

Project Information			
<b>Project Acronym</b>	REMAP		
<b>Project Title</b>	REMAP – Using a digital repository to support the embedding of records management and digital preservation within the institution		
<b>Start Date</b>	01 April 2007	<b>End Date</b>	31 March 2009
<b>Lead Institution</b>	University of Hull		
<b>Project Director</b>	Richard Heseltine (former Director: Ian Dolphin)		
<b>Project Manager &amp; contact details</b>	Richard Green <i>r.green@hull.ac.uk</i>		
<b>Partner Institutions</b>	Spoken Word Services (Glasgow Caledonian University)		
<b>Project Web URL</b>	<a href="http://www.hull.ac.uk/remap">www.hull.ac.uk/remap</a>		
<b>Programme Name (and number)</b>	Repositories and Preservation Programme: Digital Preservation and Records Management		
<b>Programme Manager</b>	Neil Grindley		

Document Name			
<b>Document Title</b>	Final Report		
<b>Author(s) &amp; project role</b>	Richard Green (Project Manager)		
<b>Date</b>	31 March 2009	<b>Filename</b>	Remap-finalreport-v10.pdf
<b>URL</b>	<a href="http://www.hull.ac.uk/remap/downloads/remap-finalreport-v10.pdf">http://www.hull.ac.uk/remap/downloads/remap-finalreport-v10.pdf</a>		
<b>Access</b>	<input type="checkbox"/> Project and JISC internal		<input checked="" type="checkbox"/> General dissemination

Document History		
Version	Date	Comments
v0.1	090210	First draft
v0.2	090220	Second inc SL comments
v0.9	090527	CA comments
V1.0	090603	



THE UNIVERSITY OF HULL

with

**SPOKEN WORD SERVICES**

**REMAP Project**

---

**Final Project Report**

**Richard Green**

**March 2009**



## The REMAP Project

<b>Project Director:</b>	Richard Heseltine, Director of Academic Services (r.g.heseltine@hull.ac.uk)
<b>Project Manager:</b>	Richard Green (r.green@hull.ac.uk)
<b>Repository Domain Specialist:</b>	Chris Awre (c.awre@hull.ac.uk)
<b>Technical lead (Hull):</b>	Robert Sherratt (r.sherratt@hull.ac.uk)
<b>Software developer (Hull):</b>	Simon Lamb (s.lamb@hull.ac.uk)
<b>Archivist:</b>	Judy Burg (j.burg@hull.ac.uk)
<b>Records Manager:</b>	Vicky Mays (v.mays@hull.ac.uk)
<b>Project Director for GCU:</b>	David Donald (d.donald@gcal.ac.uk)
<b>Project Manager for GCU:</b>	Iain Wallace (iain.wallace@gcal.ac.uk)
<b>Technical lead for GCU:</b>	Caroline Noakes (caroline.noakes@gcal.ac.uk)

## Table of Contents

Acknowledgements	5
Executive summary	6
1. Background	7
2. Aims and objectives	8
3. Methodology	8
4. Implementation	9
4.1 Development of the RepoMMan tool	9
4.2 RMDP	10
4.2.1 Triggers	11
4.2.2 Processing the response from the calendar server	12
4.2.3 Responding to events	12
4.2.4 E-mail vs RSS	13
4.2.5 PRONOM and DROID	13
5. Outputs and Results	13
6. Outcomes	15
6.1 Aims and objectives	15
6.2 Stakeholders	15
6.2.1 Institutional managers	16
6.2.2 Archivists or those with responsibility for digital preservation	16
6.2.3 Records managers	16
6.2.4 Repository managers and the repository community	16
6.2.5 End-users	17
6.2.6 The Fedora community	17
7. Conclusions	17
8. Implications	18
9. References	20

## Acknowledgements

The REMAP Project, subtitled “Using a digital repository to support the embedding of records management and digital preservation within the institution” was funded by the JISC Repositories and Preservation Programme under its Digital Preservation and Records Management strand.

The main work of the project was carried out at the University of Hull with assistance and advice from the Spoken Word Services team at Glasgow Caledonian University. The Project manager was Richard Green, an independent IT consultant, working as a subcontractor.

We acknowledge with thanks the help and assistance given us by the digital repositories community world-wide. In particular we would like to acknowledge the contributions made by the Muradora team at Macquarie University in Australia, under the leadership of Chi Nguyen; Ben O’Steen at the Oxford Research Archive; contributors and delegates to meetings of the Sun Preservation and Archiving Special Interest Group (PASIG); contributors and delegates to the Open Repositories 2008 conference; and staff at what is now Fedora Commons. In the latter stages of the project we have been fortunate to work closely with Tom Cramer and Lynn McRae from Stanford University, Tim Sigmon and Ross Wayland from the University of Virginia and Thornton Staples from Fedora Commons on the Hydra Project which promises to take the work started in REMAP to a much wider audience.

