Repository Case History

University of Strathclyde Strathprints

Alan Dawson, Centre for Digital Library Research, University of Strathclyde
and Alan Slevin, Andersonian Library, University of Strathclyde

alan.dawson@strath.ac.uk

28 February 2008

Organisational context

The University of Strathclyde is a large, well-established university located in Glasgow, Scotland, with about 20000 students (full-time equivalent) and the usual mix of teaching and research. It is probably best known and strongest in business and engineering.

The repository's mission

Strathprints is an institutional eprint repository for making research papers and other scholarly publications widely available on the Internet.

Building a business case

The repository is not a business and its purpose is not to make money. The implicit plan was to keep costs low but make the repository valuable and effective quickly by adding a large body of content and demonstrating substantial usage ('if we build it, they will come'). The only additional cost to the University to date has been one full-time member of staff and one standard PC with 120Gb disk space (the server). The ideal is that the repository is viewed as a showcase for the University's research output and an important element in its overall objective of being a centre for research excellence.

Hosting and support

Hardware, operating system and backups are managed by the system support team of the Department of Computer and Information Sciences (CIS). It is a minor matter for them as they run several similar servers and the support load is minimal. Eprints and usage statistics software have been installed, configured and managed by the Centre for Digital Library Research (CDLR), which works closely with CIS and the University Library. One key issue here is that staff have direct access to configure eprints and underlying systems if necessary (though this is rarely required), as the hosting is local rather than with an external group.

Overview of current contents

Strathprints had 4746 records as at 27 Feb 2008, most of which describe published research papers: 3522 articles, 472 book chapters, 395 conference items, 177 monographs, 141 books, 12 theses, 9 patents and 18 other items. 1012 records (21.3%) currently have a full-text document available; 3734 have only metadata. Efforts are being made to increase this ratio. Most documents are in PDF, with only a few in Word or HTML. There are currently a further 156 records in the submission buffer and 145 in the deletion table.

Overview of current deposit activity

There are 202 registered users, of which 44 have either deposited an item or have had ownership of a previously deposited item assigned to them. Most records are uploaded to the submission buffer in batches by the Strathprints administrator in CDLR, using the `import_eprints` program. Corresponding department and LCC subject codes are also entered in batches, directly into the relevant tables of the MySQL database. The XML files required for import into eprints are generated automatically from an Access database, which in turn is used to import and augment records from departmental databases. Each record in the submission buffer is checked, edited and completed by library staff before being added to the live repository.

Title, author(s), type and year fields are regarded as essential for all items. For most records, the following fields are also required: title of journal, book, conference or equivalent (in which the item will appear), the journal volume and number, where applicable, and an official URL, which is regarded as especially useful where the full text is not held in Strathprints. An abstract and full text are regarded as highly desirable; locating and adding them is an important task for cataloguing staff. The University department (of the first author) and the subject classification (LCC) are mandatory and are added during bulk upload or by cataloguing staff. Other fields that are standard in Eprints are added if the metadata is readily available (page numbers, ISSN etc) but are regarded as less useful. Uncontrolled keywords are often added but their quality and value is questionable.

Developmental phases

Approval to proceed with an official institutional repository was given by the University's Head of Information Resources (Professor Derek Law) in May 2005. The task was assigned to an existing part-time member of CDLR staff (Alan Dawson). University approval was given for a two-year post of Institutional Repository Coordinator. A server was purchased and installed in June 2005, Eprints was installed and configured, and the first (genuine) record was submitted by a member of CDLR staff in July 2005. A small informal steering group of five people was convened in September and the first meeting held. The repository was open for registration, usage and harvesting in October 2005, with no publicity other than word of mouth. The Institutional Repository Coordinator (Alan Slevin) began work in November 2005, based in the University Library. A second steering group meeting was held, a service name chosen and a basic inclusive collection policy agreed. A development archive (Strathdev) was created in March 2006 (using the same installation of Eprints), to run alongside the main archive and enable testing of new interface designs, bulk imports and configuration changes such as the departmental browse interface. Strathdev was very useful initially but is only used now for testing occasional configuration changes.

Institutional embedding

Support and enthusiasm from a senior member of staff (Derek Law) has been important. Appointment of a full-time Institutional Repository Coordinator was a crucial step. When the initial two years expired in November 2007, this post was converted to a continuing one. By the time the first article about Strathprints appeared in the internal University newsletter, the repository already had 1000 records, and an award was given by the Principal to the depositor of the 1000th item. Presentations have been given to Library staff to ensure that Strathprints is viewed as an increasingly important new collection alongside other information resources, and the prominence of links to Strathprints on the Library and University websites has been enhanced to reflect this.

Faculty engagement

There was little or no general publicity for the first year of operation, but the IR coordinator contacted faculties and departments individually as part of his advocacy role. This encouraged several departments to supply documents,

spreadsheets and databases of publications, though the metadata in these was patchy and never included abstracts or full text, so each one required metadata cleaning and enhancement. The departmental browse structure (added in September 2006) also helped, as it displays the numbers of records belonging to each department. The inclusion of Strathprints records in Google Scholar has been useful.

Although the IR was not used to host the RAE collection, Library staff have worked closely with the RAE team to ensure the metadata was accurate. This had the bonus of ensuring access to the RAE metadata. General awareness of the download statistics in Strathprints and speculation over the new metrics system to replace the RAE has also resulted in more formal contacts being established with some departments, including Mechanical Engineering, Physics and Photonics. Departmental research managers have made more formal arrangements to transfer their metadata and author final drafts to the IR coordinator on a regular basis. This is an encouraging development although the deposit process is still mediated.

