The REMAP Project: steps along the way to a repository-enabled information environment

Richard Green and Chris Awre

[A slightly edited version of this article was published in Ariadne:59 (April 2009): http://www.ariadne.ac.uk/issue59/green-awre/]

Introduction

This article centres on the recently completed REMAP Project undertaken at the University of Hull, which has been a key step toward realising a larger vision of the role a repository can play in enabling and supporting digital content management for an institution. The first step was the Joint Information Systems Committee (JISC)-funded RepoMMan Project that the team undertook between 2005 and 2007 [1]. RepoMMan was described at length in Ariadne 54 (January 2008) [2] and will only be dealt with in summary here. The second step has been the REMAP Project itself; JISC-funded again, this second two-year project further developed the work that RepoMMan had started. The third step, more of a leap maybe, is a three-year venture (2008-11), the Hydra Project, being undertaken in partnership with colleagues at Stanford University, the University of Virginia and Fedora Commons: Hull uses the Fedora repository software, its development is undertaken by the not-for-profit organisation Fedora Commons [3]. Hull will also be working with King's College London on the CLIF project to December 2010, work that will run in parallel with and complement Hydra.

In the Ariadne article describing the work of RepoMMan we wrote:

> "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."

This remains the case but with REMAP we added in notions of records management and digital preservation (RMDP) once the materials were placed in the repository. Thus the repository can play a key part throughout the lifetime of the content. It turns out that others share this vision of repository-enabled management over the full lifecycle of born-digital materials, a concept that some are calling the "scholar's workbench". (Others are calling it the "scholars' workbench": the community has not yet decided quite where the apostrophe belongs!)

RepoMMan: a short review

The RepoMMan Project developed a browser-based interface that allowed a user to interact with a private repository space ('My repository') where they could safely store and manage digital works-in-progress of any kind. The user has the ability to treat the workspace as a digital vault, accessible from anywhere that they have access to the internet, but one that natively supports versioning of the materials that they develop. Thus, at any stage in the process, it is possible to revert to an earlier iteration of their work. It was always the intention that the RepoMMan tool would

eventually allow a user to publish appropriate material from this private space into the University's public-facing repository and to that end the project investigated a range of options for automatically providing various items of metadata with which to describe it. The project stopped short of actually implementing the publishing process.

REMAP: the publishing process

The REMAP Project (2007-2009) took over where RepoMMan left off, but it was not simply an exercise to enhance the RepoMMan tool with a 'publish' function. Whilst working on RepoMMan we had realised that the process of publishing an author's material to the public-facing repository was actually an opportunity to embed within it triggers that would help repository staff to manage the material over time and, potentially, assist in its long-term preservation. Thus the repository has an active role to play throughout the lifetime of the materials.

When an author decides that an item in their private repository space should be submitted for publishing they select it and click on a button labeled 'publish'. This starts a set of processes that will take a copy of their file, whatever it may be, lead them through the process of adding metadata to describe the material and will finally create a new digital object which goes into the accessioning queue for the repository proper to await approval.

From the author's standpoint, clicking the 'publish' button starts what we might call a 'publish wizard'. The purpose of the wizard is largely to take the author through the process of gathering metadata with which to describe their work. As things stand at present with Hull's repository this will generally become a Dublin Core (DC) metadata record within the repository's digital object (a move to the richer MODS metadata schema is likely soon) with the exception that our theses use UKETD_DC [4].

It will be helpful if we take a concrete example, so let us assume that one of us wishes to place a version of this article in Hull's repository and that it exists in a private repository as a Word (.doc) file. The author is first asked to describe the material that they are publishing by choosing from a dropdown menu. Our article is clearly 'text' and from the subcategories available we would choose 'article'. The next wizard screen asks us for the author, the subject of the article and an abstract, however it does not appear with three blank text boxes; we might reasonably hope that all three are filled in through automated means. When the author clicked 'publish' a set of processes (technically, Web Services) was invoked which retrieved a copy of the Word file and, in the background, ran it through a metadata extraction tool [5]. Amongst other things the tool attempts to identify these three items (author, subject and abstract) from the file; in our experience it is quite effective. Thus the author sees prefilled text boxes which he or she can alter if they feel that the content is wrong. The process proceeds with pages pre-populated as much as possible.

