also because the data is likely to be used for Leeds' Research Excellence Framework submission. As Symplectic is set up to email individual authors directly, we potentially have a new mechanism for reaching out to authors and reminding them to deposit their research outputs. It is planned that Symplectic will be used to generate researcher publication pages; WRRO can supply the full texts to populate these pages. Clearly, there is potential for mutual benefit but the exact working relationship – including how "visible" WRRO will be from within the Symplectic system – is very much under development.

This period of change and uncertainty has made it particularly challenging to carve out a clear role for WRRO within the wider research management landscape at the three partners.

4.3.3 arXiv

On the assumption that arXiv users were unlikely to change their depositing behaviour (see Section 3.4.3 above) we developed a plug-in for arXiv. As works after often deposited prior to publication, we found that the metadata in arXiv was often incomplete. We also found that affiliation data can be absent - so there is no obvious way to identify all content from White Rose authors. We were aware of other repositories adding large volumes of arXiv data by hand, searching for works author by author. We estimate there are a minimum of 2,800 items from White Rose authors in arXiv (based on institution and postcode search) but quite possibly many more than this. Ideally, we wanted a process which was at least semi-automated.

[22] White Rose Grid e-Science Centre http://www.wrgrid.org.uk/

arXiv have made an API available to enable data extraction. Some metadata fields are unambiguous within arXiv and therefore straightforward to extract whereas others – particularly journal title, volume and issue number – are more problematic. We utilised the Biblio::Citation::Parser::Jiao perl module (written by Zhuoan Jiao and Ported to Biblio interface by: Mike Jewell at the University of Southampton) to parse citation data from a given reference. Pre-existing EPrints plug-ins (for PubMedID and PubMedXML import) were modified to create the arXiv plug-in for EPrints. It is necessary to know the either the arXiv ID of the publication or the name of the author to utilise the plug-in (multiple IDs can be entered into the plug-in).

4.3.4 ESRC

During discussions with ESRC over which deposit scenario's were feasible, it became apparent that the Economic and Social Science Research Council were actively exploring ways to make it easier for funded researchers to deposit their outputs and the impact of their research work. WRRO recommended extending the functionality of the ESRC website http://www.esrcsocietytoday.ac.uk/ to support the SWORD protocol.

Supported by WRRO, the Research Council has completed a proof of concept project for extending the ESRC Awards and Outputs repository. ESRC are currently in active talks with Microsoft about the use and implementation of their new open-source Research Output Repository Platform. It is their aim to deploy SWORD functionality in the near future and to allow any SWORD Compliant repository to deposit ESRC funded outputs into their systems. WRRO will continue to maintain this close relationship with the ESRC. The IncReASe project web site will be updated with any further developments.

The identified scenarios and further consideration of the SWORD protocol is included in Appendix 3.

4.3.5 RePEc

We found that RePEc require the creation of a simple local archive using a metadata format called ReDIF (Research Documents Information Format). The papers themselves are housed and fully described in WRRO as normal, and the WRRO URL for each paper is included in the ReDIF record. Full instructions for creating an archive are available from the RePEc site. RePEc allocates an archive code and provides templates for the various required files. A fuller description of our local RePEc archive is available on the IncReASe web site[23]. The process is not complicated but requires the maintenance of an extra metadata set. We created and maintain our small RePEc archive manually but there are converter scripts available[24] for EPrints (and DSpace and Digital Commons) to facilitate ReDIF creation. A current example of an EPrints installation automatically creating RePEc compliant output is the Munich Personal RePEc Archive.[25]

4.4 Metadata

4.4.1 SWAP

There was a good fit between known requirements for WRRO and potential solutions offered by SWAP implementation. In particular:

- Some of the metadata we enter about an item describes the item itself; other metadata describes a different, published version of the item. It would be useful to be able to assign appropriate metadata to describe a work and show its relationship with other version(s) of the work. The SWAP approach of defining "entities" and "relationships" fits well with our requirement.
- It is possible to append more than one file to a single record within Eprints: for example, the metadata describes a published work, a version of that work is attached but also one or more supplementary files may be attached. It would be useful to have a way to describe each of these attached files and the relationship between them. Again the SWAP model provides a solution.
- We want to associate items in the repository with funder and grant data. These are additional fields suggested by SWAP.

[23] IncReASe RePEC case study http://eprints.whiterose.ac.uk/increase/eprints_repec.html
[24] RePEc Scripts http://ideas.repec.org/s/rpc/script.html
[25] MPRA http://mpra.repec.org/

- Under pressure to grow our repository but struggling to populate it with full text, we are likely to want to incorporate metadata only records but need to be able to differentiate records where we offer full-text from those where we don't; plus it will be useful to differentiate those records where we offer full text and open access and those where we offer restricted full text. For example, embargoed works. Again, this requirement would be addressed through SWAP implementation.

In terms of implementation, we were unsure of the fit between the essentially "flat" structure of EPrints and the hierarchical structure of the SWAP model. From Version 3.1, EPrints software has implemented some additional metadata fields, as specified in SWAP, but the software changes do not address the scenario described in the first bullet point above. We were concerned that meaningful implementation of SWAP would require significant customisation of our EPrints installation and that this could have knock-on effects when we came to upgrade to subsequent EPrints releases. We concluded that the most productive approach to SWAP implementation is for the developers of the repository platforms to incorporate it into their core code.

We have implemented a Funder Information field which holds the name of the research funder and the relevant grant number. This is a repeatable field and will hold multiple funders. Ideally, we would like to draw research funder data from local systems at the three partners. This capability is some way off. In the short term, we are introducing a controlled list of funders based on the Research Information Network (RIN) list of funders.[26] We are investigating the automatic creation of an additional acknowledgement field according to RIN's recommendations[27] i.e.

> *This work was supported by the Wellcome Trust [grant numbers xxxx, yyyy]; the Natural Environment Research Council [grant number zzzz]; and the Economic and Social Research Council [grant number aaaa].*

It may be that this is best handled as an enhancement to core EPrints code if most institutions / depositors wish to capture this information.

4.4.2 Metadata quality

We have tried to maintain good metadata quality and consistency within WRRO but we have variation in name formats for authors and, to a lesser degree, journal titles and publishers. Sources of metadata are often imperfect; repositories need to make a realistic assessment of the resource needed to improve metadata and decide whether the added value justifies the cost of the resource needed to achieve it. Repositories need to think about how the metadata will be re-used and strike an appropriate balance between speed of dissemination and quality of metadata. It may be that tools to work in conjunction with bulk metadata ingest (e.g. to identify empty fields or potentially anomalous metadata) could help improve metadata quality. Quality tools may work well when there is a defined standard to work to: for example, requirements for harvesting by DRIVER[28] or requirements for REF data submission (once known).

5. Outputs

We hope the IncReASe web site[29] provides a helpful overview of the areas we addressed during the project. We have made available the questionnaire distributed to potential depositors (also included here as Appendix 1), an analysis of our questionnaire finding, summaries of our interviews with researchers, a Services Report[30] outlining areas for possible service development and some general observations on researcher behaviour.

[26] *Major funders of research in the UK* http://www.rin.ac.uk/files/List-of-major-UK-research-funders.pdf
[27] RIN (2008) *Acknowledgement of Funders in Scholarly Journal Articles: Guidance for UK Research Funders, Authors and Publishers*
http://www.rin.ac.uk/files/Acknowledgement%20of%20funders%20full%20guidance.pdf
[28] DRIVER Digital repository infrastructure vision for European research http://www.driver-support.eu/
[29] IncReASe project web site http://eprints.whiterose.ac.uk/increase/
[30] http://eprints.whiterose.ac.uk/increase/milestone14_services_report.pdf

IncReASe contributed to an overall increase in content for WRRO and helped us to develop mechanisms for bulk import. The bulk import of RAE2 data for York should provide a useful platform for further advocacy at that institution. A summary of our experiences importing databases can be read on the web site and we have made available the perl script[31] we used to parse references from a researcher's publication list. Some consideration of metadata issues and a short checklist of issues to consider when bulk importing are also available[32].

