RepoMMan Project

R-D14

RepoMMan User Needs Analysis

Richard Green and Chris Awre

March 2007

Version 1.0 07-03-21

The RepoMMan Project

Project Director:	Ian Dolphin, Head of e-Strategy, University of Hull	(i.dolphin@hull.ac.uk)
Project Manager:	Richard Green	(r.green@hull.ac.uk)
Technical Lead:	Robert Sherratt	(r.sherratt@hull.ac.uk)
Repository Domain Specialist:	Chris Awre	(c.awre@hull.ac.uk)

The Repository Metadata and Management Project (RepoMMan) at the University of Hull is funded by the JISC Digital Repositories Programme. The project is being carried out by the University's e-Services Integration Group (e-SIG) within Academic Services.

Introduction

The RepoMMan Project was established to provide a user interface to a Fedora digital repository which would allow a user to interact with it as an integral part of his or her workflow. In order to understand the requirements of such interactions, the project undertook survey and interview work with three groups of people: those involved in academic research, those involved in teaching and learning (T&L), and those involved in administration. In practice it is accepted that some potential users fit into more than one of these roles.

A number of staff at the University of Hull from each of these three groups were interviewed. These interviews typically took 60-80 minutes each and were later transcribed verbatim. Each of the interview transcripts has been reduced to a 'scenario' which encapsulates the user needs described. It is not necessarily the case that there is a 1:1 correspondence between an interview and a scenario; some material has been merged. In addition to the interviews, the RepoMMan team conducted an on-line survey of researchers at Hull and elsewhere and assisted the CD-LOR (Community Dimensions of Learning Object Repositories) Project with a similar survey of those involved in T&L.

The University of Hull takes a very broad view of the way in which a digital repository might be used. In addition to being a showcase for 'finished' digital material, we see potential for the repository to support a user in the production of that material. The investigations thus sought to tease out the essentials of each person's daily work and working practices, as related to potential digital objects, the type of digital objects that they create, or might create in the future, and thereby to reach an understanding of how a digital repository might help them in the process.

Research

Interviews

The project has produced a report on the 'research user requirements interview data'.[1] It concludes by reducing the research interviews to five summary scenarios:

Scenario 1

Steven uses a range of specialised search tools to inform the development of his research papers. He organises the materials that he finds in a highly structured manner using sophisticated cross-referencing and he uses advanced indexing software to help him find references on his computer's hard drive. When his draft is far enough advanced he consults colleagues at Hull and further away who comment using pen or Word 'track changes'. When the article is finished he submits it to a publisher electronically. During the development of the paper Steven starts a new copy of the file each time there is a significant structural alteration; he takes periodic backup copies on CD. At the end of a research project he retains all his research materials. Steven generally signs any copyright agreement with his publisher without looking at its detail.

Scenario 2

Tony uses a small number of search tools to inform the development of his research papers. He organises the materials that he finds in a simple structure. As he develops a paper Tony uses version numbers which advance each time he makes more than a

[1] Green R (2005) *R-D4 Report on research user requirements interview data* University of Hull at http:// www.hull.ac.uk/esig/repomman/downloads/R-D4-rsch-int-data-11.pdf

trivial change; he normally has at least one backup and usually two. He does not generally share his work until it is at an advanced stage at which point he presents his thoughts at a conference to obtain feedback. When his paper has been published Tony retains all his research materials. Tony generally reads and signs any copyright agreement with his publisher but has not considered the implications of this in the context of making his work available by other means, such as a personal website or an institutional repository.

Scenario 3

Charles uses general and subject-specific search tools to inform the development of his research papers. He keeps the materials that he finds in an ordered manner on his computer but frequently prints them in order to work with and annotate the text. Once the paper is well advanced he may share it with colleagues elsewhere for comments which are generally made using the Word 'track changes' facility or else by telephone. As he develops a paper he uses version numbers which advance each time he makes more than a trivial change; there is always at least one backup of the current version. At the end of a research project he may keep his research materials on disk for a time but is more likely to print them out and file them. Charles carefully reads any copyright agreement with his publisher and may challenge its provisions if they do not suit his purpose.

Scenario 4

Darren uses a small number of search tools to inform the development of his research papers but increasingly tends towards Google Scholar as his tool of first choice. He keeps the materials that he finds in an ordered way. As he develops a paper he shares it with colleagues elsewhere for comment which is generally done using the Word 'track changes' facility. Every time he alters the developing paper he gives it a new filename which includes the date; there are always multiple copies for backup. At the end of a project he makes his paper available via the Departmental website and would like to be able to provide accompanying data. He retains his research materials on disk in a structured way but also keeps printed copies. Darren reads any copyright agreement with his publisher and signs it but believes that there is generally a private understanding that he will also post a version of his paper on the Departmental website.

Scenario 5

The last research scenario is similar to the third although, being about medical research, there is an important addition about record keeping.

Peter, a medical researcher, uses general and subject-specific search tools to inform the development of his research papers. He keeps the materials that he finds in an ordered manner on his computer but frequently prints them in order to work with and annotate the text. Peter develops a lot of his work using empirical research data about patients and is subject to strict rules about the way this data is handled, stored and retained. Once the paper is well advanced he may e-mail it to colleagues elsewhere for comments which are generally made using the Word 'track changes' facility or else by telephone. As he develops a paper he uses version numbers which advance each time he makes more than a trivial change; there is always at least one backup of the current version. At the end of a research project he may keep pdfs that he has downloaded as background on disk for a time but is more likely to print them out and file them. Records and data relating to his own research are carefully preserved in their original paper form. Peter usually reads any copyright agreement with his publisher and has a general idea of its provisions.

On-line survey

The on-line survey of researchers[2] did nothing to contradict the findings from the interviews, rather it confirmed that many of the views expressed in them were common at universities elsewhere. The survey provided considerable information on the types of digital object produced by researchers and on their methods of managing these from conception through to completion. If the RepoMMan Project is to provide a tool to meet needs in this area, the repository must be capable of storing, and the tool capable of handling, a wide variety of formats. We shall return to this theme later.