When the author finishes with the wizard a further set of processes is invoked to take a copy of the author's file and the metadata that they have provided and to build from it a new digital object which will go into the repository accession queue. This object will conform to a standard 'content model' for articles. In other words, as an aid to long-term management, it will have exactly the same internal structure as all other digital objects holding articles in the repository. So what will this structure be?

An object in a Fedora repository contains a number of so-called 'datastreams'. Some of these are required by the repository, others are defined by the user. Thus our object here must have datastreams that deal with the minimal Dublin Core metadata required by Fedora and the object's relationships to other objects in the repository (the formal expression of the content model is itself a repository object). In addition, all Hull's simple text objects will have a delivery datastream for the text and for any archive version of it:

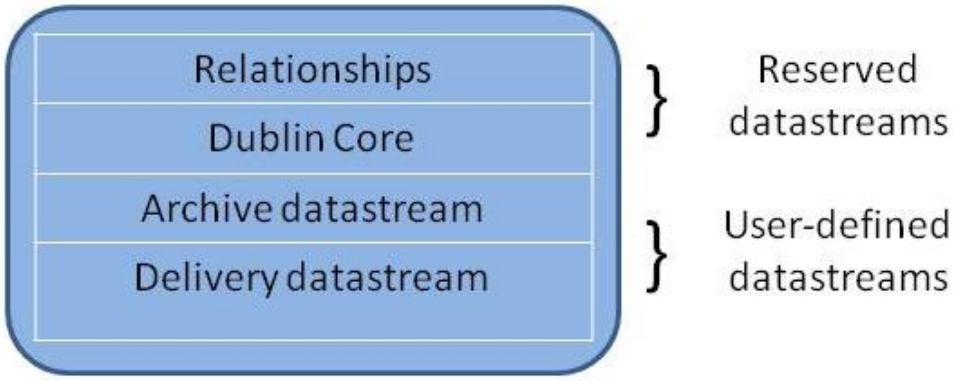

Figure 1: A simplified structural view of a text object in Hull's repository

Hull's preferred format for text in the repository is pdf; thus a background process will be run that converts the Word file. The original goes into the object as the archive datastream whilst the pdf becomes the delivery datastream and the metadata enhances the DC datastream. Essentially this is all that is needed for the object to become part of the institutional repository's content. The object can now go for checking and approval and should appear in due course. The 'creation to publish' cycle envisaged at the start of the RepoMMan Project has been completed – but REMAP goes further.

Before moving on to describe the RMDP aspects of REMAP there are a few further comments that should be made about the publishing wizard and the process more generally.

We have tried to adopt an intelligent approach to the gathering of metadata from the author and the subsequent creation of digital objects. Hull's use of Web Services to underpin the tools means that we can run quite complex, non-linear processes behind the scenes. The metadata wizard is sensitive to the type of content being submitted; had we declared our text to be a 'thesis or dissertation', for instance, we would have seen a quite different sequence of screens because these use a different metadata schema (UKETD_DC). Had we submitted an image of some sort, whilst DC would again have been used, additional requests for information would have been made notably for filesize and image dimensions but these would have been pre-filled by calling another Web Service (in this case one invoking a locally installed copy of the JHOVE tool from Harvard [6]). The process of creating an image object would have resulted in a number of datastreams containing 'derived' images in a range of sizes from 'full-size' to 'thumbnail' and the transfer of the original file to an archive datastream (it may have been that this original was not in a browser-compatible format, perhaps a TIFF file). All of this goes on with the author largely unaware that anything terribly complex is occurring.

REMAP: the RMDP work

At the beginning of the previous section we talked about embedding 'triggers' in the digital objects that the REMAP technology creates. This formed a significant part of the REMAP work, taking the team into new territory.