In the context of reviewing how our repository might work in conjunction with subject repositories, we created four arXiv plug-ins for EPrints – described in Appendix 2 and we hope these will be helpful to other repository services looking at importing content from this source. The plug-ins have been made available from the IncReASe web site[33] and from the EPrints Files repository[34]. A description of how we created a RePEc archive for The York Management School is available[35]. We also worked with ESRC to provide test data to explore how locally deposited research outputs might be fed to ESRC's Awards and Outputs repository: further information is available from the web[36]. The potential deposit scenarios we created during this work with ESRC are included at the end of this Report as Appendix 3.

During IncReASe we designed a new poster and leaflet; these are available from the project web site[37]. These will be utilised in a publicity campaign following completion of the project – vital in the light of the low levels of open access awareness highlighted by the IncReASe questionnaire.

6. Outcomes

The overall aims of the project, as set out in Section 2.1, were partially met. We made progress in all the area we planned to. The overall increase in content could have been higher; having said this, we are now experiencing a sharp upturn in repository growth. Steps have been taken towards better embedding of the repository within the three partner institutions. As outlined elsewhere in the report, what we have been able to do has been influenced by factors external to the project, such as changes in local IT systems and research management processes. Work undertaken during the project will provide a helpful springboard for further service development.

Section 2.2 lists the original project objectives and outlines why some were dropped or amended. The remaining objectives are considered below.

1. Survey sources of metadata and full text across the White Rose Consortium

We have produced a useful summary map for the three partners. This will be used to help identify further target departments for inclusion in the new advocacy plan for each partner, to be agreed by summer 09. In particular, it is useful to know the baseline of archiving activity in order to plan the type of advocacy that is required and the level of ingest support that may be needed. We know that, where databases exist, import may look attractive but, in practice, tends to be resource hungry.

2. Test mechanisms for bulk ingest to the repository

A number of EPrints plug-ins have been tested and we have imported bulk data from departmental databases, from a centralised publication database and have imported the majority of York's RAE publication data. Close liaison with the departments with publication databases has proved a useful means of increasing overall awareness of WRRO and of informing our own understanding of publication data management requirements at the departmental level. The level of staff support needed to handle bulk import should not be underestimated – particularly if the repository wishes to maintain a good level of metadata

[31] http://eprints.whiterose.ac.uk/increase/scraper.html
[32] http://eprints.whiterose.ac.uk/increase/import_guidelines.html
[33] http://eprints.whiterose.ac.uk/increase/arxiv.html
[34] http://files.eprints.org/
[35] http://eprints.whiterose.ac.uk/increase/eprints_repec.html
[36] http://eprints.whiterose.ac.uk/increase/esrc.html
[37] http://eprints.whiterose.ac.uk/increase/publicity.html

quality and consistency. Our experience with the Philosophy Department at the University of Leeds outlines that bulk import alone will not lead to continued deposit momentum; even though many researchers contributed their works to project staff, they did not continue to deposit once these staff were no longer actively seeking content from them. For maximum impact, any bulk import of this kind needs to be accompanied by:

(i) ongoing advocacy to the department in question
(ii) regular reports of downloads to emphasise the positive benefits of depositing
(iii) clear instructions for researchers on populating the repository "post project"; ideally endorsed by the departmental research committee or similar
(iv) engagement of a local champion to keep deposit on the agenda – for example, this could be a researcher, an administrator or the appropriate subject librarian.

Although sources of bulk metadata for legacy publications may look attractive, careful consideration should be given to whether the bulk of data imported will have a positive impact on researchers' engagement with self-archiving. It may do, but only if accompanied by an appropriate supporting framework of advocacy and liaison. There is always the danger of bottlenecking, where records appear slowly as they are improved by repository staff, and that the efforts to deal with legacy data take resource away from promoting the repository. In case it's of use to others, we have collated a short set of guidelines – or, rather, issues to consider – when approaching a bulk import of data. This is available from the project web site at http://eprints.whiterose.ac.uk/increase/import_guidelines.html

3. Investigate whether it is possible to achieve economies of scale by organising the repository buffer by publisher

We did not organise the repository by publisher, however, we have adopted different methods of organising bulk data. By default, the EPrints buffer is a flat list of deposited items and can be difficult to navigate where there are larger numbers. EPrints 3 allows greater customisation of what fields are listed in the buffer and allows reordering of records within the buffer e.g. by depositor. Each WRRO partner contributes administrative time to the repository and so there can be a number of different staff, at different institutions, accessing the buffer and processing records. We introduced a new notes field, viewable in the buffer, which allows the repository administrator to indicate whether the item is being dealt with e.g. if we are awaiting a response from a publisher or awaiting a file from the author.

It is WRRO policy to prioritise any items self-archived by authors; we do not want these items to be "lost" in the middle of bulk imported data. To address this, we have created "user accounts" with full administrative privileges for the various bulk imports, processing these records outside the main repository buffer.

4. Investigate enhancing metadata using CrossRef

Import via the EPrints plug-in, using single or multiple DOIs, works well. The main drawback with CrossRef is the lack of author data. We have used CrossRef as a base source of metadata but not to enhance metadata in records already created within the repository.

5. Identify strategies for scaling the repository from a pilot service based on central mediated deposit to a hybrid deposit model repository capable of ingesting and making available research outputs from the Consortium

This is a large and ongoing challenge. Our investigation of bulk import has some relevance to this objective. We need to create ingest models which are scalable across three large institutions. Clearly, centrally mediated deposit, without significant additional resource, is not an appropriate ingest model. Effective interaction with core university IT services, for example, any other systems collection research output metadata, is crucial if the repository is to be a credible, sustainable service. We have made progress in this area but there is still a long way to go. We have also looked flexibly at different ingest models. Author self-deposit is still our ideal scenario – so long as there is sufficient central resource to apply any locally agreed quality control measures and to advise on/ assist with copyright checking.

6. Review the relationship between institutional repositories and national and subject repositories, and explore the workflow implications for the population of Research Council repositories in particular

This is a major piece of work and an objective which was, realistically, too ambitious for a small scale project like IncReASe. Nonetheless, we have improved our understanding of the issues surrounding PMC interaction and have made some progress in enhancing the relationship between WRRO and arXiv, RePEc and ESRC's repository (explored in various sections of this report).

7. Explore issues for academic and research staff around the research and publication lifecycle, and make recommendations for the optimal point at which research outputs should be deposited in both subject and institutional repositories

See Section 4.1.

8. Investigate what services could be offered back to depositing researchers in order to increase the utility of the repository and a feeling of greater ownership by the depositing community

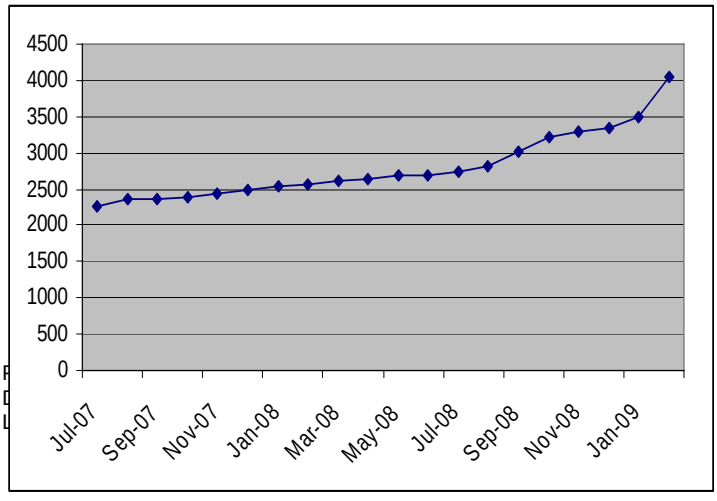
We have identified a number of areas for potential service development, the most popular being provision of tailored statistics. A few interesting suggestions arose from our researcher interviews – though not necessarily feasible ones (e.g. providing a translation service). Defining WRRO's role within the research management landscape of the partners is proving challenging. Our consortial arrangement introduces an additional layer of complexity when positioning WRRO as a local service provider. It's possible that the collaborative nature of the repository, existing as it does within a well known, well established consortium structure could act as a selling point and perhaps offer new avenues for service development. The identification of local collaborators from within the consortium was one of the original reasons for our shared repository and this is an avenue for further investigation in the future. Our work on import options and exploration of how WRRO fits with central/subject repositories may persuade more researchers to engage with the repository. The service which has most direct impact on researchers and on the likelihood they will self-archive (themselves or by proxy) is copyright advice and checking. The most successful selling point of the repository for University managers is the possibility of increased citation rates.