Teaching and Learning

Interviews

As with the researchers, the RepoMMan team conducted a number of interviews with people whose interests centre on T&L.[3] The verbatim transcripts of these interviews were reduced to scenarios. Although these scenarios are drawn from interviews with members of the teaching and learning community, each of those staff interviewed had administrative responsibilities of one sort or another. Accordingly these scenarios are not wholly about T&L issues.

Scenario 6

Peter is interested in making available learning objects consisting of between five and nine 'information objects', as he calls them, which each represent about five minutes of T&L activity. The learning objects would be searchable for their information components, and these information components would be capable of extraction and reconstruction into new learning objects. An information object might consist of a brief introduction, some facts that are being taught, and an assessment. It may also contain media of some sort, perhaps an image, a Flash object, or a short video clip. Peter also sees advantages of having aggregations of learning objects stored in his private user area - each perhaps corresponding to a lecture of some sort. In saying this he is thinking beyond lectures to full-time students and thinking about his need to be able to mount a lecture or short course for an external group at short notice.

Stephen, who works with Peter, likes these ideas and sees much potential advantage in being able to identify components of the learning objects, being able to extract them, and then being able to recombine them, perhaps with additional material. He sees benefits not only in re-using Peter's 'information objects' but in being able to extract components from them - perhaps the image or video clip. He also sees benefits in students having access to all these components so that they can directly use part of the learning material, perhaps in the assessment work that they do in response to it.

Stephen would like to have all the components of his teaching materials available to him on-line in a repository. This would enable him to locate a specific item in his private repository space very quickly, to the extent that it might be feasible to locate it in response to an unexpected question from a student in a lecture. He would greatly value the ability to very quickly find and display a short video-clip, say, in this manner.

Stephen can also see considerable value in sharing his materials with others and in users being able to annotate the sub-components. Such comments might say that "this image was really useful in trying to explain X to my students today": in other words

[2] Green R (2005) *R-D3 Report on research user requirements on-line survey* University of Hull at http://www.hull.ac.uk/esig/repomman/downloads/R-D3-research_survey_data_11.pdf

[3] Green R, Awre C (2007) *R-D12 Report on admin and teaching & learning user requirements interview data* University of Hull at http://www.hull.ac.uk/esig/repomman/downloads/R-D12-TLA-user-needs.pdf

offering to others the context in which it was found useful. In sharing objects, Stephen would like the description available to others to be very broad so that an Economist (say) might be able to locate, recognise and use a Flash animation that he had produced for his subject, which is actually Geography. He could also see much benefit in being able to maintain a 'wish list' (à la Amazon) for components that he would like but has not yet found. The list would clearly need to be visible to others.

Scenario 7

Keith works with students who produce a lot of media material: images, video and some sound. At present the finished items reside on a server in his department, which is to say that they are not available for wider use. He regards this as a shame because he and his team are careful about copyright and other permissions so that in most cases the images and clips could be more widely used within the University of Hull community if they could be made accessible; a smaller number of these materials could be made available to the public. However, he accepts that this would not be a trivial task because whilst many of the materials have some metadata associated with them this is in a proprietary database at present and it is not clear whether this could be transferred to a repository in an automated way. The materials vary in file size from digital video (dv) format files measured in gigabytes to podcasts and some images which may be only a few megabytes.

Keith can see the future benefit in using a repository such as the University proposes. Students could place their finished materials in the repository and provide appropriate metatagging. He is very interested in the ideas of being able to provide metadata for video that references scenes in the clip by timecode and of being able to annotate images or video. He is not quite so interested in having students use the repository as a development space because he encourages the use of 'social' storage on the web where the drafts can be widely shared with others for their comment, that said he can see potential uses for the inherent versioning capability of the repository, particularly because he and his colleagues encourage their students to reflect on the various stages of the development process.

Backup is a serious concern to Keith at the moment and the repository could potentially offer him a solution to this problem, however already Keith's materials consume almost two terabytes of storage and this could be an issue.

Scenario 8

Timothy works with students to produce simulations of industrial processes. This involves data modelling leading to simulations, programming, and the construction of bespoke computers to service particular needs.

He and his students have a number of relatively unusual storage requirements. Departmental projects tend to build on what went before, rather than to be parallel developments. This means that in preserving previous material which may be extended in later years it is necessary to preserve all aspects of the hardware and software that have gone into a development: programs, data, component specifications, printed circuit board design, overall design and a complete image of the system's disk(s). In terms of file size for storage, some of these components are relatively trivial, however disk images can amount to tens of gigabytes and the data sets used to develop a simulation can be similarly large. The component parts of the project materials would need to be reliably collected together and simple to retrieve.

Timothy can see use for the repository during the development of, most especially, simulations. The development of the data and the model which provides the simulation can be a long process measured in months. It is not unknown to make a fundamental

misjudgement in the development which will eventually mean having to revert to a version from many weeks earlier. Automatic versioning in a repository would be useful here. He can also see use for the repository as a backup for development work in progress. The University does not normally provide him with centralised backup although it does provide some centrally maintained systems - which is to say that it manages the basic computer disk image and desktop. Experience has led him to conclude that he should have his own backups of the computers that his students use to develop project work in order to guarantee quick restoration in the case of problems. In this regard he is thinking about putting complete disk images in the repository; these would necessarily be many gigabytes in size, even when compressed.

In his administration role within the department, Timothy can see merit in using a repository to store students' completed work, including perhaps electronically marked materials.

Looking across the department as a whole, Timothy recognises that adoption of technology to support learning & teaching is not wholesale. The use of a repository to support his work will benefit from the involvement of other members of staff and some level of advocacy work will be required to facilitate this.

Scenario 9

Pradesh manages a resource centre for academics and their students. The centre offers printed materials, some of which it develops itself, a question bank, images, video and software. Currently this material is made available to users through a number of websites and Pradesh can see benefits in having them available from one. He sees a repository, such as the University envisages, as an appropriate place for these materials for two reasons: the repository might be of use whilst developing some of the, especially printed, materials; and some of the materials produced have long-term value and should continue to be available even should his Centre close.