As we noted in our introduction, the vision for an institutional repository at Hull was one that saw the repository at the heart of the University's information architecture; the storehouse for much of its digital material of whatever type. Thus it was that we could envisage in the future a repository holding hundreds of thousands, even millions, of objects of widely varying file type. Such a vision could easily become a nightmare for repository managers. How would we manage so many disparate objects effectively? The idea behind the REMAP work was to investigate how, in a sense, the repository could contribute to its own management.

The work started with a user needs analysis during which we interviewed a variety of key University staff in order to understand the lifecycle of the materials that they dealt in and how that might translate into repository terms. In addition we spent time with our partners in the REMAP Project, the Spoken Word Services team at Glasgow Caledonian University [7] who were investigating the potential for a repository of audio materials which envisaged a particularly complex lifecycle. Thus informed, we identified in our user needs report [8] a range of triggers that might usefully be built into a repository object.

It may help to explain the process through a concrete example. Consider a Departmental Secretary who each year submits to the repository a departmental prospectus for potential students. The students might retrieve it directly from the repository or via a link from the departmental website. The prospectus needs to go through the normal repository checks before being exposed and, once available, has a lifetime of just a year before it needs to be replaced. Triggers might usefully be placed in the object to drive two processes. In the first, a trigger is placed in the object to say that on or immediately after the date of publication the Secretary should be e-mailed that publication has successfully taken place and giving the URL of the prospectus. In a year's time, the repository should e-mail the Secretary that the prospectus is now out of date and will be hidden from public view unless action is taken to prevent this. In this last example we see the repository becoming proactive in its own management. We have used e-mail in the examples but the process could equally well use RSS feeds or some other information tool.

How is this functionality achieved? When the new object is constructed for the repository from the author's original, an additional datastream is created within it. This holds information about the triggers in a format commonly used with personal calendars and scheduling software. From there the information is copied to a 'calendar server' which is used to actually deal with it. In fact, the information goes onto the calendar server as 'to-do' jobs and the system is periodically polled for outstanding jobs to be dealt with. Once each task is complete that fact is recorded against the trigger entry in the object so that an audit trail is built up.

Consider a second example involving information that needs to be preserved in the long term. The minutes of key University Committees need to be stored 'for ever'. When these go into the repository they would be processed, again via Web Services, through a locally installed copy of the DROID tool from The National Archives (TNA) in the UK [9]. This analyses the file containing the minutes and attempts to identify the format exactly, producing a DROID file signature. The

information from the DROID tool is stored in the digital object. The unique file signature is then transmitted to TNA's PRONOM service which returns information about the format and any risks associated with it. Whilst we would not expect it at the moment, in the future the PRONOM response might recommend a format migration or other preservation work to ensure that the content can be successfully accessed for a further period. Whatever the PRONOM response, it is stored. Driven by embedded triggers in our object, the PRONOM service will be re-queried periodically so that any changes can be noted and necessary actions taken.

Seen in overview, the complete RepoMMan and REMAP structure looks something like this:

