The RepoMMan Project: automating workflow and metadata for an institutional repository

Richard Green, Ian Dolphin, Chris Awre, Robert Sherratt
University of Hull

This article is published as:

Green R, Dolphin I, Awre C and Sherratt R (2007) The RepoMMan Project: automating workflow and metadata for an institutional repository *OCLC Systems and Services* Volume 23 Number 2 pp212-215

The RepoMMan project[1] has been funded from June 2005 to May 2007 as part of the Digital Repositories Programme managed by the United Kingdom's Joint Information Systems Committee (JISC). RepoMMan is closely aligned with the deployment of the Fedora digital repository system within the University of Hull, and seeks to address two areas of functionality that have generated wide interest in the repository arena: workflow and automated metadata generation. In this article we describe the work of the RepoMMan project, and place this development in the context of wider infrastructural developments within the University that aim to establish the repository as an underpinning part of work and study. The RepoMMan project is being led by the University of Hull's e-Services Integration Group (e-SIG).

The open source Fedora digital repository system[23] is a framework upon which many repository services can be built. This flexibility requires a degree of development and configuration to meet the particular requirements at hand, but also provides a powerful basis for enabling technology to meet user needs rather than having the needs adapt to meet technology. Fedora is built around a powerful digital object model[4], where individual digital objects can be a combination of any number of data streams: these digital objects may be local to the repository or referenced elsewhere. Because of this high level of granularity within the repository, Fedora allows detailed management of associated metadata and relationships between digital objects. This approach, coupled with a detailed security architecture and compliance with the OAIS Reference Model, suited the desire at the University to implement a repository that could be used for a variety of purposes to support the institution's work.

Managing a flexible repository does, though, bring with it the need to provide ease of use across a number of user groups, and establish agreed workflows. In considering how a repository is going to be used in any particular use case there will be a number of steps involved. Each task will have its own series of steps, and these need to be followed to achieve the desired aim. Commercial repository software can build in workflow capability that allows these steps to be followed seamlessly and simply, but this is often quite specific to the areas of functionality provided by that software. The RepoMMan project is developing an open standards-based, flexible workflow tool through which users can interact with Fedora in a configurable way. The tool is based on a combination of approaches: it is making use of Fedora's Web Service API interfaces[5] and orchestrating a series of Web Service calls to these interfaces using BPEL (Business Process Execution Language)[6]. The project intends that access to the workflow tool should be through the University's institutional portal (based on uPortal[7]) and the Sakai Collaboration & Learning Environment[8].

It is central to our approach at RepoMMan that the workflow tool should reflect user needs. To this end a survey and a set of interviews with researchers have provided evidence of how research is carried out and how research documentation and information is managed, from research initiation through to publication. The development of the workflow tool is centred on how the repository can be provided as a working tool on a day-to-day basis throughout the research process, and not simply as a store for documents at the end of this. Information will also be gathered from two further groups: those in the teaching and learning community and colleagues in Administration. Acknowledging that the University's Library staff may well eventually have a role in managing parts of the repository workflow, a working party has been set up in parallel with the project to involve them from the earliest phases of the design work: this group also includes representatives from staff involved in Archives and Records Management.

The interviews with researchers at Hull attempted to identify how they "did research", how they used IT tools in the process, and thus how a repository might be able to help them. A range of researchers from different subject backgrounds took part. In some cases the researchers were interviewed singly, sometimes in pairs. The interviews, which took anything up to 90 minutes each, identified a number of areas where researchers felt that a workflow tool and repository might aid them as they went about the development of a new idea; from these the RepoMMan team has attempted to distil a set of common needs[9]. In addition to the interviews, an on-line survey was developed to explore a similar range of issues with researchers elsewhere. This received almost 300 responses, largely from researchers at institutions within the UK, and these tended to reinforce the results of our interviews[10].

An outline of results from this user requirements work can be described as follows. Consider the example of a researcher developing a paper in collaboration with a colleague at another university.