The Centre produces a number of publications each year. At the moment, these tend to be drafted by one of Pradesh's team but then checked and commented on by a number of others. A private repository area that allowed collaborative working could be useful in this process.

Other materials that the Centre provides are not so much developed by them as by others. However, Pradesh sees the repository potentially as a useful showcase where all the materials can be brought together. He accepts that these materials would need to have good metadata to aid search and discovery. Many of them do have such metadata at the moment, but it is not directly 'attached' to the items, rather it is in an associated database. The question bank poses a slightly more complex problem in that access to it, should it be done through a repository, would have to be subject to flexible yet absolute security.

Pradesh understands the copyright issues that would be involved in providing his Centre's materials through a repository and is in a position to address them.

On-line survey

The RepoMMan team assisted members of the CD-LOR Project in conducting an on-line survey of the T&L community. This is reported separately.[4] The aim of the survey was "to help determine how individuals find, create, store and share their educational resources and how they collaborate on the development of these resources with colleagues [elsewhere]." The

[4] Margaryan A (2006) *Report on Personal Resource Management Strategies* CD-LOR Project, Glasgow Caledonian University

term ‘educational resources’ covered almost anything that could be used to support T&L. Whilst it is not appropriate to quote fully the conclusions of that report here, some paragraphs are of direct relevance:

> *“Sharing and storing work-in-progress*
>
> “A very high level of sharing work-in-progress for comment and collaboration was identified in our sample, which confirms the view that repositories could play a useful role in supporting such collaboration, although they don’t appear to do so at present. In terms of subject areas, Education and Arts appear to be most active in terms of sharing.
>
> “An important element of storage is backing up work-in-progress. The vast majority of the respondents use one or more strategies/devices for backing-up work-in-progress, the most popular of which are pen drives and university network.
>
> “Version control is another important aspect of storing work-in-progress. The majority of the respondents in this study utilise one or more methods of version control. The most popular methods are indicating version and date in filename or document.
>
> *“Sharing, delivering and storing completed work*
>
> “Completed work is even more widely shared than work-in-progress. This could be interpreted as that people tend to be less willing to “go public” with work-in-progress and more confident in sharing completed work. It may also be explained by the fact that repositories of teaching and learning resources are not yet set up to handle or support sharing of works in progress, but are seen as storage areas for completed works.
>
> “The completed work is predominantly made available via departmental, institutional or personal websites. This is not surprising given that most of sharing tends to be with departmental or institutional colleagues. This could imply that there could be a larger scope for institutional LORs [Learning Object Repositories] than other types of LORs. A relatively large number of respondents indicated that they utilise institutional, national or subject-specific repositories to share completed work.
>
> “In terms of delivering completed work to students, institutional VLE [Virtual Learning Environment] is the most popular medium. Displaying resources electronically in classroom, as well as distributing resources in paper-based format are also popular ways of delivery. When making the educational resources available electronically, they tend to be both uploaded to the delivery mechanism (e.g. VLE) and linked to the external location. Repositories do not appear to be used for this task. However, given these findings perhaps repositories should be linked with VLEs.
>
> *“Ownership of educational resources*
>
> “A widely held perception among the respondents is that copyright is owned by institutions rather than individuals, although one fifth of the respondents indicated that they owned the copyright. This could imply that many would perceive that they wouldn’t be free to distribute the resources through channels outside institutions, for example via national or subject-specific repositories.
>
> “Notable is the number of respondents who indicated that they didn’t know who owned the copyright for the resources they developed. This is indicative of the current lack of clarity and in many cases the lack of explicit policies with respect to ownership of resources, both within the UK and internationally. This lack of clarity is a barrier for wider uptake of repositories in teaching and learning.

"Reusing and repurposing resources

"Overall, people mostly tend to collect materials developed by others when developing their own educational resources, but they often do not repurpose or reuse these materials. These are mostly text-based resources, images, diagrams and URLs... Also, the majority tend to archive on computer all or some of these educational materials once the particular resource for which they were gathered has been delivered.

"Resources are often created completely from scratch. In a few cases they are based on some existing materials. In vast majority of cases they are repurposed from own *(sic)* materials. In contrast, repurposing of others' materials is currently low. In addition, current levels of reuse of resources created by others are very low.

"Finding and selecting materials to reuse

"Finally, current practices related to finding and selecting materials to reuse were investigated. In searching for materials to use when developing educational resources, most popular strategies are search by subject keyword, by type and file format. Principal criteria by which such materials are selected include recommendation by a trusted colleague and trustworthiness and reputability of the source where the materials are found. Thus trust appears to be a major factor when selecting materials for potential repurposing or reuse."

Note that these paragraphs are selected from the conclusion of the CD-LOR report; the reader should refer to the original document for the full conclusions, properly in context.

These points do not contradict any of the findings from the RepoMMan interviews; quite the opposite, they confirm many of them. In addition, there is interesting reference to the use of these materials from within a VLE, a process that RepoMMan will be investigating as part of its own work.

Administration

The third group interviewed by the RepoMMan team were primarily administrators. The same approach, that of reducing the verbatim transcripts to scenarios, was taken. As with the T&L scenarios, where student contact is involved there is here sometimes a level of crossover in individual roles between the main area of administration and T&L activities.

Scenario 10

Tabitha works in an academic department with students, but her role centres significantly on administrative tasks related to the courses that they provide. She organises programmes of study which normally involve students spending significant time in the workplace, and she is involved in the quality assurance of these courses - something which is a particular concern given the level of external contribution from the workplace teams.

A large part of Tabitha's work is the bi-annual Quality Assurance report that she writes. This involves gathering together a significant volume of information but then she writes the document without collaboration. This is a long process which often involves working at home or whilst on the move. Tabitha is interested in the idea of using a repository to store this work-in-progress so that versioning is done for her, the document is available to her via the web, and so that automatic, routine backup of the developing draft takes place.