One drawback of the "institutional" repository is that it is can be perceived as collecting data for reporting and monitoring purposes and as being a system which is predominantly administrative in nature rather than a system which offers benefits to the depositor and their academic discipline. Depositors rarely seem excited by the idea of the repository or the prospect of the availability of their work within it. It may be that we need to find additional ways to add value to the deposited work and/or build links with other services which enrich repository content – for example, linking to relevant primary data – before we start to see "greater ownership" of the repository within the depositing community.

9. Produce reports, workflows and case studies of general interest to the repository community

A range of materials have been created and made available from the IncReASe web site. We publicised the results of our questionnaire but could perhaps have been more proactive in promoting / seeking comment on other areas of project work.

10. Double in size over the course of the project



At the original start date for the project (April 07), the repository held somewhere over 1,600 items. Taking this as the baseline, we have exceeded our target. However, as we delayed the official project start date to allow for staff recruitment, if we take our figure from July 07, we have fallen slightly short but will meet the target approximately 1 month post-project.

As can be seen from the graph, the growth rate has been much stronger in the latter half of the project.

11. Be capturing 20% of research outputs across the consortium by the end of the project

This target has not been met and we have some way to go before it is achieved. Progress has been made. We have a deposit mandate from one large Faculty at the University of Leeds and we anticipate that developments in central publication management systems will have a positive impact on the overall capture of research outputs. It remains to be seen which ingestion model is most effective for each of the partners (self archiving, (or archiving by a nominated proxy) via a central publication system; direct self-archiving into WRRO; archiving via departmental based editors; central archiving by repository staff). Close monitoring of the different ingestion methods currently in operation will be required to improve the effectiveness of research output capture. Across the partnership, we estimate nine-ten thousand items falling within repository scope are produced per annum. Eventually, we need to be ingesting / be capable of ingesting over 200 new items each week; this excludes the "mountain" of legacy metadata and publications which could potentially be added to WRRO.

12. Maintain a high proportion of full-text outputs, with at least 80% full text content

This target has been met. For the majority of its life, WRRO has had a high proportion of full text records (90- 95%). At the close of the project, approximately 82% of items have a local full text openly accessible copy of the research outputs; an additional 5% or so link to a full text open access works outside the repository. The proportion of metadata only records is increasing because of the addition of the University of York's RAE data and other bulk imports. It is anticipated that the proportion of full text items will fall to 60% for a short time but that the proportion of full text will then start to recover.

13. Offer services back to White Rose depositors: these could include tailored statistics, feeds for local databases and personal page generation

Service development is still underway as the project comes to an end. We have made limited progress on introducing statistics to the repository – beyond basic Google Analytics reporting – but development resource is earmarked for this work during 2009. The role of the repository in other types of service delivery remains unclear. Other university services may become the main generators of personal web pages. Whatever the architecture of the local systems, WRRO is likely to have a continuing role at all three partners in adding full text content to any automatically generated publication pages.

14. The three partner institutions will adopt and promote a formal open access policy

University of Leeds has taken its first steps towards a "patchwork mandate" (see Sale 2007); the Engineering Faculty has introduced a requirement for staff to deposit published journal and conference papers produced since January 2008. A university wide policy has not been adopted but there has been active discussion with senior management about this possibility. University of Sheffield has introduced a requirement for PhD candidates registering from 2009 onwards to deposit electronic copies of the theses with a view to making these openly available; similar policies are under discussion at University of York and University of Leeds. University of Sheffield has a long standing policy on Open Access (since 2005)[38] - though the policy currently encourages rather than requires deposit.

The project was probably over ambitious in the number of different areas it attempted to address and elements of the project were heavily contingent upon cooperation and availability of staff in academic departments. There is some feeling that the project was "development heavy"; perhaps more emphasis should have been placed on ongoing advocacy. It has sometimes been difficult to balance project-driven activity with the day to day needs of the repository. Having said this, the ongoing WRRO service has benefited from the IncReASe project in several ways. The additional staffing supported by IncReASe has enabled investigating and customisation of our repository software, allowed us to liaise more effectively with internal and external technical staff and provided an increased resource for

[38] *Guidance for Departments on dissemination of research papers and research results to avoid breach of copyright* http://www.shef.ac.uk/content/1/c6/03/42/48/Copyright.pdf

the investigation of depositor requirements and attitudes. The project really started to come together effectively in its last few months; spin out work and development will continue well after the official "project" end date.

7. Conclusions

Awareness of open access and of WRRO/ repositories is low across the partners – lower than we hoped. This extends to awareness of research funder open access requirements. There is much advocacy and awareness work still to do. Post-IncReASe we have more import and export options to offer depositors and we have strengthened our links with both central research support and central IT systems at the partner institutions; this is helping us to improve the profile of the service and link it with other, key central services.

There are likely to be personal and departmental sources of metadata suitable for bulk import at most /all HEIs. The metadata within such systems may well be inconsistent and incomplete. We found import to be more time-consuming than we hoped. A high degree of manual intervention was required: mainly to supplement incomplete metadata or add full publication details to imported "in press" items. Unless effective ways can be found to automatically check and improve bulk metadata this type of import may be a false economy and may not be the best way to grow the repository sustainably nor to embed into researchers' workflow. An alternative approach would be to identify sources of pre-quality checked metadata – possibly from commercial sources – to create a back-catalogue of publication metadata.

Our discussion with researchers suggests that a comprehensive service – essentially, a publication database - is probably an easier sell than a pure "open access" repository (echoing the conclusions previously drawn by, for example, the TARIS project); its raison d'être is clearer and the possibility for providing services back to researchers in the form of full listings of research and detailed information on traffic to individual works, is increased. Currently, this is not the direction being taken by WRRO; rather, because other central services are likely to fulfil the publication database function, the emphasis remains on external dissemination of open access outputs. However, there is a danger that an overly exclusive collection policy may be stifle growth and researcher buy-in. It is probably best not to be too dogmatic in approaching repository design and content; a hybrid mix of metadata and full text may be inevitable for repositories as they continue to evolve.

There is a good chance that dual deposit into a local repository and ESRC's repository will become unnecessary where both repositories implement data-sharing – possibly utilising the SWORD protocol. ESRC was cooperative and willing to investigate options for the mutual benefit of depositors, IRs and ESRC's growing social science repository.

There is probably no simple "optimum" deposit point for research outputs; however, in the short term, capturing papers at the point of acceptance for publication is probably the most realistic option. The emergence of desktop capture/deposit tools may facilitate earlier capture and assist with version control. Capturing the most appropriate version of a work continues to be an issue; all efforts should be made to inform researchers about the "accepted version" and its importance in the open access landscape. It is likely to be helpful to instil this awareness in early career researchers and PhD students by including open access / scholarly communication elements in training.

