RepoMMan: delivering private repository space for day-to-day use

Richard Green and Chris Awre, e-Services Integration Group, University of Hull

Introduction

In the spring of 2005, the University of Hull embarked on the RepoMMan Project[1], a two-year JISC-funded[2] endeavour to investigate a number of aspects of user interaction with an institutional repository. The vision at Hull was, and is, of a repository placed at the heart of a web services architecture: a key component of a university's information management. In this vision the institutional repository provides not only a showcase for finished digital output, but also a workspace in which members of the university can, if they wish, develop those same materials.

The RepoMMan project set out to consider how a range of web services could be brought together to allow a user to interact easily with private workspace in an institutional repository and how the web services might ease the transition from a private work-in-progress to a formally exposed object in the repository complete with metadata. Three key decisions had been taken before the project proposal was submitted and will not be further discussed here: that open source software should be employed for the project, that the web services should be orchestrated by an implementation of the Business Process Execution Language (BPEL)[3] and that the Fedora repository software[4] should be used.

The remainder of this article describes the work of the project through to its eventual conclusion in September 2007 and briefly looks forward to follow-on work that will realise other parts of Hull's vision for an integrated repository.

User needs

The e-Services Integration Group (eSIG) at the University of Hull has a tradition of placing user needs at the centre of its development work (in contrast to some organisations where the technical work is carried out and then users are 'persuaded' that it is what they wanted). Consequently RepoMMan started with two key investigations to discover how potential users currently managed their work and what benefits they might gain from using a private repository space as part of their 'toolkit'. The first such investigation involved interviewing members of University staff. Colleagues were drawn initially from the research community but also later from the learning and teaching (L&T), and administrative communities. The interviews each took typically just over an hour and the findings were eventually synthesised into a first user needs report.[5] The second investigation involved a web-based survey covering the same ground but necessarily in less detail; this was initially restricted to Hull staff but then widened to the Higher Education community generally. This second piece of work allowed us to check that the needs expressed by our interviewees were typical of those in the wider world. The findings of the survey were reported separately.[6] Eventually, the findings of all the investigations were synthesised into a combined user needs report.[7]

The user needs work identified four overlapping areas in which a repository was seen as being of potential benefit to individuals: storage, access, management and preservation.

Users, particularly those in the research and L&T communities, were keen to have a storage solution that allowed them to access their works in progress from anywhere they could access the internet; this rather than have versions of their materials scattered variously across their office computer, their home computer, their laptop, memory sticks, CDs and (in a frightening number of cases) floppy disks. Further, this easy availability potentially meant that they could access their work in an ad hoc manner away from their normal work environment, perhaps at meetings, conferences, or in a lecture theatre. Many of those we talked to had well developed strategies for managing developmental versions of work, others were keen to have a system that would allow them easily to identify versions and, if necessary, to revert to a previous stage. Finally, many could see the potential in a system that would easily take their newly finished work, help them add metadata to it and publish it into a repository space where it could be managed, and potentially preserved for the long-term.

The RepoMMan tool

The early technical work of the project concentrated on developing what became known as the RepoMMan tool; this is a browser-based rich internet application interface that allows a user to transfer work to and fro between their local machine and the repository. It was a key design goal of this work that the interface should be intuitive to use and, to this end, it was deliberately modelled on a generic design common to ftp clients. The finished version is shown at figure 1.