Tabitha's department has to deal with a range of written assignments, dissertations and theses. At the moment many of these are, by statute, paper documents which are marked by more than one assessor. Following graduation many of these documents might usefully be used by others but at present they are stored in a somewhat obscure office. Tabitha can see that allowing students to submit these materials in electronic form might have a number of advantages: the electronic form could be made easily available to multiple markers; once marked, the document could be made more widely available for future reference in a repository; and the document could be preserved over time.

Scenario 11

Mary frequently works within the University committee structure. She is responsible for providing agendas, minutes and working papers to committee members and for making them available to other interested parties.

Although the team in which she works is given sharing facilities on the University network, which also provides backup, she would like to have in place more flexible sharing arrangements, which encompass not just the team with which she works, but authorised members of University staff. She can see how a repository might help with this, and also how it might address some of her longer term access and preservation needs. (Some of the documents she deals with are the official records of University business and need to be preserved indefinitely.)

Mary may be provided with a report which is to go to a sub-sub-committee. This arrives from its sponsor in a finished, digital form and she associates it with an agenda and other papers for an upcoming committee meeting. (There will probably be a set of previous meeting minutes for the committee too.) The report must be clearly identified as the report that went to this committee. Mary would like to be able to make these pre-meeting papers available in digital form for members of the committee to access, and to inform a small group of other interested parties.

Following the meeting, the report may be accepted, sent back to its sponsor for revision and resubmission or passed up to the next committee level. In these last two cases, the cycle starts again in that the (possibly) revised paper goes into the collection for the next committee meeting in a version that must be clearly associated with the new meeting (and not the previous one). If the report is going to a new committee the new document needs appropriate access permissions for its members. It is possible that the report would work its way up to the highest committee for final approval and this would result in, say, five possibly distinct versions each of which needs to be clearly identified, associated with a particular committee meeting and preserved. At whatever level the report is signed off, the final version may need to be made available to a much wider audience.

The agendas and minutes from all these meetings are normally made available to staff after the event, with the exception that business is divided into Part A and Part B business: the agendas for both parts are published, but only the minutes and associated papers from Part A are published after the meeting. Part B materials contain confidential information, some of which may always be so, other of which may be made public after some time. The status of any such confidential material would be reviewed each time a request was made to see it.

The security of papers pre-meeting relies on the University Computing Service maintaining up-to-date lists of committee members for access by other University systems.

Scenario 12

Lauren works with a team responsible for codes of practice within the University. Although the team in which she works is given sharing facilities on the University network, which also provides backup, she would like to have in place more flexible sharing arrangements which encompass, not just the team with which she works, but staff who contribute to the development of and compliance with the codes; these staff are not necessarily at the University of Hull but may be employees of a partner organisation. She can see how a repository might help with this.

Lauren and her colleagues develop a new document from version 0.1 onwards. Each sub-version is retained. At stages in the development process she would like to make the drafts available to others outside her department for comment. She would like to do this in a linear fashion so that the second contributor can see the comments made by the first, and so on. This is an iterative process which eventually becomes version 1.0 which is then made available to members of staff at the University and in partner institutions. In due course the document will come up for review at which time version 1.0 may be subject to minor modification and become 1.1 (say), or it may need major revision in which case a new iterative process will start with a view to producing version 2.0 and this process will probably use 1.0 as its starting point. 1.0 would need to be preserved for the historical record. Lauren can see how the automatic versioning offered by a repository could help with all this and how a collaborative facility could be used effectively.

The documents produced in the manner outlined above are complex. Lauren would like to see them exposed to their end-users as a collection that is easily and effectively searchable so that relevant documents and sections can be located quickly.

Lauren and her colleagues in the team are also responsible for monitoring compliance with the University's Codes of Practice. This involves to-and-fro work with departments on documents and Lauren would like to have access to some of these works-in-progress before their submission without having to keep asking the department for the latest version.

Scenario 13

Julia is an administrator whose job involves reporting each year on the teaching in her area of the University. This report is made against the background of Codes of Conduct, course specifications and other internal systems.

In order to prepare the report each year, Julia and members of the group she services require access to the latest documents describing University practice and requirements. Not all of the group are members of University staff, some work for partner institutions. At present she and colleagues sometimes find it difficult to locate what they want over the web. Julia hopes that the use of a repository to store these materials would allow quick and effective search and discovery. She understands that effective metadata would be required to achieve this but hopes that the authors of the papers will see the merit in providing it. As a *quid pro quo* she would be willing to spend a small amount of time making sure that her own documents had effective metadata in order to facilitate their use by others.

Current repository usage

The RepoMMan Project Plan indicated that this report should identify current repository usage amongst those who contributed to the needs-gathering process. In the event, this is minimal.

Researchers who were interviewed were aware of various repositories available nationally or internationally. None of those interviewed contributed to these as a matter of course.

None of those interviewed for T&L said that they contributed to repositories although one made regular use of *News Film On-line* as a source of materials and two others were aware of *Jorum*. The CD-LOR survey, quoted above, found that "A relatively large number of respondents indicated that they utilise institutional, national or subject-specific repositories to share completed work." In fact, from 247 responses 48 (19.4%) identified their use of an institutional repository, 17 (6.9%) a national repository, and 29 (11.7%) a subject- or discipline-based repository. (Note that these responses were non-exclusive; a particular individual could have selected one, two or three of these responses.)

No-one interviewed for the administration group used a repository at the present time; this was the least repository-aware group.

User needs

Analysis of the scenarios developed enables the development of a fairly simple set of user needs, although the implementation of them may be less simple.

The RepoMMan team embarked upon the interviews with open minds, but knowing that the University was seeking a repository that would allow development of materials as well as exposure of the final 'products'. This underlying requirement focused the interviews and the subsequent analysis of user needs slightly so that, for instance, the second user need identified below assumes something about the structure of a potential repository in that it should have personal as well as public spaces.

- we take in as a *sine qua non* that a repository interface should not make it difficult to do something that is currently achieved easily. A number of interviewees made this point but several could see ways in which a repository might make processes more easy that they currently need to undertake, and make possible some processes that they would like to undertake.