Uncertainty around copyright is likely to continue to be an issue for academics; central advice and support in this area is valued by academics and likely to be necessary even in a well-embedded repository with widespread self-archiving.

Capturing grant and project data relevant to research outputs is likely to increase in importance; this data can help maximise the value of repository content for both research and administrative purposes.

WRRO's mission is to handle research outputs; during the course of the project, emerging digital systems at the partner institutions – for example, the creation of the Leeds University Digital ObjectS (LUDOS) service[39] and the York Digital Library[40] – mean that there is an increasing capacity at the

[39] Leeds University Digital Objects (LUDOS) http://ludos.leeds.ac.uk/ludos/

partners to handle diverse digital materials. It is becoming less meaningful to see WRRO as separate from these services and to consider the handling research outputs in isolation from other categories of digital material. Rather, the three partners are moving towards a more integrated, holistic approach to managing digital content.

8. Implications

Our experience suggests researchers' awareness of and understanding of open access issues (open access publication models, self-archiving in repositories, research funder requirements) remains low. Targeting relevant information to PhD candidates, as the next generation of researchers, may well contribute to longer term cultural change. For current researchers, some emphasis on REF benefits may increase their interest in the repository but there is, perhaps, a significant amount of work to do in creating repository services which are meaningful within different subject disciplines.

If repositories look at wholesale import of metadata from local or external sources, there may be an increasing need for easily deployable tools to identify metadata shortcomings across a dataset and to facilitate rectification of these shortcomings. For example, it could 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. The NZ KRIS service provides error alerts via RSS feed (Nichols et al 2008).

One potentially interesting area for investigation is whether effort is best directed towards making "the repository" as a specific service, high profile and more engaging, with added value features, attractive presentation and so on; or better directed towards making "the repository" so well integrated with other services, ideally ones which are integral the researcher's everyday work environment, that it is effectively "invisible".

Even where self-archiving is achieved, there is likely to be a central role for repository staff in checking that the correct version of a work has been uploaded and in liaising with both authors and possibly copyright holders. Where this is the case, the administrative load can be quite high. Good systems are needed to help repository staff manage this administration. A repository platform designed around self-archiving and assuming minimal central intervention may not lend itself to providing this administrative scaffolding. Perhaps there is a need for repositories offer better generation and tracking of administrative actions. Similarly, for administrative purposes – but also to reassure depositors and copyright holders – better mechanisms for ensuring a clear audit trail of permissions relevant to a deposited work could be helpful. For example, capturing the relevant RoMEO entry for the date the work was deposited.

The development of desktop deposit mechanisms – particularly if the captured work can be pushed to more than one location – could be helpful.

Further exploration of the scholarly workflow with academic publishers, including the clearer identification of the "accepted version" of a work, would be beneficial for authors.

The central, subject and institutional repository landscape remains fragmented and significant development is needed to knit these systems together in ways that serve the needs of academic communities.

Our work with ESRC suggests that there will be a solution, in the foreseeable future, which will enable institutionally based repositories help their ESRC funded researchers fulfil ESRC's deposit requirements. It looks likely that the SWORD protocol will be relevant in achieving this solution. Institutional repositories should be looking at how best to acquire funder and grant metadata to effectively service ESRC's – and other funders' – deposit requirements. The funder acknowledgement format laid out by RIN could form the basis of a recommended minimum metadata standard for institutional repositories to implement.

[40] York Digital Library (YODL) http://www.york.ac.uk/library/electroniclibrary/yorkdigitallibraryyodl/

References

Björk, B-C., Roosr, A. & Lauri, M. (2008) Global Annual Volume of Peer Reviewed Scholarly Articles and the Share Available Via Different Open Access Options. Proceedings ELPUB 2008 Conference on Electronic Publishing - Toronto, Canada - June 2008, p. 178-186
http://elpub.scix.net/data/works/att/178_elpub2008.content.pdf

Davis, P.M. & Connelly, M.J.L. (2007) Evaluating the Reasons for Non-use of Cornell University's Installation of DSpace. D-Lib Magazine, 13 (3/4).
http://www.dlib.org/dlib/march07/davis/03davis.html

Harnad, S. (2006) Publish or Perish - Self-Archive to Flourish: The Green Route to Open Access. ERCIM News 64
http://www.ercim.org/publication/Ercim_News/enw64/harnad.html

Nixon, W & Greig, M. (2005) Populating the Glasgow ePrints Service: A Mediated Model and Workflow
https://dspace.gla.ac.uk/handle/1905/387

Sale, A. (2007) The patchwork mandate. D-Lib Magazine, 13 (1/2).
http://www.dlib.org/dlib/january07/sale/01sale.html

Simpson, P. (2005) *TARDis Project Final Report.* Southampton UK, University of Southampton, University Library, 14pp.
http://eprints.soton.ac.uk/16122/

Swan, A. & Brown, S. (2005) Open access self-archiving: An author study. [Departmental Technical Report].
http://cogprints.org/4385/

Xia, J. (2008) A Comparison of Subject and Institutional Repositories in Self-archiving Practices. The Journal of Academic Librarianship, 34 (6), p 489-495.
http://dx.doi.org/10.1016/j.acalib.2008.09.016

Appendices

Appendix 1- Researcher questionnaire

This questionnaire is for academic / research staff and research students from the Universities of Leeds, Sheffield and York.

Please help us understand what information about research publications you put online by filling out this questionnaire. This will allow us to assess how best White Rose Research Online can serve researchers' needs. The questionnaire will be followed up with some interviews and case studies. If you would like to be involved further please do let us know below. You can also enter our free prize draw and win an iPod Shuffle or £50 gift voucher. All the questions are on this one page. Please click the Continue button at the bottom of the questionnaire when you have finished.

Background Information

1. Which institution are you based at?
University of Leeds : University of Sheffield : University of York

2. Please state the name of your department / research centre.

3. Which of the following best describes your role at the University?
- Professor, Reader, Senior Lecturer, Senior Research Fellow
- Lecturer, Research Fellow
- Graduate Student, Post-doctoral Researcher
- Other (please specify):

Research Publications Online

4. Do you list your research publications on an individual staff webpage?
No (please go to Q.7) : Yes : Other (please specify):

5. What information about your publications do you have on your staff webpage? (select all that apply)
- Bibliographic information
- Links to full text (e.g. PDF or Word files you have attached)
- Links to full text elsewhere
- Other (please specify):

6. How often do you update the research outputs information on your staff webpage?
- Immediately
- Every month
- Every semester
- Annually
- Other (please specify):

7. What types of publications are listed on your department's website? (select all that apply)
My department does not list publications on its website (go to Q.10)
- Journal articles
- Book chapters
- Books
- Theses
- Working papers
- Conference papers
- Research reports
- Other (please specify):

8. Whose responsibility is it to update information on the departmental website? (select all that apply)

- Authors themselves
- Administrative staff
- Other members of the research team
- Don't know
- Other (please specify):

9. How often is the information updated on the departmental website? (select all that apply)
- Immediately
- Every month
- Every semester
- Every year
- Don't know
- Other (please specify):

10. Do you submit details of your publications to any other University, faculty or departmental database? (select all that apply)
- Yes
- No
- ULPD (University of Leeds staff only)
- Somebody does it on my behalf (please state who below)

If yes, please give details of where you deposit. If someone else deposits on your behalf please give details of their role.

11. Do you submit details of your publications to any systems or databases outside the University?
- No (go to Q.12)
- Yes
- Somebody does it on my behalf (please specify who below)
- Other (please specify):

If yes, please specify where you submit and why. If someone deposits on your behalf please specify their role.