We also acknowledge particular help and assistance provided by other members of staff at the University of Hull: Carl Barrow, Richard Garbutt, John Higham, Gary Thompson, and Wayne Thompson in the e-Services Integration Group, and David Jarvis and Alex Sharaz in Computing Services. At Glasgow Caledonian University we were grateful for the assistance and advice of Graeme West during the first year of REMAP.

Last, but certainly not least, we acknowledge the assistance of the JISC Programme Managers who have overseen the REMAP work, Helen Hockx-Yu and Neil Grindley, and the continued support of Neil Jacobs our Programme Manager for the precursor RepoMMan Project.

## Executive Summary

The REMAP Project set out to build on a previous JISC-funded project, RepoMMan. The RepoMMan Project had developed a tool which would allow a user to interact with a private 'My Repository' space in a Fedora repository to develop materials of whatever kind: text, images, databases etc. In addition, the RepoMMan tool investigated ways in which metadata could be automatically generated for the materials if and when the author decided that they should be 'published' to an institutional repository where others might benefit from them. RepoMMan stopped short of developing the 'publish' process itself.

The main aims of this REMAP Project were to develop the publishing process that would allow an author to make materials available through an institutional repository using an extended RepoMMan tool. In addition it was planned that REMAP would investigate embedding into the digital objects thus produced 'triggers' that would help with the ongoing management and possible preservation of the materials over time, a process that we called 'records management and digital preservation' (RMDP).

The first phase of the REMAP Project consisted of user-needs analysis with a number of groups, each offering a potential use case for the toolset that would be developed. From these use cases a smaller, manageable, selection was made to serve as exemplars for the REMAP work.

The technical development that followed first worked towards a usable 'wizard' that an author could interact with to publish material. The wizard, as completed, is adaptive to the type of material that the user is depositing to the repository and RepoMMan's metadata auto-completion work is exploited to pre-populate forms in the wizard where possible. Subsequently the content of the digital objects created in the publish process was extended to contain a set of RMDP triggers and extensive work took place to see how these might best be processed. A decision was made to employ a 'calendar server', a system designed to process timed alerts in a variety of ways. The REMAP team is convinced that, in the long-term, this will turn out to have been the correct decision, however in the short-term – the duration of the project – it caused considerable problems. These problems were compounded by difficulties in obtaining an upgraded version of the software used in RepoMMan and REMAP to sequence the steps required to build a digital object and to process some of the RMDP alerts.

In the closing stages of the project the team integrated work done by The National Archives in the UK which significantly enhances the potential for preservation of digital objects created by the REMAP tool and stored in an institutional repository.

The work done by RepoMMan and REMAP over four years has been recognised at an international level and this has led to the University of Hull being part of an international collaboration: the Hydra project. Hydra will develop an "end-to-end, flexible, extensible, workflow-driven, Fedora application kit." Put another way, it will develop interfaces to the Fedora repository software that will allow institutions to build custom tools like those from RepoMMan and REMAP adapted to local circumstances; it will also provide a customisable search and discovery interface to the repository. In this way, we hope that the ideas developed in the REMAP Project will live on long after the project itself has been forgotten.

## 1. Background

The JISC has, over the past few years, funded a wide range of projects in the records management and digital preservation (RMDP) arena,<sup>i,ii,iii</sup> and established much good practice and knowledge. Many of the projects have worked on the basis of there being an acceptance that interaction with a repository for RMDP is a good thing. The work carried out has certainly demonstrated that this is the case. However, there remains a potentially damaging disconnect between the repository and the people expected to use it on a regular basis.

At the time REMAP was proposed, work in Australia had sought to address this in one way, by developing tools that allowed files to be transformed into XML, currently viewed as the most stable format for preservation purposes, for example the XENA<sup>iv</sup> and ICE<sup>v</sup> projects. DSpace had developed its Lightweight Network Interface (LNI)<sup>vi</sup> based on WebDAV to enable external interaction with a DSpace repository. The JISC-funded PRESERV<sup>vii</sup> project had, though, identified Web services as potential candidates for enabling the development of distributed preservation services,<sup>viii</sup> recognising that not all preservation functionality can be delivered locally within a digital repository. The National Archive's (TNA) PRONOM service is a tool that has its origins in web-based and desktop delivery, but was then being developed further to deliver the PRONOM file format services as Web services. This work built on previous work within the PANIC project at the University of Queensland<sup>ix</sup> on a Web services architecture for preservation services, and the work of the EU PLANETS project<sup>x</sup> (in which TNA is a partner). Web services development to support records management, a specific stage in preservation, had thus far tended to be through the provision of commercial systems, though there was potential applicability of preservation Web services to be focused around specific records management needs

Web services, whilst offering great potential, do not in themselves offer integration between a digital repository and the desktop applications used to create digital materials. This gap can, though, be addressed through notifications, alerting material owners and/or creators to specific records management and/or digital preservation tasks that need to be carried out. The repository can thus enter into an engagement with the end-user in order to encourage RMDP processes over a period of time.