- the repository interface must allow structuring of a user's personal storage space and have the capacity to hold potentially large numbers of objects, possibly of a range of differing types, and possibly of multi-gigabyte size, for each user

- the repository interface should provide an easily usable versioning facility (it must be easy to version a file *and* to revert to an earlier version). It is noted that versioning of objects containing large digital 'payloads' may have significant implications for storage space

- the materials in the repository should be available through a web browser 24/7, given an appropriate internet connection

- the repository should allow sharing of a private document with a closed group of collaborators or readers and should provide some sort of locking facility so that conflicting revisions cannot occur. Some of the types of sharing discussed would require a responsive approach to changing the membership of access groups; this may be a function managed by others and outside the scope of the repository, however the need is noted

- the contents of the repository should be subject to a reliable, regular backup regime

- the repository must make effective public exposure of content easy and controllable, taking account of digital rights issues as part of that process. Part of this 'easy' process

should be the automated generation of metadata, where possible, and the ability to draw on contextual data of various kinds

- digital objects exposed through the repository must take account of relevant copyright and Digital Rights Management (DRM) considerations and, in some cases, the needs of additional legislation such as the Freedom of Information (FOI) Act

- the University of Hull's institutional repository will contain a proportion of objects to which appropriate records management and preservation processes should be applied. Objects in the repository should contain 'digital flags' to aid this process and stimulate administrative intervention at appropriate stages in these processes

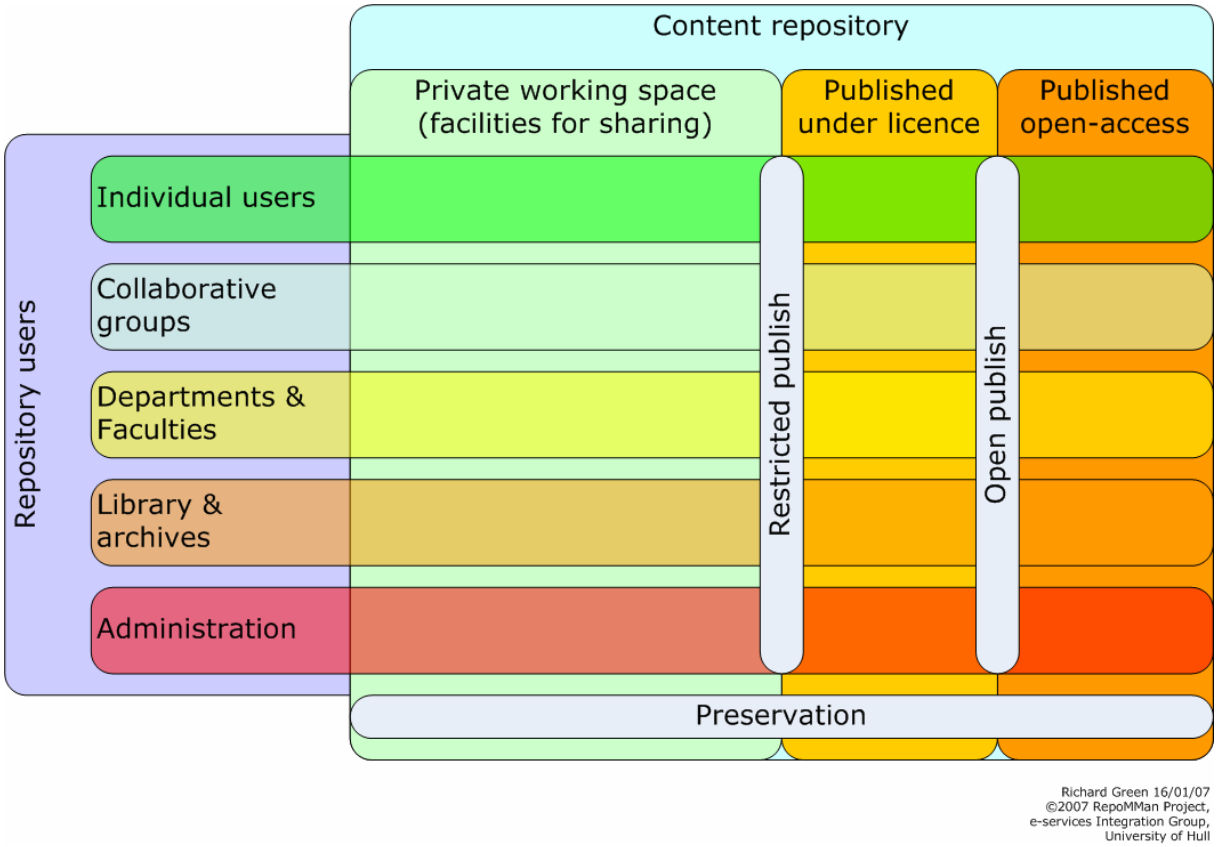
Mapping the repository process

Given the user needs identified above, it is possible to map out functionality that should be supported in the University's repository. Whilst this draws heavily on needs expressed by potential users from the University of Hull, the outcomes from the two surveys in which we have been involved lead us to believe that our proposal is of wider applicability, a belief further supported by informal contact with many involved in repository work both nationally and internationally.

It is important to understand that the requirements detailed here go beyond the brief of the two-year RepoMMan Project to consider the longer-term needs of the University. In practical terms, the RepoMMan Project will not provide more than the basic functionality of the deposit tool.

Repository structure

The following diagram represents our current model for a fully developed institutional repository at the University of Hull.



The repository will have many contributory users: these might be individuals such as researchers or (eventually) students; some might form somewhat *ad hoc* groups and these groups may be granted repository space; the groups might be more formalised - the staff from a department or faculty, say; the repository will have special provision for service providers, like the University Library; and there will be provision for formal groups involved in administration, for instance the Committee Section or the Quality Office.

Each of these users or groups will have a private working space that can be used for the development and storage of digital content of whatever kind. Where appropriate there will be facilities for the 'owner(s)' of the space to make content available to individuals or small groups outside the normal ownership list on an object-by-object basis. These facilities will allow the user to invite comment or collaboration on objects in the space - this may, or may not, imply write-access to the object.