Open Access Repositories
Open Access archives, or repositories, provide free online access to full text versions of research publications. Researchers upload their own papers and add some descriptive data (often called metadata), such as author, journal title, volume, etc. These research publications can be found using conventional web search engines, such as Google, as well as other specialist search services. Users are then able to read, download, and print these papers without charge. The majority of papers have already been published, or accepted for publication. Many of the publications have been peer-reviewed. Open Access repositories are fully observant of copyright laws.

12. Would you like to have your research publications freely available online?
- Yes
- No
- Don't know
- Please say why / why not.

13. Do you have any subject specific Open Access repositories for your discipline area?
- Yes
- No
- Don't know

If yes, what are the subject specific repositories for your discipline area?

14. Had you heard of White Rose Research Online (http://eprints.whiterose.ac.uk/) before you received this questionnaire?
Yes : No

15. Have you deposited research publications in White Rose Research Online? (select all that apply)
- Yes - deposited myself
- Yes - someone has deposited on my behalf
- Yes - have deposited on behalf of someone else
- No
- Other (please specify):

If yes, what might encourage you to deposit more frequently?
If no, what might encourage you to deposit?

16. What types of materials would you like to be able to deposit in White Rose Research Online?

17. When, during the lifecycle of your research, would you like to deposit in White Rose Research Online?

18. What services might encourage you to use White Rose Research Online (more frequently)? (select all that apply)
- Statistics about your publications (e.g. number of times your papers have been downloaded)
- RSS feeds (e.g. notification when new papers are available in your subject area)
- Different export options (e.g. export search results to Endnote)
- Automatic feeds to other systems / repositories (e.g. deposit in WRRO means your papers are automatically deposited in your funder's repository)
- Customised reports (e.g. the ability to prepare reports on the output of your dept or colleagues)
- Links to your papers from your personal website
- Other (please specify):

19. What would make you want to deposit your research into White Rose Research Online on a regular basis?

20. If the university required you to deposit copies of your articles in White Rose Research Online or another open archive, what would be your reaction?
- I would comply willingly
- I would comply reluctantly
- I would not comply

Funding Bodies
21. Who are the major funders for your research area?

22. Are you aware of your funders' policies on Open Access deposit of research outputs?
- Yes
- No
- Some but not all
- Other (please specify):

If yes, please summarise what the policies are.

23. If your funder required you to deposit copies of your articles in White Rose Research Online or another open archive, what would be your reaction?
- I would comply willingly
- I would comply reluctantly
- I would not comply

And finally...
24. Would you be willing to be contacted by the project team in the future? This places you under no obligation to be involved in any further stages of the project.

- Yes
- No
- Other (please specify):

25. Do you want to be entered into the free prize draw?
Yes (please enter name and email address in Q.26)
No

26. Please enter your name and email address (optional).

27. Do you have any other comments or suggestions about the questionnaire, White Rose Research Online or access to research?

Appendix 2 – arXiv plug-in

An arXiv ID number is entered into the text box (multiple ids can be entered with a carriage return after each ID) or can be uploaded from a file (Figure 1). This then retrieves all the metadata associated with the particular citation ID (Figure 2).

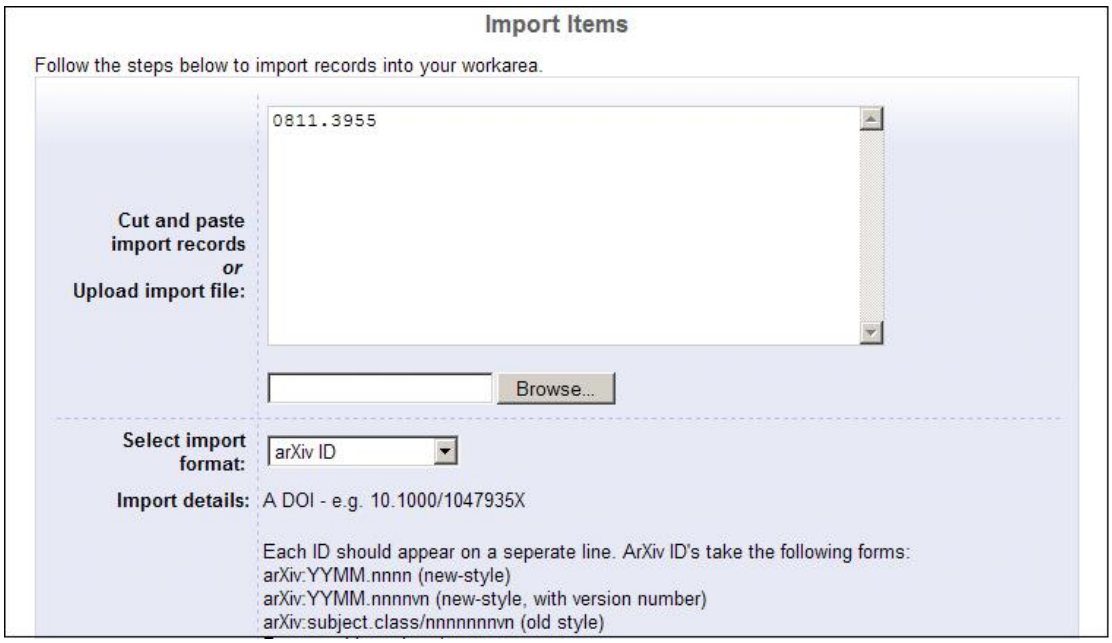


Figure 1. arXiv ID import text box from EPrints.

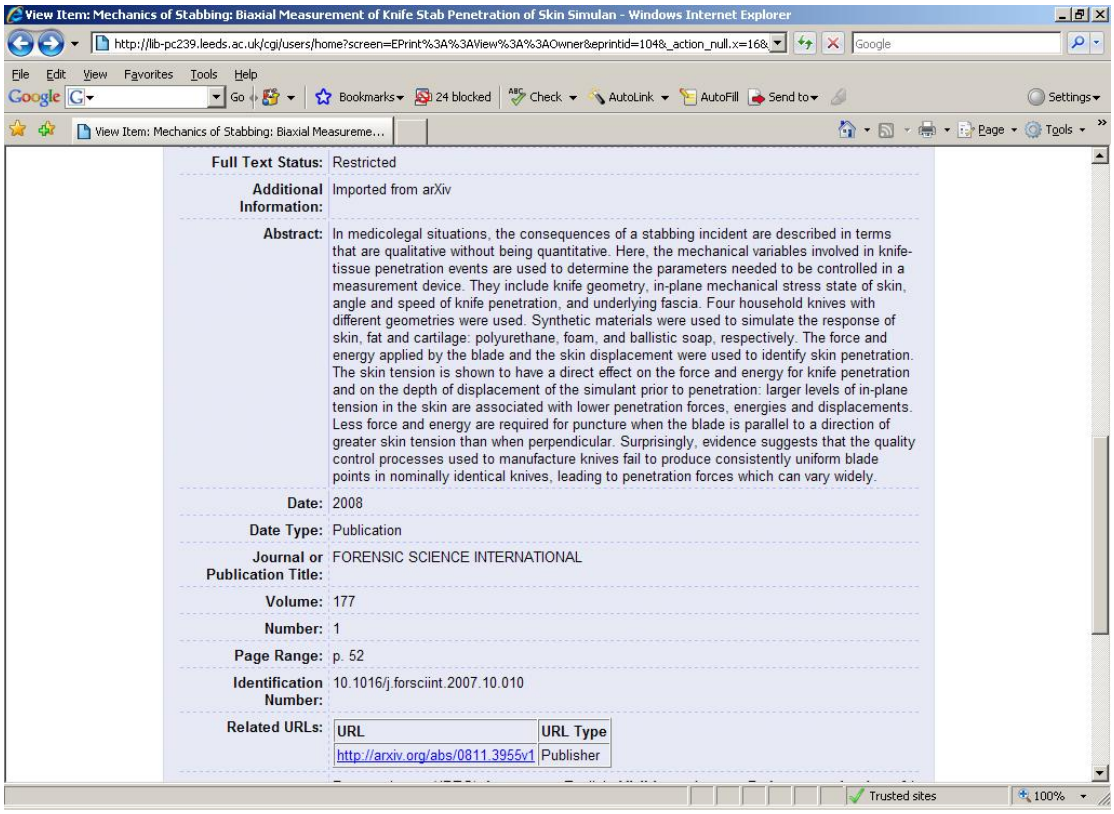


Figure2. Metadata imported from the arXiv ID import plug-in.

arXiv XML

The ArXiv XML plug-in maybe of potential benefit to users who want to search using the arXiv API and then save the resultant XML file. This file can either be uploaded or cut and pasted in the text box for import into EPrints(Figure 3).