The JISC-funded RepoMMan project<sup>xi</sup> based at the University of Hull had by early 2007 developed a tool to orchestrate Web services using WSBPEL over a Fedora repository.<sup>xii</sup> The REMAP project sought to enhance this tool to work with the PRONOM file format Web services, and others where available, including policy acceptance (e.g., through use of the PLEDGE work at MIT and the San Diego Supercomputer Center (SDSC)<sup>xiii</sup>) and metadata submission and creation. This work was to be based on user requirements for institutional RMDP and the development of related models describing RMDP workflow. The project anticipated that it would develop a notification layer as part of the overall orchestration, using RSS or other messaging to inform users of repository tasks that required their attention and encouraging their engagement with the system. Such notifications would be provided through incorporation of a persistence layer to underpin the orchestration as a whole and address RMDP requirements over time.

## 2. Aims and Objectives

2.1 The REMAP project had the following aims:

- To develop institutional RMDP workflow(s) models in order to understand how a digital repository can support these activities
- To embed digital repository interaction within working practices for RMDP purposes
- To further develop the use of a WSBPEL orchestration tool to work with external Web services, including the PRONOM Web services, to provide appropriate metadata and file information for RMDP
- To develop and test a notification layer that can interact with the orchestration tool and allow RSS syndication to individuals alerting them to RMDP tasks
- To develop and test an intermediate persistence layer to underpin the notification layer and interact with the WSBPEL orchestration tool to allow orchestrated workflows to take place over time
- To test and validate the use of the enhanced WSBPEL tool with institutional staff involved in RMDP activities

2.2 These aims will feed into and help achieve the project's objectives:

- To raise the profile of records management and digital preservation and how it can become a part of regular working practices through interaction with a digital repository
- To better understand how WSBPEL can be used in a real world scenario to support records management and digital preservation
- To test and demonstrate how the Fedora digital repository system can be used to support records management and digital preservation within institutional practices

These aims and objectives did not change in any significant way during the course of the project although the balance of priority between them may have done.

## 3. Methodology

The project used a methodology that has been tried and tested in past successful projects carried out at the University of Hull. This places the user at the centre of technical development, to ensure that the development that does take place is relevant for the purpose at hand.

REMAP's first phase started with an investigation of the user requirements for embedding RMDP into regular day-to-day working practices. The Project Plan identified the institutional Records Manager and its Archivist as two people with a responsibility for addressing these issues at the University of Hull. As the project got under way a number of other people became involved as the project team sought to identify a number of processes as use cases that could be used to ground the work of the project in the real world.

The first set of processes was external to the REMAP Project but based at the University of Hull:

- Preparation of committee papers by Committee Section ◆
- Dealing with past undergraduate examination papers and providing them as learning resources ◆
- Learning and Teaching Programme Approvals





- Maintaining the University's Register of Policies and Procedures
- Maintaining the University Quality Handbook


In addition there were a number of processes in which members of the REMAP Project team were directly involved:

At Spoken Word Services (SWS), Glasgow Caledonian University:

- Development of RMDP for a large repository of audio materials 

At the University of Hull:

- Development of the RepoMMan tool to 'publish' materials 
- Development of repository functionality to support electronic theses and dissertations (ETD) 
- Development of procedures specifically to support digital archives
- Development of an integrated web presence for the newly formed Centre for Spirituality Studies

This offered the project a wide range of opportunities, far more than could be accommodated in its work. In the event, it was the processes marked  that were specifically followed up whilst, in general, the other processes were given a lower priority. (Several are now being addressed outside the scope of the project.)

For the first group of processes, interviews were carried out by members of the project team with staff involved and were documented. A meeting held over two days was devoted to considering the SWS requirements, later summarised in a report. For the final group of processes a brief report on each was produced by the team member(s) involved. In some of these cases, though not all, the outline workflow was represented in diagrammatic form.<sup>xiv</sup> The report went on to synthesise from this work the functions that an RMDP-enabled repository might fulfil.

Thus informed, the technical work on the project began. The project team came to the work with some starting points given, amongst them that its work would be based on open standards wherever possible and that the Business Process Execution Language (BPEL) would be used to orchestrate the Web Services used.

## 4. Implementation

### 4.1 Development of the RepoMMan tool

The RepoMMan project which preceded REMAP had developed a tool which allows users to interact with a Fedora repository in such a way that they can use it as part of the private storage associated with their work; many people see it as a sort of 'digital vault'. However, RepoMMan did not have

the time to add into that tool a process to publish an object to the public-facing repository. (This is not to suggest that the public would necessarily have access to the content, rather that a wider audience with appropriate authentication and authorisation would have access to it.) The first stage of the REMAP project was thus to enhance the RepoMMan tool to carry out this function and, in doing so, produce an RMDP-capable proto-object.

To publish material, a user selects an item in their private repository and clicks the 'publish' button. He or she is asked what the object is (image, text etc) and then for more detail (if text, for example, is it a book chapter, a report, a thesis, ...) The responses are used to set values in the metadata of the proto-object and also to determine which of a number of metadata formats to use. The user is then offered a set of simple metadata forms which are pre-filled as much as possible. (RepoMMan spent some time investigating processes for generating descriptive metadata automatically and deriving personal metadata from the users computing environment.) Once the user has supplemented or altered the metadata offered, a BPEL process is invoked which builds a new object and places it in the publishing queue for the public-facing repository; the user retains their object. A future development of this work will ask the user whether they wish to publish only the latest version of the material or previous versions as well.