Objects in this space, as in the rest of the repository, will support versioning and ideally record locking where collaborative work is done.

Users may wish formally to publish one of their objects. This may be a restricted process, in that the object becomes available only to, say, members of the University; or it may be a full-publish in which case the object will be available to anyone with a web-browser. At the point of publication a number of processes may be involved. The object would need to acquire appropriate metadata, it would need to be restructured in order that it conform to a standard University content model for dissemination purposes, and its ownership would need to be changed such that it belonged to the repository rather than an individual. These last two processes would be carried out on a clone of the original object in order that the original owner could retain a copy of the work.

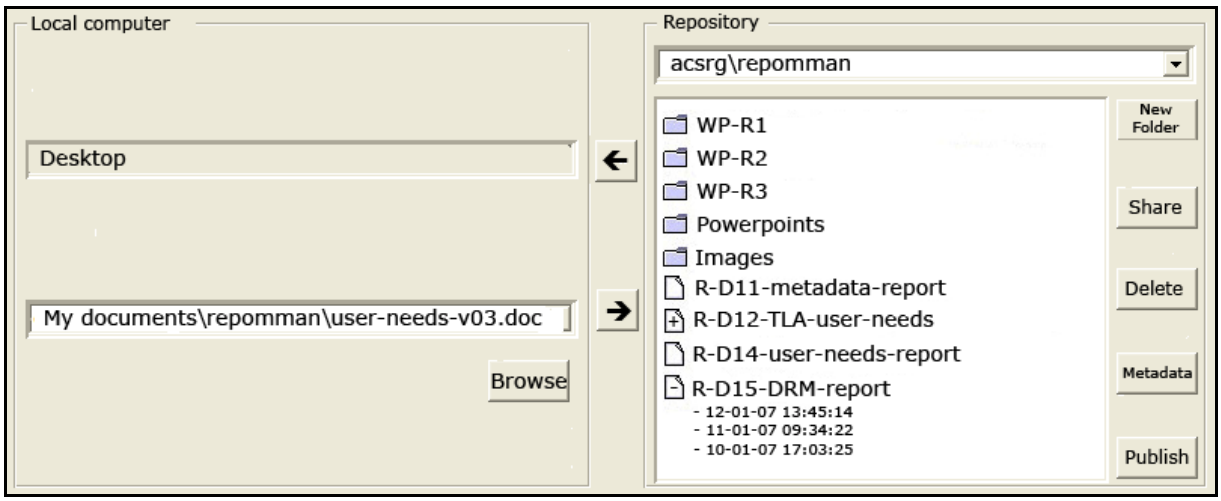
Underlying all this is the possibility of building in processes for on-going records management and preservation where that would be appropriate. Published objects in the repository would carry a small number of flags in their metadata which would be used to alert appropriate staff of the need to perform particular management or preservation functions from time to time.

The entire repository will be fully backed up on a regular basis, and the content will be available to those with appropriate rights 24/7 through a web browser.

Interaction with the 'private' repository space: the RepoMMan deposit tool

In designing a tool for users to interact with the repository, the RepoMMan team was keen to use existing and well-understood paradigms. Thus, it is no accident that the materials stored in the private space of the repository are represented to the user in a display very similar to that used for file browsing on a Windows PC, or that the larger display mimics a well-known and heavily used ftp client. It was felt to be important that as many users as possible should be able to understand and manipulate the interface 'instinctively' and without extensive formal training.

The design mock-up of the interface is shown below.



The left-hand side of the interface allows a user to control the location of the 'digital payload' from a repository object (usually the file that the user stores in the digital object) on their own computer. In the case of a file to be uploaded to the repository this is the current location from which to 'get' it, in the case of a download it is the location where the file should be 'put'. The buttons in the centre of the interface initiate the 'get' or 'put' routines.

The right-hand side of the interface appears to represent a file structure on the repository but it is important to understand that this is not actually the case. In fact, the items represented as folders (eg WP-R1) are repository collection objects which can be expanded to show their contents, the items represented as files (eg R-D15-DRM-report) are digital objects and these, unlike the conventional representation of a computer file, may further be expanded to show versions of the digital payload accessible through the object. The facility exists to create a new 'folder' or 'sub-folder' to any level of complexity, though in reality these are collection objects within the repository.

The 'put' (➔) button attempts to create a digital object from the file identified at the left-had side of the interface in the current 'folder' at the right. If an object of the same name and MIME type already exists a new version is added to it, otherwise a new object is created.