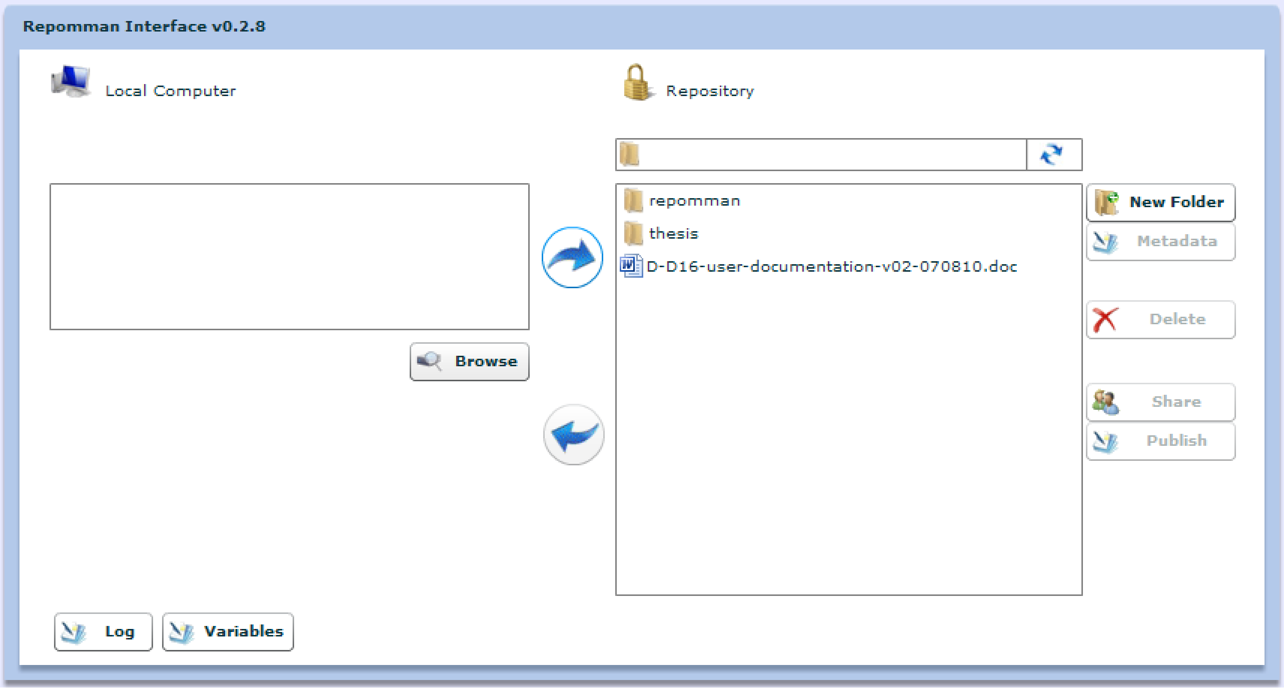

Figure 1: The RepoMMan browser interface

The left-hand side of the screen allows the user to browse their local computer whilst the right-hand side of the screen shows their private repository space, represented as a familiar file and folder structure. This is actually a visual metaphor for what is happening in the repository because the ‘files’ are actually digital objects, whilst the ‘folders’ are collection objects within the Fedora repository. The two ‘arrow’ buttons in the middle of the screen cause the files to be transferred in one direction or the other. ‘Files’ in the listing at the right-hand side can be double-clicked to expand a list of versions available and these versions can be manipulated individually.

When a user clicks to transfer a file to the repository, a BPEL engine orchestrates a number of web services to create and save a digital object on the repository server. This object consists largely of the datastream definitions, relationships and metadata; the binary content associated with the digital object (the actual document, image, or whatever) is time-stamped and transferred to the University’s storage area network (SAN). This time-stamping allows relatively easy storage of the ‘same’ document in multiple versions. Using the SAN for digital content, rather than the repository server itself, gives us access to a very large, extensible and secure file store.

Metadata

As noted in the introduction, another aspect of the project's work was to consider how the transfer of an object into a public-facing repository space could be eased by the RepoMMan approach. A crucial element of this transfer is the addition to the object of appropriate metadata; often this has been a stumbling block in repository workflows. Authors are generally not good at 'doing metadata' and, faced with a blank form to fill in on their screen, may well do the bare minimum.