The new object encompasses the user's material but is a completely new construction conforming to the appropriate content model(s) of the main repository. It is 'owned' by the repository (not the author) and includes an RMDP datastream appropriate to the content type, a DROID datastream and a PRONOM datastream. The RMDP aspects of the object are dealt with in the next sub-section of this report. At the time of writing, Hull uses the Muradora interface to manage and expose its repository. The publish functionality of the enhanced RepoMMan tool eventually places the new object into the submission queue of Muradora from where it can be moved into the appropriate collection of the repository (see also section 4.2.2).

The enhanced RepoMMan tool, as it stands at the end of the REMAP project, is not capable of batch processing a group of objects but this functionality may be added shortly. However, section 5 ('Outputs and Results') will describe an exciting future for this work which may curtail much further development of the tool as it now stands.

## 4.2 RMDP

The requirements gathering phase of the project has been adequately described in the previous section: Methodology. Here we shall concentrate on describing the work that was done to progress from a set of user requirements towards a toolset that addressed those needs.

The *Records management and preservation requirements* document<sup>xv</sup> identified a range of RMDP alerts that could usefully be associated with repository objects of one sort or another. Whatever the use case, these fell into four identifiable categories: events ('this has just happened' and, in some cases, 'you need to do something'), dates (maybe 'you asked for this to happen xx days ago and it has not yet happened' or 'this document has reached its intended lifespan, what do you wish to do?'), and status reports ('the repository contains xx documents in file format zzz which is now deprecated. Migrate?'). Clearly, the examples given have been abbreviated and simplified. In addition the report identified a number of functions that would be useful to a user responding to these alerts ('take default action' or 'snooze', for instance). The first phase of the technical work

involved deciding how, in principle, how best to store and process the triggers, how to notify the intended recipient of the alerts, and how to process their response.

#### 4.2.1 Triggers

As noted elsewhere, the work envisaged by REMAP assumes the presence of a digital repository using the Fedora software developed at Cornell and the University of Virginia.<sup>xvi</sup> Digital objects in a Fedora repository consist of a number of 'datastreams', some of which refer to content (for example the pdf file represented by the objects) whilst some refer to – or indeed contain - metadata (perhaps Dublin Core or MODS), yet others are used by Fedora for internal functions of the repository. It was always envisaged by the REMAP team that RMDP events and processes should be serviced by adding one or more further datastreams which would record the necessary information to support them. Early on in the project (and its documentation) we referred to the primary RMDP datastream as the 'flags' datastream until it was pointed out that flags in the world of computing are generally a binary concept whilst we were expressing something much richer. It is now called the 'RMDP' datastream.

Fedora datastreams expect their content to be structured. Initial discussions and investigations considered whether the project should write its own pseudo-XML in which to express the alerts or whether (preferably) we could find an open standard which would allow us to express them. In the event, and after many discussions with others in the repository community, the team centred on a small group of calendaring standards, CalDAV, iCAL and xCAL. These are generally used to express information in conjunction with calendaring services such as Microsoft Exchange/Outlook or Google Calendar but, on investigation, turned out to offer the majority of the functionality that we anticipated using.

With the issue of the representation of events essentially solved, discussions and investigations turned to the problem of how to persist them in such a way that they were easily processed. The team was somewhat torn between the straightforward idea of persisting them in, and processing them from, a database or, again, taking a more standards-based approach and using a conventional calendar server. The latter choice won out and, whilst in the long term it will probably prove to have been correct, in the shorter term – the duration of the project – it has given rise to a number of problems that limited what could be achieved in the available time. Initially development proceeded using a Darwin calendar server, developed originally by Apple.

Processing the alerts now, in principle, becomes straightforward. One writes a query using an appropriate language to find out 'what needs to be done today?' say and a response comes back. To communicate programmatically with the Calendar server we wanted to use the Calendaring Extensions to WebDAV popularly known as CalDAV. The appropriate Java API turned out to be CalDAV4j<sup>xvii</sup>, but this was not, indeed is not, nearly so mature and robust as we were led to believe and a great deal of time and effort has been spent trying to solve myriad small problems that its use has thrown up. Late in the project (January 2009) we became involved with the lead developer of CalDAV4j and, at his recommendation, switched from Darwin calendar server to one called Bedework on which his code was routinely tested.

Our move to, and experience of, Bedework – described as an open-source 'Java-based, enterprise-wide calendar system' may yet be of use to others in the JISC and related communities. The Bedework system has recently (April 2009) been adopted by the JASIG incubator, a process that will provide it with support and community engagement toward making it a fully-fledged JASIG project,

akin to uPortal. This will place it on a firmer footing as a tool that can be used within other applications.