A 'typical' researcher works on a new idea in any number of places. Sometimes they work on it in their office at the University, often work is done at home, sometimes work is done in other places. At present, this flexibility requires the researcher to carry around the latest version of the paper, maybe on a laptop but frequently on some form of portable storage - most often a USB memory stick of some sort. Interviewees told us that they would welcome a system that would allow them access to their current document from anywhere with an internet connection without the need to carry a copy with them. Hull's repository will provide them with a private development area (a 'My Repository,' if you will) in which they can keep their work-in-progress. Using the workflow tool they will be able to access the contents of this area via a browser. The fact that their files will be stored within the repository also addresses one of their other concerns, that of backup. The University repository would be subject to the University's regular backup regimes and thus researchers could be assured that backup was being dealt with effectively.

The results highlighted the potential value of a repository for general day-to-day purposes. The RepoMMan workflow tool is being designed to adapt to these purposes as required. The tool will particularly be of benefit in the case of a piece of collaborative research involving one or more co-authors at other institutions. The tool will allow the Hull researcher to give others access to the document being developed, automatically dealing with issues of versioning and locking. When each version of a document is saved into the repository it will be time- and date-stamped; older versions are not automatically deleted, rather they are kept as part of an audit trail and can be recalled on request. If a co-author decides to work on the paper, the other potential user(s) are warned of this and the version in the repository is effectively locked until the co-author saves his changes. This process guards against the possibility of conflicting revisions.

Whilst we have used the example of a research paper to illustrate some of our plans it should not be thought that Hull's repository is intended to be used only for documents. The same principles can be applied to almost any form of computer file because Fedora offers great flexibility in the type of content that a digital object can contain. The survey work that was conducted with researchers identified the need to cater also for presentations, images, spreadsheets, statistics files, database files, multimedia and web pages as well as a range of less frequent needs. *(See Green (2005) [2] p7 ff)*

It should not be thought that all the researchers we talked to were equally enthusiastic about the possibility of this flexible workspace. Some had very well developed research methods and, understandably, did not want to alter them. There is no intention that researchers should be pressured into using this workspace, only that it should be available to those who want to make use of it. Others, though, could see how the organisational possibilities of a 'My

Repository' could help them collect and collate their research material and retrieve elements easily using its search facilities.

By whatever methods the research is developed, there will come a stage when it is ready to be included in the public-facing area of the repository and be made available on open or restricted access. Access will be by browse or search and make use of metadata associated with the digital objects.

The second area of functionality being investigated by RepoMMan is automated metadata generation to ensure the digital objects can be found when made public-facing. The creation of quality metadata is essential for the proper management and use of the public-facing repository. It is, though, also clearly recognised that it is not viable for this metadata to be entirely human-generated. Many different types of metadata can be created automatically. For example, technical metadata about digital objects can be gathered through tools such as JHOVE[11], whilst administrative metadata can be gathered through institutional profiles. If, as we intend, interaction with the RepoMMan workflow tool is managed through the University portal, the portal already knows who the user is and can pass this metadata on to be included in digital objects ingested through the portal into the repository. The third major category of metadata, descriptive metadata, is less easy to deal with; automated generation of this is still a holy grail for the most part. RepoMMan is investigating the different approaches that can be taken to enable automated descriptive metadata generation and will be testing possible alternatives to gather further information on requirements and practice.

Ideally, it would be wonderful to be able to pass the researcher's paper to a metadata tool that could analyse it and which would extract a useful set of keywords (albeit that the tool may first need to undergo some form of 'training'). Our research thus far indicates that this may not be the best approach. The alternative is to pass the paper to a tool which analyses it against some sort of controlled vocabulary, perhaps subject based, and generate a list of keywords. It seems that this latter approach is generally believed to produce metadata of a much higher quality. Indeed, the one serious contender that we found for a tool of the first type[12] has recently adopted the controlled vocabulary approach and its developers have published a comparison showing the improvement in metadata quality.