The RepoMMan tool has been designed for use primarily from within the University of Hull's portal (based on uPortal[8]) or virtual learning or research environments (in the process of transfer to Sakai[9]). For the purposes of the project, work was undertaken to develop approaches for pre-populating administrative metadata (author's name, department etc) from the portal environment; because the user is logged in to University systems through the portal, web services can be used to derive useful metadata from them. When an author does have to provide metadata, perhaps about a specific research project, this is stored in their repository space so that it can be re-used with future objects.

The greatest challenge to our team was to provide a set of services that would add meaningful descriptive metadata into a digital object. When the project proposal was written we were aware of a number of initiatives to develop metadata extraction tools and we had imagined that, by the time we needed one, at least some of these would have borne fruit. In practice we found only one tool that seemed to do a reasonable job of providing metadata for an unseen document without the provision of a controlled subject vocabulary, this is the iVia metadata tool, a component of the Data Fountains suite.[10] We have deployed this tool as a web service and use it to prepopulate the descriptive metadata form that is presented to an author. It is our belief that an author is far more likely to correct and/or expand a filled-in form than to do a good job of filling in a completely blank one. Once the author 'saves' metadata for an object it can be accessed again for amendment but this does not re-run the web services unless the author specifically requests it; thus once an author has customised metadata it is not overwritten by our toolset. For non-text content we currently still need the authors to provide descriptive metadata themselves.

Investigated, but not actually implemented in the lifetime of the project, was a further web service which will use Harvard University's JHOVE tool[11] to populate a hidden object datastream with technical and/or preservation metadata - depending on the content type.

Problems? What problems?

Set out in a few brief paragraphs, it might seem that the RepoMMan Project proceeded in a fairly straightforward and trouble-free manner but the careful reader may have noticed in the introduction that this 'two-year' project appears to take more than two years; it did and much has been learned along the way.[12] It transpired that, to a large extent, the project had to pioneer the use of SOAP-based[13] web services with Fedora repository software. Almost inevitably, we encountered a number of problems, which took a considerable time, and a considerable amount of joint effort with the Fedora community worldwide, to solve. We again visited uncharted (and somewhat troubled) waters when we became one of the first teams to try and set up Fedora for use by a very large number of independent users using an LDAP server[14]. Last, but not least, we were the first to try and deploy the iVia metadata tool on a modest 32-bit computer (rather than the 64-bit system on which it had been designed); some time was spent working with developers in the US to make this work. All these processes took time to resolve and, in the event, an unfunded extension was granted from the JISC in order to complete the work. It is good to be able to report that the improvements made to Fedora and to the iVia tool have now been made available to the wider community and that the experience of interacting with the various development teams was generally very positive.

Software

The RepoMMan tool uses the open source ActiveBPEL engine to orchestrate a range of web services in order to achieve its tasks; some of these web services are necessarily unique to the University of Hull and for that reason it is not possible to offer the tool as an off-the-shelf package to others. That said, the RepoMMan team will be more than willing to share its experience and code with groups who are interested in following a similar approach to repository work. Whilst not necessarily directly transferable, our work may be a useful starting point for others.

Future work

The RepoMMan tool, as developed to meet the requirements of the JISC project, stops short of the actual process by which an object can be transferred from a private into a public-facing repository. This work, represented in simplified overview at Figure 2, is currently being undertaken in-house as part of the wider work that is developing the University's institutional repository infrastructure.