The daily triggering of sending CalDAV queries to the Calendar server, was achieved by using the Java based open source scheduler Quartz. This provided 'cron' like functionality, enabling daily or even hourly calls to the Calendar querying services.

#### **4.2.2 Processing the response from the calendar server**

When a response comes back from the calendar server that requires action a BPEL process is initiated. BPEL was used in the RepoMMan project which preceded REMAP to orchestrate web services in order to process a workflow. Since that work the language and standard (BPEL 2.0) have been somewhat enhanced to encompass not only machine processes but also human tasks and processes as well (BPEL4People) – something clearly needed for REMAP. With conspicuous success, RepoMMan used the open-source BPEL engine from a company called Active Endpoints, then in version 4; since then a version 5 has been produced which implements many of the enhancements. Writing BPEL 'by hand' is a complex business and Active Endpoints provided a graphical tool with which to design processes and with which to create the necessary code. Since RepoMMan, Active Endpoints, now Active VOS, has had a change of management and although the open-source version 5 BPEL engine was released it is not supported; version 6 (in no way open-source) has been released and the free designer client has been withdrawn. It took us some time to understand what was going on in this previously very helpful organisation and to discover, at a crucial point in REMAP's development, that they would no longer support us on the previous basis and that, on the surface of it, we were left without a viable BPEL engine. (Other, generally commercial, BPEL engines exist but with a price tag quite outside the range of a JISC project.)

What to do? To cut a long story short we eventually reached an accommodation with Active VOS that they would provide us with a development licence for Active BPEL 5 (albeit superseded) and, as part of that, a designer client. This agreement required us to consult with our JISC Programme Manager in order to re-purpose a substantial sum of money for the licence. As with the issues around CalDAV4j, this was a time consuming process and deprived us of many weeks development time.

Consider a simple example: the final stage of the generic publishing process places the appropriate digital object in the public-facing repository: embedded in its RMDP datastream is a 'published' date and time which is copied to the calendar server. The next time the calendar server is queried for jobs that need doing, this entry, derived from the RMDP datastream, will trigger an e-mail to the original author of the content confirming that the object is now available and providing him or her with the appropriate persistent identifier with which it can be accessed.

#### **4.2.3 Responding to events**

As noted immediately above, the project ran out of development time that would have allowed us to produce a client, or group of clients, designed to help specific groups of users respond to events. However, as we shall see in section 5 ('Outputs and Results'), this work will be taken up elsewhere. Nevertheless, it is worth describing briefly here the sort of functionality envisaged. One of the use cases for REMAP mentioned earlier was that of past undergraduate examination papers. The University of Hull wished to improve its provision of these materials to students to aid the learning and revision processes. The REMAP processes as they stand are capable of publishing these papers

and in doing so an RMDP trigger is built in to alert staff after five years. It is envisaged that after that time a department will normally want to hide the past papers from students as they will be becoming 'dated' and may not accurately reflect current syllabuses or teaching. A customised client would enable a departmental secretary to say "take default action" (causing the repository to automatically change the permissions of the relevant objects so that they are hidden to students) or "hide these but not those"; in other words to deal with the group of objects in an intelligent way. Such an approach is important given that some departments produce thirty or more different papers at a sitting – for preference, a user will not want to deal with each individually. Customised clients, addressing specific needs, had been envisaged for a range of user groups.

#### 4.2.4 E-mail vs RSS

The preceding subsections make repeated reference to sending e-mail notifications; the astute reader may have noted that our Project Plan, parts of which were quoted earlier, refers rather to RSS. As a matter of pragmatism it was easier for REMAP to use e-mail in its initial work. It was always envisaged, and remains a goal, that a user could choose how to receive notifications from the RMDP-enabled repository. Certainly the technologies we have chosen to use can support RSS as well as e-mail, but in choosing to work with the calendaring standards that we did there are other possibilities, for instance the potential to inject items into, say, an Outlook task list because it shares the same underlying technology.

#### 4.2.5 PRONOM and DROID

Thus far we have described work associated with the RMDP datastream which relates largely to events and Web Services that can be managed within the University. However, our view of RMDP goes beyond that and it was always part of the REMAP planning that we should seek to use Web Services provided externally to the University as appropriate, either by calling the remote Web Service or by implementing a local copy of it.

It is clear that the work done by The National Archives (TNA) on format identification is something that a project such as REMAP would wish to embrace rather than re-invent. The process of publishing an RMDP-enabled object creates, not just an RMDP datastream but also one for DROID and one for PRONOM. A local implementation of the DROID service wrapped in a Web service is used to analyse the content associated with each object as it is published; this information is transferred to a DROID datastream. Still within the publishing process, the DROID signature is then used to query TNA's remote PRONOM service to retrieve information about the format identified by the signature. The information returned is stored in a PRONOM datastream from whence it can be processed as and when required. By keeping a local copy of the return we ensure that the University repository is potentially capable of work on format migration even should TNA's services be unavailable.