When our example researcher completes their paper, whether using 'My Repository' or not, the file is offered up for transfer to the public area. The RepoMMan tool will generate metadata in the three categories described, technical, administrative and descriptive, and display it to the user. The researcher can then change anything in the descriptive metadata that seems inadequate or inappropriate; this, we feel, is likely to produce better metadata than asking the researcher to write it from scratch. Once the job is done a clone of the digital object thus created can be moved into the public area of the repository, perhaps by way of a repository administrator for quality assurance and checking, or perhaps automatically. The clone is not 'owned' by the researcher and they have no ability to edit this public copy. By dealing with the object in this way, the administrators of the repository are given absolute control of its public content and, ultimately, the quality of that content.

We have dwelt at some length on the example of a university researcher, it is perhaps timely to reiterate that similar, tailored facilities will be offered to the University's teaching and learning community and to its administrative staff. The repository will, ultimately, hold a wide range of objects in private and public areas. The public areas specifically will be subject to a range of differing security implementations that will allow particular objects to be accessed only by appropriate groups. This is particularly relevant where the repository is being used to support restricted access to appropriate materials. Open access will also be one of the intended use cases for the repository at the University of Hull. The interviews with researchers highlighted a mixed set of views as to the value of open access, and this feature will thus be made available as required rather than enforced.

The development of a repository for the University of Hull is part of an ongoing aim to provide infrastructural components that can be used in a flexible fashion to enhance and facilitate business processes. Fedora provides a platform that can support a wide range of repository use cases and adapt to requirements. The institutional portal, itself supported by a web content management system, is another component, and the Sakai Collaboration & Learning Environment is being assessed to provide a further building block. All of these systems can be described as service-oriented in their approach, insofar that they are designed to allow services to be built with them (e.g., workflow and metadata generation on top of Fedora), rather than the systems dictate the services that will be presented. Service provision exists at the technical level, through the use of Web Services, and at the organisational level, through awareness within the University of systems that can be adapted to suit user needs. As much as users are aware of how the systems can serve them, RepoMMan is also demonstrating how Fedora can be delivered through the institutional portal and Sakai environment and how these different systems can interact. The service-oriented architectural approach is in many cases still an approach, and it is accepted that many systems do not yet fully adhere to this model. The potential such architecture offers, however, has encouraged ongoing adoption.

In conclusion, the public ‘face’ of Hull’s repository will not be a single, monolithic entity. Rather there will be many different ways of accessing the content, each designed to be appropriate to the needs of the content and the use to which it will be put. Fedora provides a service-oriented ability to work with digital content in a coherent and structured manner to support all aspects of storage, management, access, and preservation.

[1] The RepoMMan Project: http://www.hull.ac.uk/esig/repomman

[2] Staples, T., Wayland, R. and Payette, S. The Fedora Project: an Open-source Digital Object Management System, *D-Lib Magazine*, 2003 9 (4), available at http://www.dlib.org/dlib/april03/staples/04staples.html

[3] The Fedora Project: http://fedora.info

[4] Fedora White Paper: http://www.fedora.info/documents/WhitePaper/FedoraWhitePaper.pdf

[5] Fedora Web Service APIs: http://www.fedora.info/definitions/1/0/api/

[6] Business Process Execution Language (BPEL) background information: http://en.wikipedia.org/wiki/BPEL

[7] uPortal: http://www.uportal.org/

[8] Sakai: http://www.sakaiproject.org/

[9] Green R (2005) *Report on research user requirements interview data* RepoMMan Project, University of Hull At: http://www.hull.ac.uk/esig/repomman/documents

[10] Green, R (2005) *Report on research user requirements on-line survey* RepoMMan Project, University of Hull At: http://www.hull.ac.uk/esig/repomman/documents

[11] JHOVE, The JSTOR/Harvard Object Validation Environment: http://hul.harvard.edu/jhove/

[12] See the Kea Project at the New Zealand Digital Library: http://www.nzdl.org/Kea/