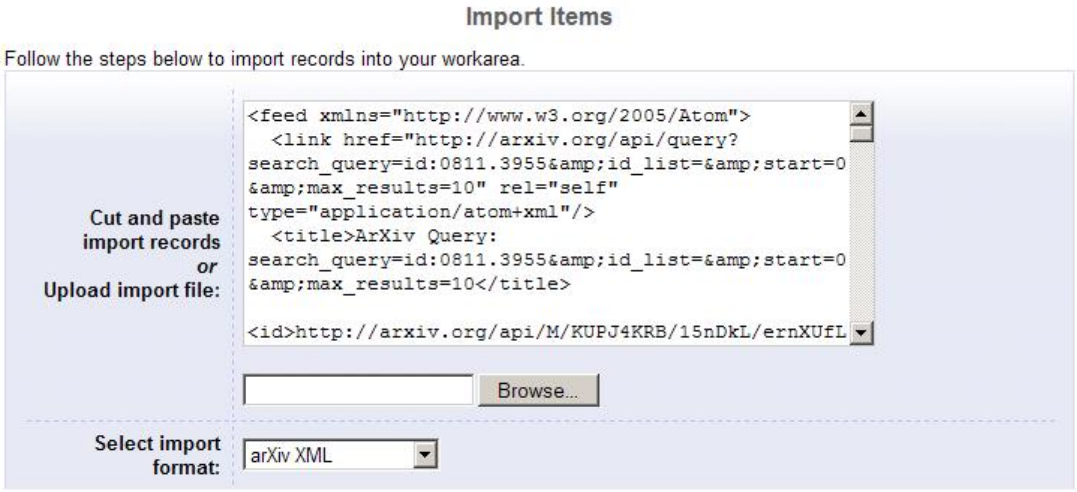


Figure 3. arXiv XML import plug-in text box.

arXiv Author

The arXiv Author plug-in can import all items associated with an author. The search can be carried out using the author's surname and can be made specific by using their initials (e.g. Name_A. [Figure 4]). The actual query to the arXiv API would be http://export.arxiv.org/api/query?search_query=au:Name_A. However, this can result in many false hits due to many people having the same surname or even the same initials. An additional issue with using names is that the use of initials in publications is not always consistent and therefore this can either result in greater number of hits some of which will be false positives or by being too specific can miss some publications.

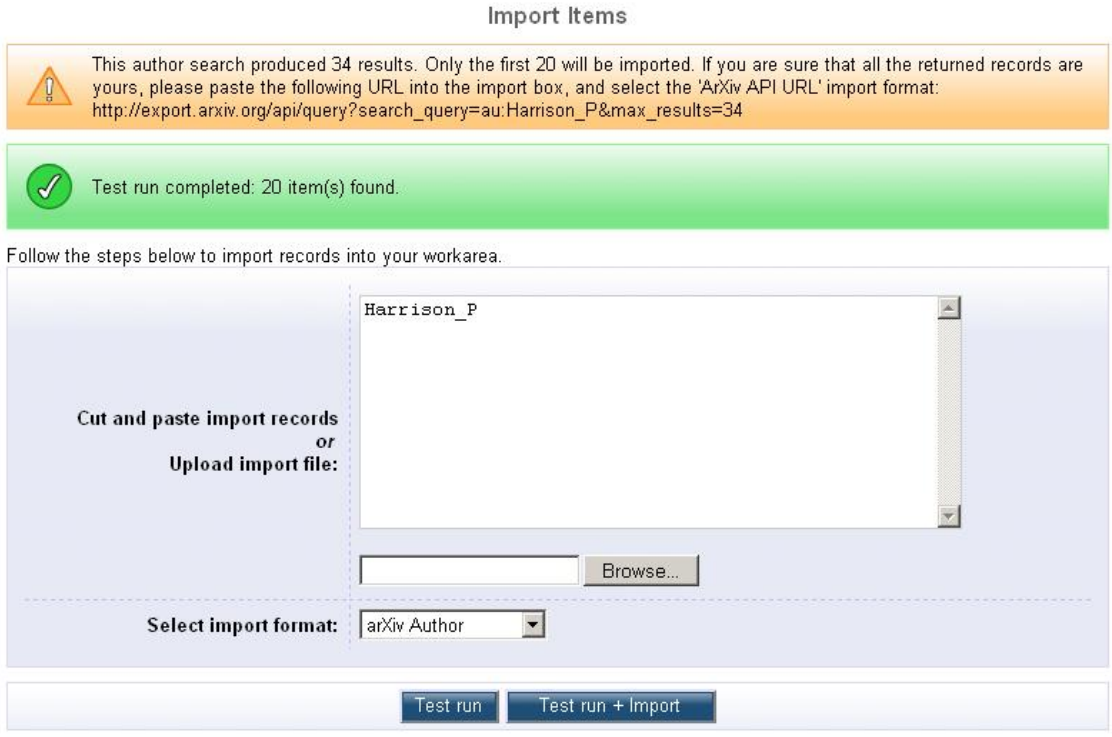


Figure 4. arXiv Author import plug-in.

The search can also be carried out using both the author's name and the institution to identify only authors from that particular institution (e.g. Name AND Sheffield [Figure 5]). The arXiv API query would be http://export.arxiv.org/api/query?search_query=au:Name+AND+au:Sheffield. Although if the authors leave the affiliation field blank then there is no way to identify the author uniquely and the only way to confirm if all the publications belong to that author would be to email them.