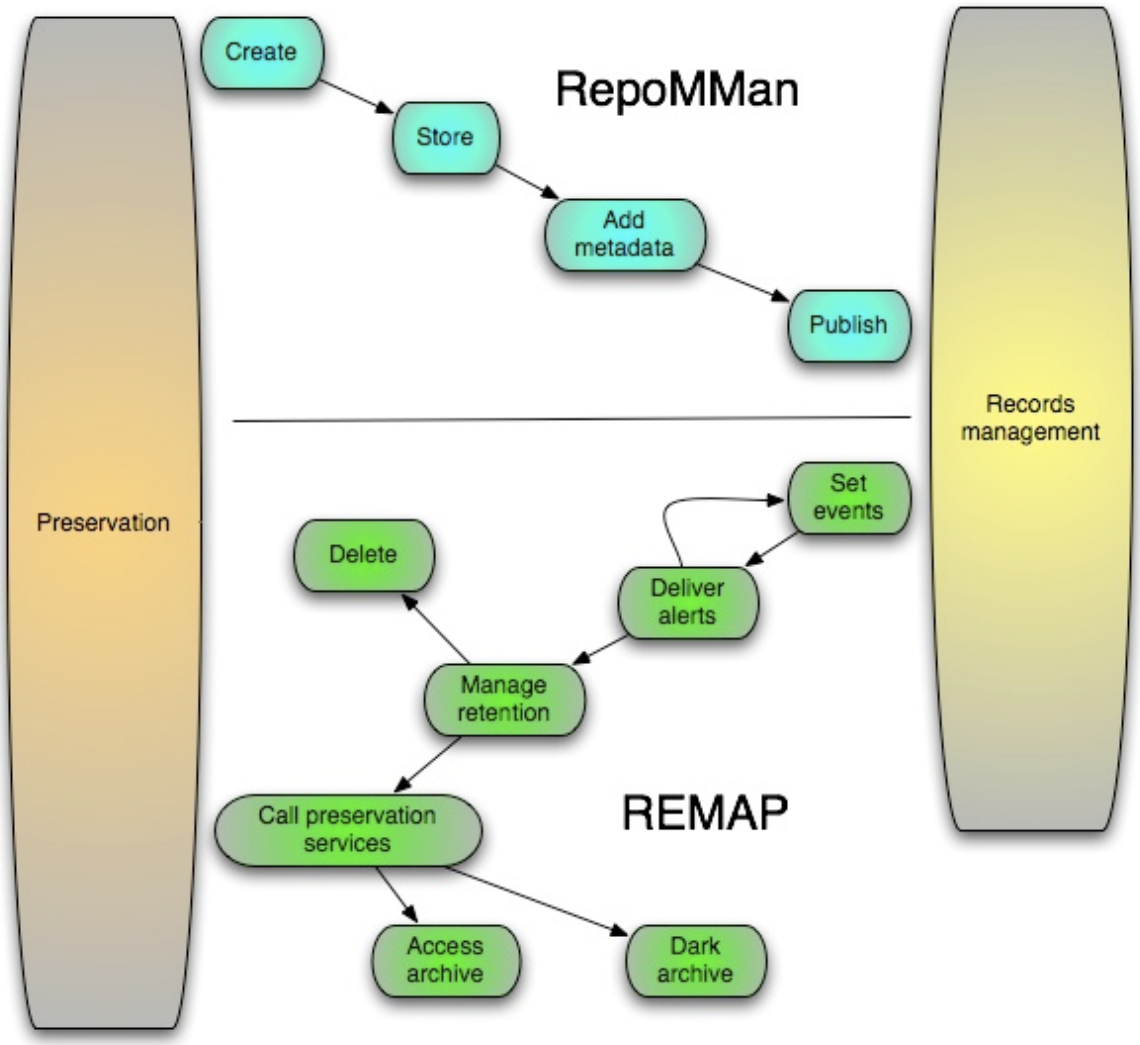

Figure 2: An overview of the RepoMMan and REMAP processes

The work undertaken with triggers has taken the REMAP team into new territory and, as so often working with relatively new standards and emerging software, the route has not been an easy one. The result of this is that, at the end of the project, our RMDP work is not as well-developed as we should have liked. That said, all the processes described above have been demonstrated in practice and will be taken forward in due course. Enter the Hydra Project.

Hydra

The work of RepoMMan and subsequently REMAP was presented at a number of conferences in the UK and the US. Following a presentation at Open Repositories 2008, the REMAP team was approached on behalf of the University of Virginia (UVa); they had a need for a specifically targeted repository but wanted to build it incorporating ideas from our two projects. It was agreed that we should hold a meeting in the early autumn and thus it was that September 2008 saw us at UVa talking with staff there. Also represented were Stanford University, who had identified a similar need, and Fedora Commons who were interested in the potential of the meeting for the development of the Fedora repository software.