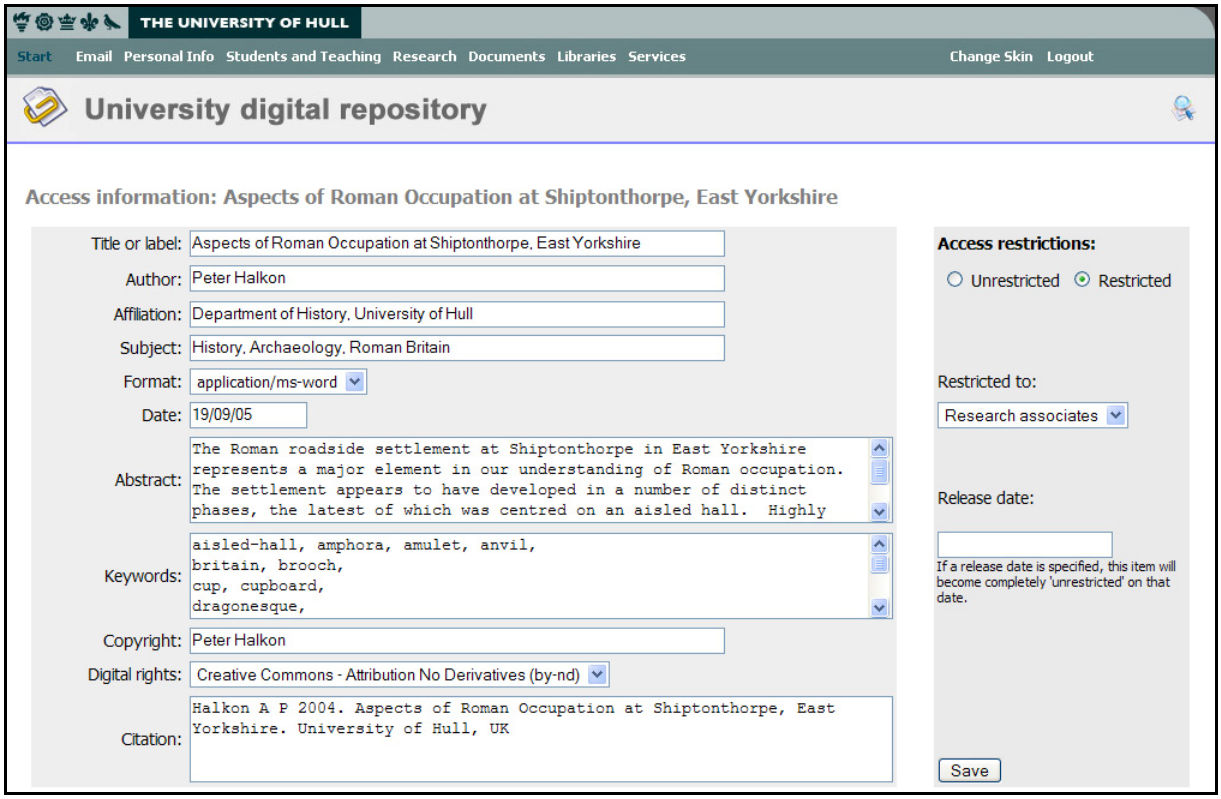
The 'get' (➔) button is enabled only at the object or version level; in the case of an object being highlighted to 'get' it is the most recent version that is provided.

Versions, objects and folders can be deleted by highlighting them and using the 'delete' button.

The digital objects in this 'private' area of the repository are owned by individuals. That is to say that they have an inherent property called 'ownerID' which is set to match the user's log-in name. This property is used by the repository's security system to limit access to the object. As at version 2.2 of the Fedora repository software (January 2007), only a single ownerID can be associated with an object. In order to allow sharing of objects with others, we need the repository to support multiple ownerIDs and this functionality should be available from version 2.3 (scheduled for release 2Q2007). Owners will then be able to select from a list colleagues with whom they wish to share the object: in principle these extra owners could be individuals or an LDAP-maintained list (say 'staff in the Geography department'). This functionality will not be available within the timespan of the RepoMMan Project. Further dependent on this functionality would be a facility to indicate when a collaborator had downloaded the content of a digital object but not yet returned it: the detail of this process needs to be further teased out, but this might be considered akin to record locking so that conflicting revisions cannot occur.

The 'metadata' button allows a user to associate metadata with a digital object. For the purposes of the RepoMMan Project, this will be limited to simple Dublin Core metadata to demonstrate proof-of-concept. In the longer term the University will wish to use the University of Virginia metadata schema[5] which will provide much richer information although this will, as a matter of automated routine, be mapped down to a Dublin Core representation.

The 'metadata' button will provide the user with a dialogue box similar to the one shown below (the image is taken from a design mock-up; at the time of writing the dialogue is not yet complete in the RepoMMan tool):



It has been an underlying tenet of the RepoMMan approach that such a dialogue box should be pre-populated; that facing the user with a blank form at this stage would be to invite poor or non-existent metadata. The RepoMMan tool will generate this metadata from a number of sources and not all of it will be available for user editing.

[5] see http://www.lib.virginia.edu/digital/metadata/index.html

In the short term, metadata about the user will be drawn from the context in which they are accessing the tool, this is most likely to be the University Portal or VLE. In the longer term it is hoped that this can be derived from a University Identity Management System. This metadata will, where appropriate, be supplemented by user-maintained metadata about their current research project or teaching course (as instances) which is itself held within a hidden repository object in their private space.

Technical metadata about the digital payload of the object will be obtained using a tool called 'JHOVE';[6] this metadata will not be presented to the user, rather the entire output of the tool will be stored with the object for potential preservation purposes and certain items (for instance filesize, dimensions of an image) will be automatically inserted into appropriate elements of the Dublin Core metadata.

Descriptive metadata is likely to be derived using a tool called 'Data Fountains'[7] and it is the output from this tool which will form the basis for pre-population of the screen represented above. The user will have the opportunity to correct or supplement this metadata. When the 'metadata' button is used, the system will need to check whether metadata has already been generated so that the user does not unwittingly overwrite any changes that (s)he may already have made.

The final button on the RepoMMan deposit tool allows the user to pass an object on for publication in either the public or semi-public areas of the repository. In fact the process will take a clone of the user's object leaving the original in the private repository space. The object will not be moved forward until it has had metadata added and, if necessary, the 'publish' process will invoke the 'metadata' routine.

The process of making the cloned object available in the repository is beyond the scope of this document and is intimately bound up in some policy decisions that have yet to be made within the University. However, one might speculate that the object goes into a 'holding pen' until it has been inspected, although a class of 'trusted' depositors might be able to by-pass this stage. The object would have its ownership changed so that it 'belonged to the repository' and could no longer be altered by its originator. Normally, it would then be processed, which is to say that it would be dismantled and restructured so that it conformed to a standard University content model to meet the needs of dissemination on the one hand and security on the other. This restructuring may involve the creation of additional datastreams within the object (for instance, surrogate images or alternative text formats). This process of ensuring conformance would be largely or wholly automated.

Storage implications

The user needs interviews have revealed that the University's repository will have to cope with a very wide range of storage needs: 'wide' both in the sense of supporting a considerable range of file (MIME) types, and also 'wide' in the sense of files ranging from a few hundred bytes to tens of gigabytes or larger.

In terms of planning, it will be useful to try and quantify these needs. The interviews have identified the range of file types (technically, MIME types) that will need to be stored. Anticipating the required storage capacity at this stage will be somewhat speculative but should, nevertheless, be attempted.