## 5. Outputs and Results

It is a matter of disappointment to the project team that there was inadequate time to complete the planned work by producing a range of user clients to process the alerts from the repository. Delays caused by issues around the BPEL engine and designer client, on the one hand, and CalDAV4j on the other limited what could be achieved. However, we feel that we have more than adequately

demonstrated 'proof of concept' and we are delighted that the work of RepoMMan and REMAP is to be taken up and built on in a way that was never anticipated.

During the four years spanned by the RepoMMan and REMAP projects, the project team has made a number of presentations at conferences, most notably at two Open Repositories conferences: OR07 in San Antonio, Texas and OR08 in Southampton, UK. It was clear during discussions there that a number of respected colleagues worldwide could see the potential of what we were doing. Following the presentation at OR08 we were approached by the University of Virginia (UVA) and Fedora Commons who asked if they could use our work to jump-start a bigger project. In the event, we met with them in Charlottesville, Virginia, in September 2008 and were joined by representatives of Stanford University. Two further meetings followed that autumn out of which has emerged the 'Hydra Project'.<sup>xviii</sup> (We gratefully acknowledge here financial assistance from the JISC who contributed towards travel expenses for these meetings.)

The Universities of Hull, Virginia and Stanford, with the active cooperation and support of Fedora Commons, have agreed to develop an end-to-end, flexible, extensible, workflow-driven, Fedora application kit. The universities share a view that a repository can potentially be used at all stages in the life-cycle of almost any born-digital material: stages from development, through exposure on the web, to long-term management and preservation. Hydra will provide a 'Lego set' of web services and templates that can be configured and reconfigured to suit a wide range of different workflows that institutions might have or might develop to manage their digital content. The material will be open-source and is likely to be actively promoted by Fedora Commons as an 'out-of-the-box' implementation of the Fedora software.

Hydra will have configurable interfaces to help authors and creators develop their content, perhaps in collaboration with others, expose it on the web, and for the institution's repository managers to manage and possibly preserve it long-term. Hydra will also have a search and discovery interface through which users can explore appropriate areas of the host institution's repository.

Whilst the approach to the search and discovery interface will owe much to UVA's Blacklight search tool and Macquarie's Muradora developments respectively, the aspects of Hydra which precede it (the 'scholars workbench') and follow it (essentially the RMDP element) will owe much to RepoMMan and REMAP. The three partner universities each have subtly different views on how workflow should be orchestrated and choreographed so that three approaches will be developed in parallel; Hull will continue to exploit BPEL, at least for the short and mid-term.

It is already clear that the Hydra project is attracting a lot of interest world-wide. The plan is to have a robust version of the search and discovery interface (including essential management functions) available for production use in the partner institutions in time for the academic year 2009/10. Work to make this adaptable and to add in the 'scholars workbench' and the RMDP element will follow in stages over the following two years. The software (and other aspects of the work) will be made available to the community somewhat behind their deployment by the partners in order to provide an opportunity for testing in a production environment.

In January 2009 it was agreed that Hull's partners in REMAP, Spoken Word Services at Glasgow Caledonian University, should be invited to assist the Hydra project by offering advice on its use with

very large numbers of objects (100,000+) and large data objects (GB+) in addition to acting as a dispassionate test site.

## 6. Outcomes

There is a sense in which members of the REMAP team wish they had been able to complete more of the work they set out to do than proved possible. Delays and frustrations, described above, consumed enough of the project's development time that the move from 'proof-of-concept' work to robust, tested end-user tools was curtailed. That being said, the work that the team did complete has attracted international recognition and, through the Hydra Project, the promise of taking the ideas born in RepoMMan and further developed in REMAP to a much wider audience than had ever been envisaged. There is no small satisfaction in that!

### 6.1 Aims and objectives

The project aims and objectives were set out in section 2 and the aims will not be repeated here. If we accept the pragmatic move from RSS to e-mail outlined at section 4.2.4 then the project aims have been substantially met albeit at proof-of-concept rather than production level.

Turning to the project objectives, the first: *"To raise the profile of records management and digital preservation and how it can become a part of regular working practices through interaction with a digital repository"* has been achieved in a way that we would hardly have dared imagine. The Hydra Project will take the promise of what we have called RMDP to a much wider audience than would have REMAP alone.

The work on REMAP has contributed, and subsequently Hydra will contribute further, to the second objective, *"To better understand how WSBPEL can be used in a real world scenario to support records management and digital preservation"*. Hull will continue to work with BPEL into the medium term although its place in the open source community must now be less secure since Active VOS has declined to further develop or support an open source product. The other approaches to orchestration and choreography being tested in Hydra may, long-term, offer a more sustainable approach.

The final objective was *"To test and demonstrate how the Fedora digital repository system can be used to support records management and digital preservation within institutional practices."* REMAP has gone no small way to achieving that demonstration and the Hydra Project will bring that work to the wider world.

### 6.2 Stakeholders

The REMAP Project Plan identified a range of potential stakeholders and their interests. It is appropriate to consider each in turn and see how REMAP's work may ultimately benefit them.