At this meeting the three universities agreed, with the active cooperation of Fedora Commons, to work together "to develop an end-to-end, flexible, extensible, workflow-driven, Fedora application kit" [10]. What this means in practice is a search and discovery tool for the Fedora repository software integrated with RepoMMan-like functionality to support authors and creators and REMAP-like functionality to support records management and preservation activities. However, this work will be done in such a way that what emerges is a toolkit from which other potential users can construct workflows around their needs using a 'Lego set' of (re-)configurable components. This work is well under way and the partner universities hope to have the search and discovery interface, the core of the system, with some useful workflows, in place for the start of the academic year 2009/10. After that the work will be broadened to enable other workflows and the RMDP work. The software will be open-source and will be released in stages to the repository community after appropriate production testing.

CLIF

The work of the projects thus far, and their continuation through Hydra, is placing the repository at the heart of the information environment within an institution. Making this environment flexible and user-friendly to use is vital, as Hydra as identified. We have also recognised that the repository within an institution is but one place where digital content management takes place, and that such management activities have taken place for some years without repositories. The CLIF (Content Lifecycle Integration Framework) Project [11], a collaboration between Hull and King's College London, will undertake work complementary to Hydra to identify how the lifecycle of digital content influences its management across systems within an institution, and the related integrations between the repository and other systems that are required.

Conclusion

Thus it is that we anticipate the work of RepoMMan and REMAP, and the original vision, being taken forward - albeit in a rather unexpected way. Little did we think, four years ago when we started out on this work, that the future held such exciting, international opportunities.

Acknowledgements

Many people, from the UK, continental Europe, Australia and the US, have contributed directly and indirectly to our work over the last four years. It would be impossible to name them all but they have our thanks. Two groups in particular, though, should be singled out:

We acknowledge with gratitude the contributions of the Joint Information Systems Committee (the JISC) in funding the RepoMMan and REMAP Projects, and in contributing towards travel costs associated with getting the Hydra Project under way.

We acknowledge too the significant contributions made to REMAP and the thinking around it by our good friends in Spoken Word Services at Glasgow Caledonian University (David Donald, Caroline Noakes, Iain Wallace and, for the first year, Graeme West). We are delighted that they have agreed to become consultants to Hydra.

References

1. The RepoMMan Project http://www.hull.ac.uk/esig/repomman/
2. Green, Richard and Awre Chris RepoMMan: *Delivering Private Repository Space for Day-to-day Use* Ariadne 54 http://www.ariadne.ac.uk/issue54/green-awre/
3. Fedora Commons http://fedora-commons.org
4. UKETD-DC http://ethostoolkit.cranfield.ac.uk/tiki-index.php?page_ref_id=47
5. The iVia metadata tool, part of the Data Fountains Project http://datafountains.ucr.edu/
6. JHOVE http://hul.harvard.edu/jhove/
7. Spoken Word Services http://www.spokenword.ac.uk/
8. Green R, Awre C, Burg J, Mays V, and Wallace Ian (2007) *REMAP Project: Records Management and Preservation Requirements* http://edocs.hull.ac.uk/splash.jsp?parentId=hull:798%26pid=hull:97
9. PRONOM and DROID http://www.nationalarchives.gov.uk/PRONOM/Default.aspx
10. The Hydra Project https://fedora-commons.org/confluence/display/hydra/The+Hydra+Project
11. The CLIF Project: http://www.hull.ac.uk/clif

Author details

Richard Green
Manager
CLIF and Hydra (Hull) Projects
c/o Academic Services
Brynmor Jones Library
University of Hull
Hull
HU6 7RX

Email: r.green@hull.ac.uk
Web site: http://www.hull.ac.uk/clif
https://fedora-commons.org/confluence/display/hydra/The+Hydra+Project

Richard Green is an independent consultant working with the IT Systems Group in Academic Services.

Chris Awre
Head of Information Management
Academic Services
Brynmor Jones Library
University of Hull
Hull
HU6 7RX

Email: c.awre@hull.ac.uk
Web site: http://www.hull.ac.uk/clif
https://fedora-commons.org/confluence/display/hydra/The+Hydra+Project