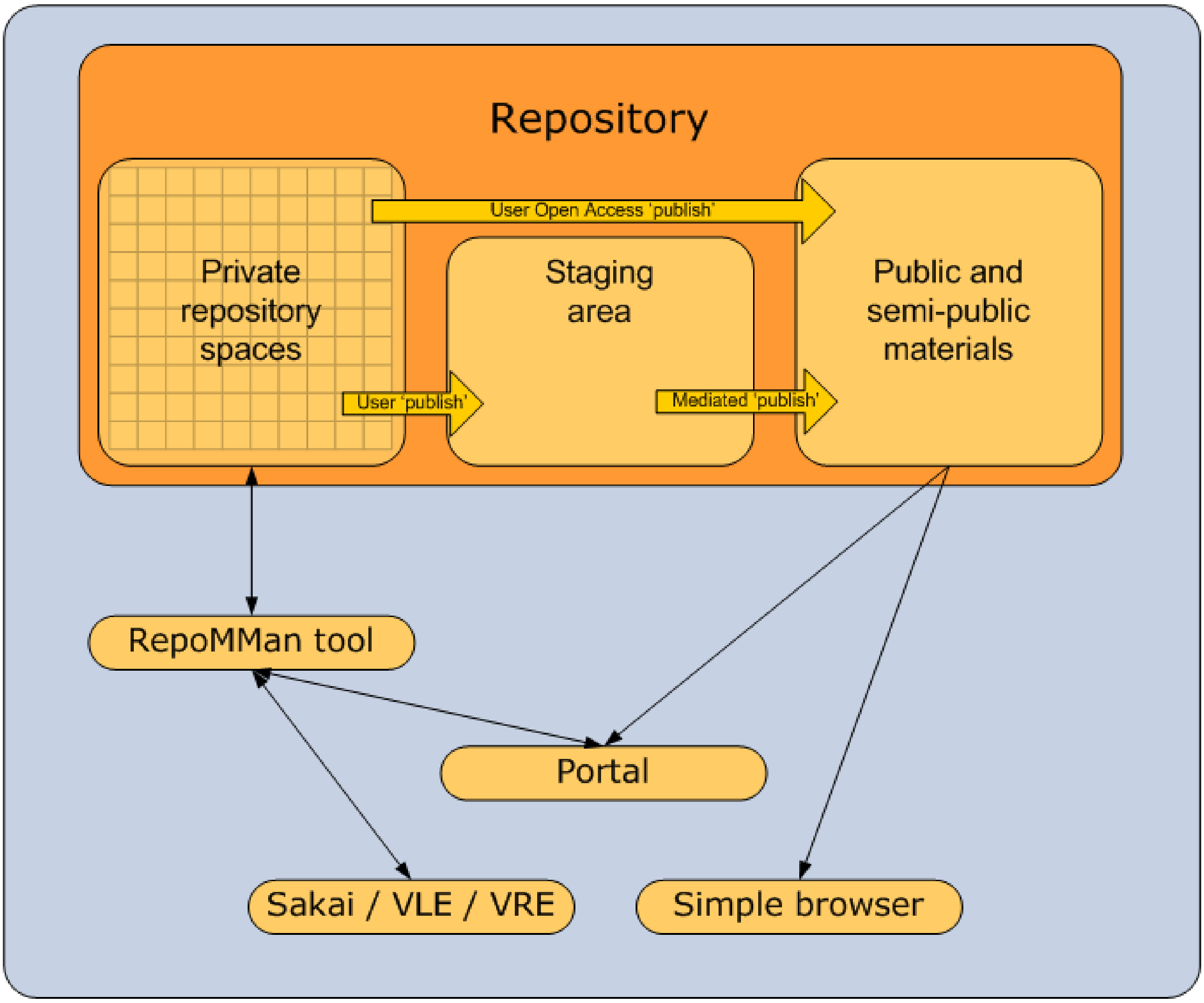

Figure 2: Overview of the process to expose authors' completed materials

The author will be able to 'publish' their work into the University repository via a staging area. (Ultimately, we shall provide a direct route specifically to support open access publishing.) After being checked over, the object will then be transferred by repository staff to the public-facing repository where it will available to others subject to appropriate access rights. This 'publishing' process, which will again be carried out by BPEL-orchestrated web services, involves copying the author's digital object and using it to construct a new object 'owned' by the repository and conforming to strict ‘object models’ (internal structures). The RepoMMan team has received a further two-year grant from the JISC to build into this process a system for ongoing records management and digital preservation (RMDP). This is the REMAP Project,[15] which is due to complete in March 2009.