### 6.2.1 Institutional managers

*Ability to enhance institutional information management and ensure digital materials and records are managed and preserved as required.*

The REMAP team have interacted with a range of middle and senior managers at the University of Hull. It has been clear from the largely enthusiastic response that the potential for an RMDP-enabled repository is understood and welcomed. Indeed, the University's Committee Section, which was involved in the user needs analysis, stopped development of an on-line storage and archiving solution in favour of using the RMDP-enabled repository as soon as it became available. Work to develop customised facilities for them continues.

### 6.2.2 Archivists or those with responsibility for digital preservation

*Ability to position digital preservation in day-to-day activity and ensure a long-term record of the institution's activity*

The Archivist at the University of Hull has contributed much to the underlying design considerations for REMAP. It was she who first pointed out the possibilities of building objects that were empty of data but which contained an RMDP datastream that could be used to guide the process of adding and checking the data; her thinking was from an archival perspective but this idea has already been taken up in a wider context at Stanford. Meetings have taken place to see how the University Archive and the repository are related and how they might best serve each other, and how a repository might eventually be used as the basis of a digital archive. Our Archivist's interest was picked up by colleagues at the West Yorkshire Archives, with whom we had a fruitful exchange of information day in July 2008. Two presentations (one to the Society of Archivists and one to a DPC meeting) are scheduled for May and June 2009, respectively to further the discussion.

### 6.2.3 Records managers

*Ability to position records management in day-to-day activity and ensure accurate recording of relevant documents for management and legal compliance*

As with the archivist, the University Records Manager at Hull has been an active adviser to the project. As a result of her involvement, University Policies and Procedures are now made available through the repository and work is ongoing to better understand the range of RMDP functions that should be built into them. This process has increased the transparency associated with these documents. Our interest and work on using an institutional repository for records management attracted the attention of Steve Bailey, Senior Adviser on Records Management at JISC InfoNet, and a presentation was delivered on this topic at JISC InfoNet's Building Bridges Conference in March 2009.

### 6.2.4 Repository managers and the repository community

*Ability to demonstrate the use of a digital repository for RMDP and embed interaction with the repository in regular working practices*

Throughout the project, the REMAP team has been promoting the idea of RMDP at meetings and conferences. The idea has generally been well received and, in some quarters, embraced – resulting most notably in the Hydra project, but also in such things as the project manager, Richard Green,



being asked to lead a repository summer school at the University of Prince Edward Island, Canada, in 2008. (This event is to be repeated in 2009.)

### 6.2.5 End-users

*Ability to engage with a digital repository to assist in managing digital materials over time through periodic notifications*

The Hull-based REMAP team has maintained an active dialogue throughout the project with partners at Spoken Word Services in Glasgow and with the community on campus. In Hull we have had approaches from a range of different groups with a view to holding material for them in the developing repository. In a significant number of cases, the RMDP element has been a factor in the request. Far from needing to search for repository content, the repository manager has struggled to deal with the volume of material coming forward in the time available to him. Glasgow's needs have been kept under review and development work has proceeded in a way that complex needs such as theirs could potentially be accommodated. It is hoped and anticipated that, in the Hydra project, Spoken Word Services may find an adequately robust and flexible set of tools to meet their needs, easily adaptable when those needs change.

### 6.2.6 The Fedora community

*Ability to demonstrate the usefulness of Fedora within an institutional setting for RMDP as a regular working practice*

All that was said in section 6.4 is equally valid here. Members of the REMAP project team have been active members of the Fedora community during the period of the project and contribute regularly to the discussion lists. Since its inception at the beginning of the RepoMMan project, Chris Awre and Richard Green have co-chaired the Fedora UK & Ireland User Group and have recently been invited to lead Fedora Commons' international 'Scholars workbench' solution community. It is clear that the work of RepoMMan and REMAP is held in considerable regard.

## 7. Conclusions

It is worth repeating here the opening lines in the 'Conclusions' section of the RepoMMan Final Report:

*"In his opening keynote speech to the Open Repositories 2007 conference held in San Antonio, Texas, James Hilton the Vice President and CIO of the University of Virginia noted that "puppies are not free". This was a key message of his talk "Open Source for Open Repositories" and was intended to remind delegates that whilst, like rescue puppies, adoption of open source software may appear to be "free" it comes with a long, and sometimes costly, commitment to maintenance and development.*

*"The experience of the RepoMMan project has been that some of the open source software adopted by the team has required a great deal of work over a period of time to put in place successfully."*

Perhaps we may be forgiven a sense of 'déjà vu'. As in that project, issues around open source software have caused significant delays and frustrations to our work. However, active involvement in the open source communities involved has help get us past the problems and allowed us to move on. We believe that such issues, no matter how frustrating, are frequently an integral part of working with cutting-edge software and evolving standards.

Whatever the difficulties encountered, it is clear in retrospect that the work we have done on RepoMMan and REMAP has struck a major chord in parts of the repositories community. Little did we think four, or even two, years ago that our work would form the basis of an important, international collaboration. It is gratifying to have our ideas, even our ideals, ratified in this way.