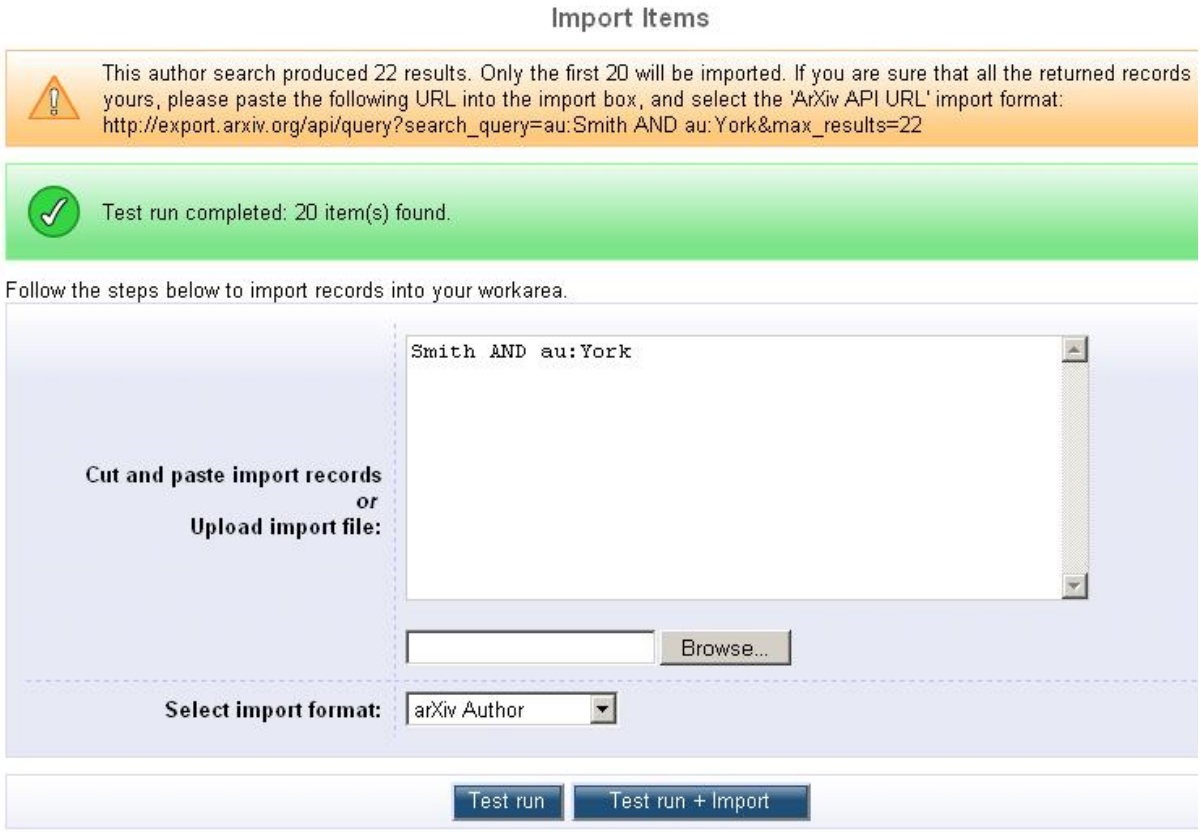


Figure 5. Importing items from arXiv using the author's name and the institution affiliation with the arXiv author plug-in.

An additional issue is that if a combined search was carried out using the author's name and affiliation for example Authorname AND Sheffield then a false positive hit would be returned if the author's name was Sheffield.

Furthermore the XML feed is set only to import a maximum of 20 items even though there may be a greater number of hits. A warning message is generated when this occurs informing users of the actual number of hits generated with that search and that they can import all the hits if desired pasting the URL (suggested in the warning message) in the arXiv API URL text box.

arXiv API URL

There is also capability to search the arXiv API directly from EPrints using the arXiv API URL. The user is required to type in the correct query URL to search the API (http://export.arxiv.org/api_help/, http://export.arxiv.org/api_help/docs/user-manual.html#query_details). This returns a list of hits, the default output is set to 10 results but if there are more hits than 10 then this figure is indicated in the results page. The user can then change the max results setting to the number of hits generated (http://export.arxiv.org/api/query?search_query=all:electron&start=0&max_results=10). They can then scan the results and discard items that are false positive and select those that are genuine and import them. Searching with this method would cut out the step of going to the arXiv API and carrying out the search first and then deciding if the results are genuine. Further work is planned to enhance this plug-in which would enable users to modify their query depending on the results and be able to submit or discard citations (by checking the little box next to each hit) for import into EPrints.

Appendix 3 – Exploring the relationship between WRRO and ESRC's repository

Aims

- To investigate deposit and/or harvesting scenarios with ESRC.
- To develop a demonstrator system showing how ESRC funded researchers can be assisted to deposit in both their local repository (WRRO) and ESRC's Awards and Outputs Repository.
- To inform the deposit process by increasing our understanding of ESRC funded researchers' workflows.
- To identify ESRC funded researchers in one or more consortium partner institutions with a view to using their research outputs to test the deposit/harvest mechanism.
- To seek feedback on the deposit mechanisms from ESRC funded researchers.
- To seek input on the proposed deposit mechanisms from JISC and the repository community.

Researcher workflow

Ideally, the depositing researcher should have ready access to relevant grant information: perhaps from a list of relevant grants linked to their authentication details or a list of grants associated with their school or faculty.

It should be straightforward to associate individual research outputs with a specific grant.

The researcher may wish to deposit the details of the work - and possibly the work itself - into a number of different places. For example, a subject repository external to their organisation (PubMed Central, arXiv); a repository associated with their research funder (PubMed Central; ESRC). Ideally, the researcher should be able to choose from a list of external deposit locations supported by the repository.

Scenarios

Five deposit scenarios were outlined: scenarios 1 and 2 assume local deposit into an institutional repository; scenarios 3 and 4 assume deposit into ESRC's repository; scenario 5 outlines desktop deposit.

Scenario 1

Local (institutional) deposit into WRRO; "post" into ESRC's awards and outputs repository using SWORD protocol
Description
An author deposits a journal article and descriptive metadata into WRRO. The metadata includes the ESRC Grant Name and Grant Number. The author indicates on a drop down menu that the work should be deposited with ESRC's Awards and Outputs Repository. This invokes a SWORD alert to push the metadata and files to ESRC (format to be defined by ESRC's SWORD service document but could be a METS file plus full text(s) in a zip file). The work is put into ESRC's A&O repository. WRRO receives a receipt from the A&O repository.
Comments
This scenario is readily achievable. ESRC will be supporting SWORD.

Scenario 2

Local (institutional) deposit; data harvested by ESRC's awards and outputs repository
Description
An author deposits a journal article and descriptive metadata into WRRO. The metadata includes the ESRC Grant Name and Grant Number.? Data is exposed so that it is harvestable using OAI-PMH. ESRC regularly harvests from WRRO and identifies new additions from their date stamp. The ESRC grant name and number metadata indicates that this is an ESRC relevant work. (Perhaps there is a different way to indicate locally which records should be harvested - particularly if ESRC will be harvesting relevant outputs which are non-ESRC funded).
Comments
OAI-PMH services are supported by ESRC, but for the immediate scope of this project it would have required additional funding to be identified in order to explore this scenario more fully. In addition, repositories may wish to share compound digital objects; OAI-ORE may be a more suitable standard to investigate in this context.

Scenario 3:

Remote deposit (into ESRC's repository); "post" into WRRO using SWORD protocol
Description
An author deposits their work into ESRC's Awards and Outputs Repository. The metadata indicates that the author has an affiliation with Leeds, Sheffield or York Universities. This invokes a SWORD alert to push the metadata and files to WRRO (format defined by White Rose's SWORD service document). The work is put into WRRO (probably via the editorial buffer rather than directly live?). ESRC receives a receipt from WRRO.
Comments
Not feasible at this time. There is very little SWORD support available for .net based systems with this feature available. (see Findings section below). Funding would therefore need to be identified to allow the development work to progress this scenario.

Scenario 4:

Remote deposit (into ESRC's repository); harvest / bulk import by WRRO
Description
An author deposits a journal article and descriptive metadata into ESRC's Awards and Outputs repository. The metadata indicates that the author has an affiliation with Leeds, Sheffield or York Universities. Data is exposed so that it is harvestable using OAI-PMH. WRRO regularly harvests from ESRC's repository using OAI-PMH - or imports a set of records using another protocol - identifying new additions from their date stamp. The WRRO affiliation metadata indicates that this is a WRRO relevant work.
Comments
ESRC has an active OAI-PMH service. If the researcher informed Leeds/Sheffield/York of their grant number, the University could harvest the records using OAI-PMH. This solution in technically feasible though OAI-PMH harvesting has not been implemented in WRRO.

Scenario 5:

Desktop Deposit
Description
Another possible workflow would see research deposited from the author's desktop. For example, there are some interesting developments coming from Microsoft which could allow repository deposit from within Microsoft Word and other common programmes.[41] This may be a longer term solution, potentially supplying content to multiple destinations from within the creation software or enabling simple desktop drag and drop tools to capture outputs. To work, the deposit destinations would have to have defined what metadata is required from the supplier and what type of authentication and verification mechanisms may be needed in order for the remote system to accept the deposit. As with the other scenarios, there is the question of the quality of metadata (whether generated automatically or created specifically by the depositor) and the extent to which processes can be automated. Would desktop deposit still require mediation? If so, would the mediation lie at the local level (the depositor's institution) or be performed remotely by those services accepting deposit?
Comments
Investigating this scenario was out of scope for the project.

