The RepoMMan Project

Richard Green, Chris Awre, Ian Dolphin, Robert Sherratt
e-Services Integration Group, University of Hull

Introduction

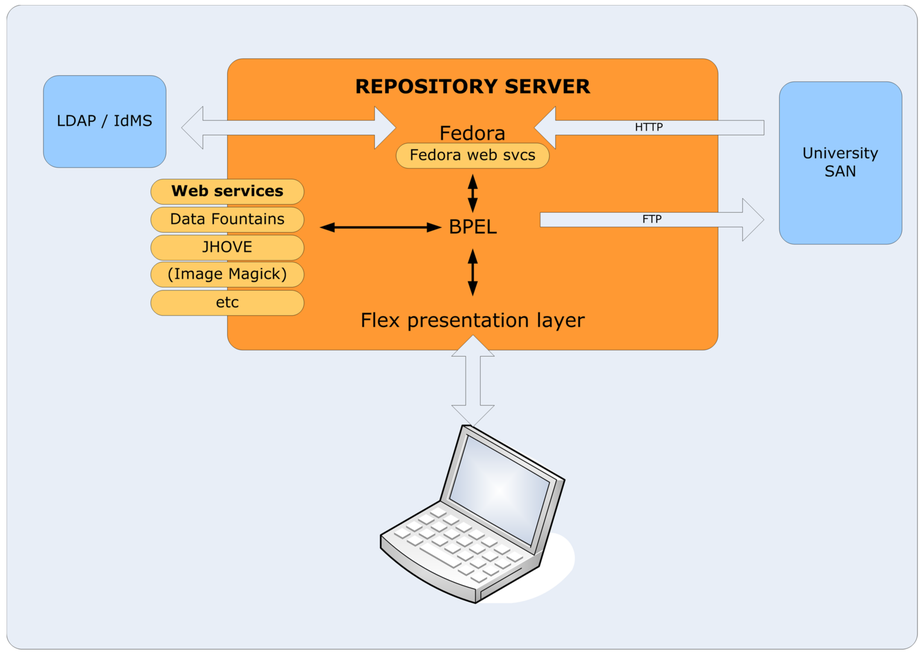
The RepoMMan (Repository Metadata and Management) project (June 2005 – May 2007) has been funded by the UK's Joint Information Systems Committee (JISC) under its Digital Repositories Programme. RepoMMan is closely aligned with the deployment of the Fedora digital repository system within the University of Hull, and has sought to address two areas of functionality that have generated wide interest in the repository arena: workflow and automated metadata generation.

The repository is the latest in a line of infrastructural services that are being assessed and/or implemented within the University to support the educational, research, and administrative needs of staff and students. The majority of these services are based on open source software, where it is recognised that non-commercial alternatives can provide much of the infrastructure an institution requires.

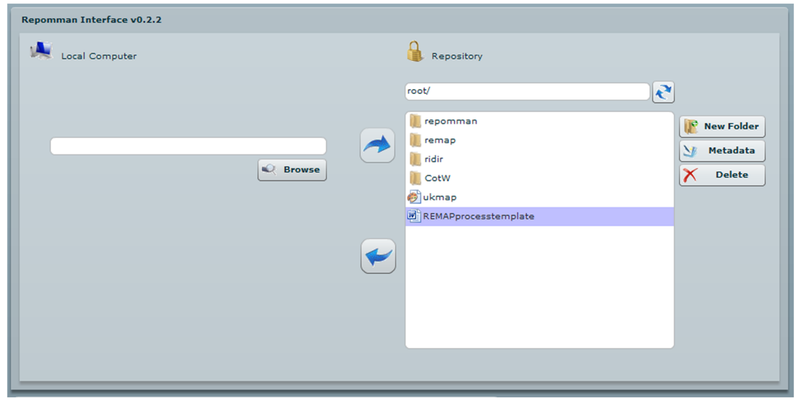
RepoMMan and workflow

Managing a repository underpinning other infrastructural services brings 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. The RepoMMan project is developing an open standards-based, flexible workflow tool through which users can interact with the repository in a configurable way. The tool is based on a combination of technical approaches: it is making use of Fedora's Web Service API interfaces and orchestrating a series of Web Service calls to these interfaces using BPEL (Business Process Execution Language). BPEL allows for multiple individual steps to be combined, allowing users to focus on the tasks that are important to them. Streamlining interaction will facilitate day-to-day interaction and support content creation as well as submission post-creation.

Interaction with the repository will be enabled through the workflow tool, allowing for streamlined ingest and access. The tool will be presented through the infrastructural services it supports: within the portal, the course management system, the virtual research environment, and through direct web access as required. Relevant content can thus be targeted at the systems that require it and made accessible where it is needed.



A representation of the processes used by the RepoMMan tool to manipulate data



Screenshot of version 0.2.2 of the RepoMMan tool
The left-hand-side of the pane represents the user's computer, the right-hand-side the Institutional Repository.
The repository contents are manipulated by the user as though they were simple files rather than digital objects.

RepoMMan and automated metadata generation

One of the steps to be incorporated into the workflow tool is metadata creation. The creation of quality metadata is essential for the proper management and use of an open repository. It is, though, not viable for this metadata to be entirely human-generated in the light of increasing digital content resources. Many different types of metadata can be created automatically. Technical metadata about digital objects can be gathered through tools such as JHOVE, whilst administrative metadata can be gathered through institutional profiles linked to logins. For example, when interacting with the RepoMMan workflow tool through the institutional portal the portal already knows who the user is and can pass this metadata on to be included in digital objects ingested into the repository.

Automated generation of the third major category of metadata, descriptive metadata, is still a holy grail for the most part. Two main approaches have been identified: a document can be analysed via algorithms to extract a useful set of keywords (albeit that the tool may first need to undergo some form of 'training' with related documents); alternatively a document can be analysed against a controlled vocabulary, perhaps subject based, and a list of keywords generated from this. Candidates exist for both approaches: the Data Fountains work at University of California, Riverside, and New Zealand Digital Library's Kea tool, respectively use the analysis and controlled vocabulary techniques. RepoMMan is assessing both approaches. In both cases, however, it is intended that the metadata created will be presented to the user for their own analysis and amendment: this, it is felt, is likely to produce better metadata than asking the user to create it from scratch.

Fedora

The open source Fedora digital repository system is a framework upon which many repository services can be built. Fedora is built around a powerful digital object model, 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. As well as direct access for this management the system has a series of well-defined Web Services interfaces through which all management and access functions can be carried out. This approach, coupled with a detailed security architecture and compliance with the OAIS Reference Model, suited our desire to implement a repository as a key infrastructural service that could be used to support and underpin existing and new infrastructure within the institution.

The Web Service APIs have acted as the basis for the development of the RepoMMan workflow tool, which orchestrates calls to the APIs through the use of BPEL.