Both the eSIG group at the University of Hull and our partners in Spoken Word Services at Glasgow Caledonian University have a history of using an approach to software development that puts potential users at the heart of a development project. Our approach of researching user-needs prior to the event has, yet again, stood us in very good stead. A number of the potential users that we interviewed have been very keen to follow the progress of the project and are looking forward to the functionality that will come from the REMAP approach when it is fully implemented. This has led to Hull's eDocs repository gaining a number of unlooked-for champions and a steady stream of material to be placed in it. Indeed, the amount of material made available has sometimes threatened to overwhelm the limited resources thus far available to deal with it. We anticipate that these enthusiasts will act as willing champions for Hull's repository facilities over time. Clearly, we commend this user-centric approach to other projects.

## 8. Implications

The work envisaged in the Hydra Project has been detailed above; it will not be repeated here.

At the time of writing, the University of Hull in partnership with King's College London has just received confirmation of funding from the JISC to carry out the CLIF Project (Content Lifecycle Integration Framework) which will incorporate aspects of the work done with RepoMMan and REMAP.

The Content Lifecycle Integration Framework project will examine the management of the lifecycle of digital content from creation through to disposal or preservation across system boundaries. It will carry out this examination through the integration of the Fedora digital repository system with two other systems used within the HE sector in the UK and abroad, Microsoft Office SharePoint Services (MOSS) and Sakai, to enable the movement of content between them at specific points in its lifecycle, in accordance with identified use case requirements: a preliminary investigation of integration between MOSS and Sakai will also be undertaken. All three systems are used to manage digital content, and each addresses different overlapping parts of the content lifecycle. Integration will use a loosely coupled, open standards-based approach so that the community could benefit from the work on three levels:

- Adoption of the integration technology developed in equivalent environments
- Adoption of the technical approach utilised to enable a comparable integration between similar systems (e.g., DSpace and MOSS)
- Adoption of the content lifecycle principles and processes developed, through use of relevant e-Framework components developed from the work

CLIF will include a research strand, investigating the content lifecycle and how systems could best support this and a technical development strand, to carry out the integration work informed by the research.

## 9. References

---

- <sup>i</sup> Digital Preservation and Records Management Programme, [http://www.jisc.ac.uk/whatwedo/programmes/programme\\_preservation.aspx](http://www.jisc.ac.uk/whatwedo/programmes/programme_preservation.aspx)
- <sup>ii</sup> Supporting Institutional Records Management Programme, [http://www.jisc.ac.uk/whatwedo/programmes/programme\\_supporting\\_irm.aspx](http://www.jisc.ac.uk/whatwedo/programmes/programme_supporting_irm.aspx)
- <sup>iii</sup> Digital Preservation and Asset Management Strand, [http://www.jisc.ac.uk/whatwedo/programmes/programme\\_preservation/programme\\_404.aspx](http://www.jisc.ac.uk/whatwedo/programmes/programme_preservation/programme_404.aspx)
- <sup>iv</sup> XENA project, <http://xena.sourceforge.net/index.html>. The XENA project at the National Archives of Australia provides a standalone tool that transforms a range of digital materials into XML. It uses the OpenOffice suite as the basis for this transformation, a good example of bringing different services together for a greater benefit.
- <sup>v</sup> ICE project, <http://ice.usq.edu.au/>. The Integrated Content Environment (ICE) project at the University of Southern Queensland takes this a step further and provides fixed templates for use in either Microsoft Word or OpenOffice Writer that can also act as the basis for transformation into XML and ingest into a repository.
- <sup>vi</sup> DSpace Lightweight Network Interface, <http://cwspace.mit.edu/docs/ProjectMgt/Reports/DSpaceUserGroupBergen2006/MIT-LNI-DSUG-Bergen2006.ppt.htm>
- <sup>vii</sup> PRESERV project, <http://preserv.eprints.org/>
- <sup>viii</sup> PRESERV project presentation, 'IRs: towards preservation services', <http://preserv.eprints.org/talks/hitchcock-jiscnew-251006.ppt>
- <sup>ix</sup> PANIC project, <http://www.itee.uq.edu.au/~eresearch/projects/panic/index.html>
- <sup>x</sup> PLANETS project, <http://www.planets-project.eu/>
- <sup>xi</sup> RepoMMan project, <http://www.hull.ac.uk/esig/repomman/>
- <sup>xii</sup> Fedora digital repository system, <http://www.fedora.info/>
- <sup>xiii</sup> PLEDGE project, <http://pledge.mit.edu/>
- <sup>xiv</sup> R Green, C Awre, J Burg, V Mays, I Wallace (2007) *REMAP Project: Records Management and preservation requirements* REMAP Project, University of Hull at: <http://edocs.hull.ac.uk/download.jsp?pid=hull:97>
- <sup>xv</sup> *Op-cit*
- <sup>xvi</sup> Now developed and maintained by Fedora Commons, see: <http://fedora-commons.org>
- <sup>xvii</sup> See: <http://code.google.com/p/calDav4j/>
- <sup>xviii</sup> See: <https://fedora-commons.org/confluence/display/hydra/The+Hydra+Project>