Progress so far

WRRO provided a sample set of records to ESRC for testing:
(i) an eprints XML file with the document Base64 encoded into it
(ii) a METS file with a URL to the document

ESRC's in-house repository runs in the .net environment. This raised some technical issues: zip files were not supported.

For the purpose of proof of concept, ASP and the Microsoft XML HTTP component were utilised, instead of .net.

Current implementations of SWORD and its test clients don't include a .net version. . EPrints and DSpace have plug-ins available to achieve support, but other repository software can require significant redevelopment. However, the advent of open source .net ATOMPub servers and the adoption of community standards, including SWORD, within Microsoft products, should make deployment of SWORD more straightforward in future.

Supported by WRRO, the Research Council has completed a proof of concept project for extending the ESRC Awards and Outputs repository. ESRC are currently in active talks with Microsoft about the use and implementation of their new open-source Research Output Repository Platform. It is their aim to deploy SWORD functionality in the near future, and to allow any SWORD Compliant repository to deposit ESRC funded outputs into their systems. WRRO will continue to maintain this close relationship with the ESRC. The IncReASe project web site will be updated with any further developments.

WRRO envisage that once ESRC are SWORD enabled then we would request a service document and if that is sent back then we would make a deposit of the file in a format that ESRC can deal with. Once a successful deposit has been made then we would get sent a reference number which we would store in the EPrints data.

[41] E.g. see Microsoft's Scholarly Communication website at http://www.microsoft.com/mscorp/tc/scholarly_communication.mspx

Some issues for further consideration

If WRRO or other institutional based repositories utilise this scenario, it would be worth considering:

(i) what local action invokes the post action to ESRC? E.g. would it be the act of publishing a work to the public area of the repository (the "Live Archive" in Eprints terms). Would it be the point at which the attached file is made live (see (vi) below)?

(ii) how will embargoed works be handled? NB ESRC's advice in their Open Access Policy (Submission Policy) is that depositors should not deposit work within the period of embargo.

(iii) will the process handle metadata only records - for example, where open access is not permitted under the terms and conditions of the publisher agreement?

(iv) if the files cannot be made live (as in (iii) above) should they be supplied in any case for ESRC's internal records? Are there any copyright implications in supplying copies for this purpose?

(v) how will amendments or withdrawals (from either repository) be handled? E.g. changes to metadata; addition of files to a repository record. Could the ESRC supplied reference number for each record form the basis for data synchronisation?

(vi) at what point in its lifespan should a journal article or conference paper be deposited with ESRC? E.g. a work arising from an ESRC grant is submitted to a Taylor and Francis journal. The pre-refereed, submitted file is submitted to the local repository. It may be made available, but will have partial metadata and may be superseded by a subsequent version of the metadata record - for example, if the paper is rejected. Would ESRC wish to handle the file at this point?

The paper is accepted for publication. A metadata record for the accepted version of the work is created. The accepted file, including changes made as a result of the peer review process, is uploaded to the repository and embargoed. According to the publisher's policy it should not be made live until 18 months post publication. At this point, the metadata may be complete or it could be that the final volume, date, pagination details for the work are not yet known. Would ESRC wish to handle the file at this point?

The work is published and metadata completed within the local repository. The embargoed file becomes "live" 18 months post publication. Is this the point of deposit to the funder repository?

NB ESRC's Open Access Policy (Content Policy) states that "deposited items may include:

• working drafts
• submitted versions (as sent to journals for peer-review)
• accepted versions (author's final peer-reviewed drafts)
• published versions (publisher-created files)"

(vii) What quality criteria will we need to meet for ESRC: who will perform data checking (depositor/local repository/ESRC); who will perform copyright checking (depositor/local repository/ESRC)?

(viii) If there is a difference between what a publisher routinely allows in an institutional repository and what may be deposited in a subject or other third party repository, how will this be handled?

(ix) Should we be looking beyond compliance with the ESRC mandate and facilitating the deposit of more than journal papers and published conference papers? Is it envisaged that depositors will be able to deposit a range of relevant materials locally (including, for example, datasets, audio-visual materials) which will be made available to ESRC in an automated way? Although ESRC's deposit mandate refers specifically to just two publication types, ESRC's depositors are able to deposit over 40 different output types into ESRC's repository.

Trusted Relationship

Currently, ESRC depositors create an account, upload work to the ESRC's repository and this is then checked, validated and enhanced as required by repository staff. The onus is on depositors to clarify copyright conditions relating to their deposit (though, in practice, checking is undertaken where necessary by repository staff).

In addition to technical aspects of establishing a link between two repository systems, there are questions of quality checking and workflow which will need to be clarified. What criteria will a supplying repository need to fulfil before ESRC will be willing to accept a deposit (or vice versa)? It will be useful to have an agreed set of item types and minimum metadata fields. Institutional repositories will vary in the level of quality checking and copyright checking they wish to / are able to provide for depositors. However, it may be helpful for the ESRC to have some idea of the level of checking already performed on a record before onward supply to their repository - otherwise there is likely to be considerable duplication of effort. Because publishers sometimes make a distinction between what may be deposited in an institutional system and what may be deposited in a funder/subject repository, double copyright checks may simply be inevitable.

Metadata

Metadata fields - but also standards applied within those fields - may vary between repositories. One example would be capitalisation conventions for titles of works. ESRC's repository staff currently apply AACR2 when creating metadata for research outputs.

Should supplying repositories ensure their metadata conforms to a minimum standard specified by the receiving repository? This could be tricky where adopted standards vary from repository to repository.

The ESRC deposit mandate refers specifically to journal papers and published conference papers. These are relatively simple item types and there is a good match between the fields we capture in WRRO and the fields required within ESRC's repository. The main difference is that ESRC's staff supplement the bibliographic metadata through the addition of keywords from the International Bibliography of Social Sciences and Humanities and Social Science Electronic Thesaurus (HASSET).

Glossary

arXiv: an online archive of openly accessible research papers in the fields of Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance and Statistics. Run by Cornell University.

API: Application programming interface - a set of routines, data structures, object classes and/or protocols provided by libraries and/or operating system services in order to support the building of applications[42].

CrossRef: a membership organization (directed by publishers) and the digital object identifier (DOI) registration agency for scholarly publications.

DOI: digital object identifier – a unique and persistent alphanumeric string used to identify digital content e.g. an online journal article or book chapter.

DROID: Digital Record Object Identification – a software tool developed by The National Archives which can perform automated batch identification of file formats.

EndNote: a software package for organizing bibliographic references.

JHOVE: the JSTOR/Harvard Object Validation Environment – identifies and validates file formats and determines the format-specific significant properties of an object of a given format.

OAI-PMH: Open Archives Initiative Protocol for Metadata Harvesting - a protocol for collecting metadata from a repository / repositories.

Perl: a programming language.

RA2: a subset of data collected as part of the RAE listing the submitted research outputs.

RAE: the Research Assessment Exercise – an assessment of research quality applied to UK Higher Education Institutions by the four national higher education funding bodies; used to calculate how much research funding each institution should receive.

REF: the Research Excellence Framework – the replacement for the RAE; likely to run in 2013 but the final details are still to be announced.

RefBase: web based reference management software.

SWAP: Scholarly Works Application Profile is an application profile to describe scholarly works or eprints using Dublin Core.

SWORD: Simple Web-service Offering Repository Deposit: a protocol, based on the Atom Publishing Protocol, for depositing into repositories and other systems.

Symplectic: a publication management system.

UKPMC: UK PubMedCentral – an online database of openly accessible research papers in biomedicine and life sciences.

[42] Definition from wikipedia http://en.wikipedia.org/wiki/API [accessed 20/02/09]