RMDP will be supported by building into the new digital object a set of date flags which will be used by an external system to trigger records management and/or preservation events. Thus, for example, when a past undergraduate examination paper has been in the repository for five years the a date flag might trigger a message to the appropriate departmental secretary that the paper is likely out of date and should therefore be hidden from future students; the secretary may wish to do this or to extend its life by resetting the flag. It may be that if the secretary does nothing the system will take a default action, in this case to hide the paper. As another example, a date flag may trigger a process to undertake a format review of the materials in the repository; this automated review may result in a message to administrators that there are some specific objects in a file format that is no longer widely supported and suggest that something should be done. The University’s REMAP team, working with colleagues in Spoken Word Services at Glasgow Caledonian University, is investigating a range of possibilities.

The Institutional Repository

Finally, and briefly, it should be noted that all this work comes to nothing without an accessible Institutional Repository at the end of the process. We have noted that Hull’s vision is of a repository at the heart of an integrated web services architecture; as such there will be many ways to access materials within it, and Fedora offers the capability of exposing them through multiple interfaces. It had been our plan to write a simple browser interface as one of the routes into repository content. In the event we have been fortunate to coincide with the development of the Mura interface[16] at Macquarie University in Australia, which so closely fitted our needs that independent development would have been unjustifiable; rather we are contributing in a small way to the development work there. A test implementation of Hull’s Institutional Repository running with the Mura interface can be found at http://edocs.hull.ac.uk and this is being used in addition to our project websites to make available selected reports from the RepoMMan and REMAP Projects (follow the link to ‘JISC Repositories Projects’).

Author details

Richard Green
Manager, RepoMMan and REMAP Projects
c/o eSIG, Academic Services
Brynmor Jones Library
University of Hull
Hull   HU6 7RX

Email: r.green@hull.ac.uk
Web sites: http://www.hull.ac.uk/esig/repomman
http://www.hull.ac.uk/remap

Richard Green is an independent consultant working with the eSIG team.

Chris Awre
Information Architect
eSIG, Academic Services
Brynmor Jones Library
University of Hull
Hull   HU6 7RX

Email: c.awre@hull.ac.uk
Web sites: http://www.hull.ac.uk/esig/repomman
http://www.hull.ac.uk/remap

References

[1] See: http://www.hull.ac.uk/esig/repomman
[2] See: http://www.jisc.ac.uk
[3] The Active Endpoints website at http://www.active-endpoints.com is a good starting point to investigate BPEL. RepoMMan uses the Active Endpoints open source BPEL engine.
[4] Now Fedora Commons. See: http://www.fedora-commons.org/
[5] Green R (2006) *R-D4 Report on research user requirements interview data* University of Hull: see http://www.hull.ac.uk/esig/repomman/documents
[6] Green R (2006) *R-D3 Report on research user requirements on-line survey* University of Hull: see http://www.hull.ac.uk/esig/repomman/documents
[7] Green R & Awre C (2007) *R-D14 RepoMMan User Needs Analysis* University of Hull: see http://www.hull.ac.uk/esig/repomman/documents
[8] See: http://www.uportal.org
[9] See: http://sakaiproject.org
[10] See: http://dfnsdl.ucr.edu
[11] See: http://hul.harvard.edu/jhove/
[12] Readers wishing further details might like to consult: Green R, Awre C, Dolphin I, Lamb S & Sherratt R (2007) *RepoMMan Final Project Report* University of Hull. See: http://www.hull.ac.uk/esig/repomman/documents
[13] SOAP: Simple Object Access Protocol
[14] LDAP: Lightweight Directory Access Protocol, commonly used to support authentication processes
[15] See: http://www.hull.ac.uk/remap
[16] Now known as Muradora. See: http://drama.ramp.org.au