MIME types

The Fedora repository software adopted by the University of Hull is largely agnostic to MIME type, which is to say that it will store anything. The file representing the user's data is just

[6] see http://hul.harvard.edu/jhove/
[7] see http://dfnsdl.ucr.edu/public-df-cgi-bin/view_page?file=public/index.html

stored as a byte-stream and one looks much like another. It is, however, necessary to store a MIME-type within the Fedora object so that secondary systems - for instance a web browser - know how do deal with the data content when it is retrieved.

Administrators from the University generally expressed a need to have the University handle Microsoft Office documents and pdf files. It was researchers and those from the teaching and learning community who had a broader spectrum of needs including, but not limited to, the following:

- archive formats (for example Zip or Stuffit files)
- audio files (eg: wav, mp3, aac)
- computer disk images (eg: gho)
- data files (bytestreams of possibly arbitrary format)
- database files (for example from SQL, MySQL, Access, Oracle)
- diagrams or CAD (for example from Visio, AutoCAD, MathCAD)
- document files (eg: doc, rtf, rtfd, pdf, xsd, ps etc)
- image files (eg: jpg, jpeg, jpeg 2000, gif, png, psd, tif, tiff, eps, raw)
- large format digital video (eg: dv)
- presentation files (for example from PowerPoint)
- scripts and files related to the design and operation of a computer or a machine (perhaps from C programming language, Java, Matlab)
- simple text files (eg: txt, xml, xslt, css)
- specialist text formats (such as LaTEX)
- spreadsheet files (eg: xls, xsc)
- statistics files (such as those from the SPSS package)
- video files (eg: wmv, avi, rm, mov, swf, mpg (and its variants))
- web pages (eg: html, jsp, php)

This list has deliberately been presented in alphabetical order so as to imply nothing about the relative popularity of one type over another. (The RepoMMan online survey of researchers did produce relative usage statistics,[8] as did the online survey of those in the teaching and learning community undertaken in conjunction with the CD-LOR project.[9])

File size

It is the nature of the files listed in the previous section that some, for instance simple text files, are commonly very small (measured in kilobytes or less) whilst some, for instance digital video files) might commonly be tens of gigabytes in size. Files of experimental data of one type or another can span this entire range.

Storage requirements

It is at this stage that we must try to quantify what these findings imply for the storage requirements of the repository. To do so we must make some fairly arbitrary judgements about take-up and usage but, whilst the figures are open to debate, the process might give us some feel for the order of magnitude that we are talking about.

Let us suppose that the repository is made available in the first instance to 2000 staff and 500 research students. This is deliberately to exclude undergraduate students from the equation for the moment.

Let us further suppose that in the first years, 20% (500) of those eligible start to use the repository in the way that we envisage. Of these let us say that 80% (400) are involved with teaching and learning or research, the balance (100) being in administration.

---

[8] Green R (2005) R-D3 Report on research user requirements on-line survey RepoMMan Project, University of Hull

[9] Margaryan A (2006) *op cit*

Researchers and those in T&L have widely varying needs. At least one T&L interviewee wishes (with good grounds, in our opinion) to deposit 1TB+. Researchers have talked to us in terms of tens and hundreds of gigabytes, other researchers, like those in administration, have very modest needs.

This table is an attempt to reconcile these requirements:

Category	Storage needs	Number	Total storage (GB)
Research/T&L	1 TB	5	5000
Research/T&L	100GB	95	9500
Research/T&L	10GB	100	1000
Research/T&L	1GB	100	100
Research/T&L	100MB	100	10
Admin	100MB	25	2.5
Admin	50MB	75	3.75
		Total	15616.25

A total approaching 16 terabytes.

It would seem then that if we exclude too many 'edge cases' that the repository may need something between 10 and 20 terabytes of storage in the early stages (maybe the first three years). The term 'edge cases' is here taken to mean users with unusual and/or extreme requirements.

As the repository is made available to more and more users there is the increasing risk that the available storage space will be abused. There are many cases known of network users at one institution or another attempting to store whole collections of video downloads which are nothing to do with their work. It may be that the repository will need to consider whether it is possible to impose limits on the storage space available to users or groups of users, and build a check for available space into the deposit process.

Storage type

It is not envisaged that the repository will have to deal with very high load factors in terms of access to storage. There has been discussion of the University Storage Area Network (SAN) being used as the basis for the repository and, indeed, initial work is taking this approach. However, it is our belief that the design of a SAN (high volume, record level access to files) makes it a very expensive storage medium and that a more modest technology - perhaps some form of network attached storage (NAS) - might be adequate for the repository's needs. This investigation lies outside the scope of the RepoMMan project, but is being investigated institutionally. This is not to suggest that such an NAS solution need be any the less robust or reliable, but that the solution adopted should be appropriate to the needs identified in this document.

Conclusion

The scenarios presented here have raised a wide spectrum of issues, many of which are examined more closely in this document. Many of them also raise additional questions, which are continuing to prove valuable in an institutional setting. The user needs identified have led to the design of a repository structure described and have reinforced our original thoughts on how a repository might support *all* aspects of work within teaching, research and administration, not just final deposit. It is also clear that a repository can support a very wide range of use cases across the user groups interviewed. We are conscious that not all potential

users have been covered, but experience to date suggests that other users will have similar needs, even if applied to different digital materials.

To pick up on one area specifically, storage is key to all use cases and user needs identified. Storage can be too easily overlooked as that element of repository development that will simply ‘happen’, a view that is probably reinforced by our acceptance that storage is available already. It becomes a particular issue where particularly large digital materials or datasets are involved (e.g., as identified in the scenarios around engineering experiments and the use of video), but needs to be considered as fully as possible to enable repositories to work effectively. Additional work in assessing requirements will be of value.

This document sits alongside others emerging from the work of the RepoMMan project. It also sits alongside valuable work that has emerged from other projects within the JISC Digital Repositories Programme, and readers are guided to this and these other documents for a full picture of repository user needs.