Policy formulation

All deposits are added to the submission buffer, and only cataloguing staff may move them to the main archive, after they have checked each record and URL, added abstracts and other missing fields, and tried to obtain the full text (usually the author's final draft). This policy has been in place since day one and there are no plans to change it. Most other policies have been reactive, i.e. we decided to just go ahead and deal with issues as they arise rather than spend time worrying about potential problems in advance. For example, cover sheets for full-text PDF documents were added after about a year of service operation. These show the university affiliation and logo, copyright statement and usage policy. The metadata, data, content, submission and preservation policies were adapted from available Eprints templates, and a Strathprints Notice and Takedown Policy has been developed.

We have kept user registration open to all, and occasionally get rogue users registering (e.g. one recently with username 'Free Porn') which are simply deleted. We have recently changed selected user profiles so that they can edit their own records after submission, once we had found out how to do this in Eprints.

Service sustainability

Costs are kept low. Hardware cost £600, software is free. Only one member of staff works full-time on Strathprints; support and cataloguing staff fit in Strathprints work along with other duties. Periodic link-checking and de-duplication takes place, and several duplicate records have been deleted. Support from the Eprints team has been occasional but very useful, either via the mailing list archive or responses to direct email enquiries. The IR Coordinator attended the Eprints training course in February 2006. Migration to Eprints 3 is regarded as a significant step and a commitment to continued usage of Eprints, at least in the short term. This is expected to take several days work and is planned for the first half of 2008.

Measuring and demonstrating success

Interest from academic departments has grown steadily over two years, partly stimulated by the RAE. Usage statistics have been very important in demonstrating substantial and increasing usage of Strathprints. The Eprintstats package was installed in September 2006 and has worked very well, failing just once in 18 months. In the first half of 2007 there were over 182,000 abstract accesses and 28,200 downloads, an increase of over 500% on the same period in 2006. The statistical package has been an important tool in advertising the repository, particularly in demonstrating to academics the increased downloads where full-text access to author final drafts are available.

Key challenges faced

At an early stage we tested uploading records in bulk from departmental databases, as it seemed unlikely that reliance on individual voluntary deposits would be productive. This required a little development work, e.g. an Access module to convert a single field of multiple authors into separate author names, and to separate forenames and surnames, as required for importing into Eprints. Another module was written to generate files in Eprints XML format from Access. This process has worked very well, but metadata quality and consistency in departmental databases remains a significant issue. Invalid XML characters, such as smart quotes and long dashes, have been a minor but persistent nuisance.

As with most IRs working with no deposit mandate, the process of ensuring that full-text author final drafts are available in the repository has been difficult. The familiar problems of making some academics comfortable with the issues surrounding open access and copyright have been evident. The reluctance to change working practices is not unexpected but still a significant factor.

Major achievements

Significant number of records and amount of usage. Fifth largest UK university repository just two years after starting up. Increasing interest from departments as usage and awareness has grown. Continuing repository coordinator post. Regular OAI harvesting by aggregators such as OAIster.

Important unresolved issues

- **Metadata workflow:** There are various departmental and RAE databases as well as Strathprints, but as yet there is no single master database or strategic view of how best to manage records of publications, at an individual, departmental, faculty or institutional level. There is no doubt that the publication metadata held in Strathprints is far more complete and accurate than that held elsewhere in the University, but it is not currently a master publication database.
- **Mediation:** Cataloguer support can be difficult to estimate as it depends on the workload of staff who have been trained in tagging Eprints. Intermittent tagging by some cataloguers can also have a negative effect on the consistency and accuracy of the metadata. We hope to address these problems in the future, possibly by deploying clerical staff to work more closely on maintenance of the repository.
- **Full text documents:** Efforts are being made to increase the proportion of full text, e.g. via faculty contacts and by emailing authors individually.
- **Unique item identifiers:** The default Eprints numbering system is used to identify records. This works well but is dependent on the software, so the URLs would change if different software were used. Identifiers in departmental databases are now added to Strathprints (a new field was added to Eprints for that purpose), which helps identify duplicates before and after uploading, but these are not unique or robust. A University-wide scheme for item identification would be useful but is difficult to specify and implement.
- **Author identifiers:** There is no option to browse by author and no system for uniquely identifying authors. There are plans to develop an authority file for author names, with Eprints 3 functionality in mind.
- **Subjects:** Subject browsing works well in some areas but is shallow (only the top levels of LCC are used), so becomes less useful as more records are added. Additional LCC subjects have been added in selected subject areas with large numbers of records, e.g. physics and some social sciences.

- **Theses:** There is currently discussion about how to handle theses, e.g. whether to join Ethos, about print and/or digital submission, how to enforce a deposit mandate, and whether to use Eprints or different software.
- **Advocacy:** With no institutional mandate for deposit in Strathprints, current advocacy efforts may only take the service so far in terms of full-text deposit of author final drafts. Therefore, certain departments where a good relationship has been developed, or who have a particular interest or expertise in developing departmental digital libraries, may continue to be the most enthusiastic contributors to the repository.
- **Integration with departmental websites:** This is starting to become an issue as some departments are generating websites and lists of publications from the university content management system, which is quite separate from Strathprints. This relates to wider University policy on a post-RAE integrated research information system.
- **Interoperability with the University VLE:** This is not an issue as yet but may become one in future.

Conclusions

It is not a bad idea to dive in and get on with it as long as there is some institutional support and initial costs are low. Eprints is ideal for this as it is freely available, relatively easy to install, reliable and highly configurable. It is possible to get a functioning repository up and running in a couple of weeks. The ability to interrogate (and occasionally update) the underlying MySQL database directly has been very useful, e.g. in extracting the figures for this report.

strathprints.strath.ac.uk