Content Repositories and Social Networking: Can There Be Synergies?

Paul Dolby[1], Adrian Fish[1]

[1]Centre for e-Science, Management School, Lancaster University, UK

Email address of corresponding author: p.dolby@lancaster.ac.uk

Abstract. This paper details the novel application of Web 2.0 concepts to current services offered to Social Scientists by the ReDReSS project, carried out by the Centre for e-Science at Lancaster University. We detail plans to introduce Social Bookmarking and Social Networking concepts into the repository software developed by the project. This will result in the improved discovery of e-Science concepts and training to Social Scientists and allow for much improved linking of resources in the repository.

We describe plans that use Social Networking and Social Bookmarking concepts, using Open Standards, which will promote collaboration between researchers by using information gathered on user's use of the repository and information about the user. This will spark collaborations that would not normally be possible in the academic repository context.

Introduction

The application of e-Science technologies and methodologies to Social Science research is still in its infancy compared to other disciplines such as the physical sciences. The uptake of these technologies is still slow and there are many Social Science researchers yet to discover the benefits of e-Science. There are many barriers to uptake and in the UK the Joint Information Systems Committee are currently funding a number of projects to determine these barriers. The National Centre for e-Social Science (NCeSS, http://www.ncess.ac.uk) is currently carrying a number of research projects with an aim to identify the barriers to uptake and to determine strategies to overcome these barriers. These projects include the e-Uptake project, which is a project carried out in collaboration with the National e-Science Centre and the Arts and Humanities e-Science Support Centre. Another of these projects is the eIUS project led by Oxford University in partnership with NCeSS.

Many barriers to the uptake of e-Social Science exist but a barrier was identified in the early days of the e-Social Science programme within the UK. The lack of education and training in e-Social Science was identified as a substantial barrier to the uptake of e-Social Science. To facilitate overcoming this barrier JISC funded the Resource Discovery for Researcher in e-Social Science project (ReDReSS) in October 2003 (http://redress.lancs.ac.uk). This project aims to provide training and awareness raising resource in e-Social Science methodologies and technologies to Social Scientists. The project has worked closely with NCeSS, which was funded by the ESRC shortly after the start of ReDReSS, and other organisation such as NeSC to facilitate this aim. More details on the ReDReSS project can be found in the paper by Dolby et al. (2007).

This paper expands upon some of the outputs of the ReDReSS project. This expansion is enabled by new Web 2.0 technologies and ideas, which provide methods for enhancing resource discovery. These techniques were in their infancy at the start of the ReDReSS project but today they are widely used and can be exploited by the ReDReSS project to enhance the discovery of the resources that the ReDReSS project has to offer.

In this paper we will discuss the application of Web 2.0 techniques to our bespoke subject repository, The Learning Space Catalogue (LSC, http://redress.lancs.ac.uk/Learning_Space/). This is a metadata repository containing hundreds of links to resources that raise the awareness of e-Social Science and provide training in e-Social Science.

We will first start by introducing some of the concepts behind the work described in this paper, together will a brief overview of projects/commercial organizations who are implementing these concepts to put this paper in context. In the final sections of this paper we will take a look at the Learning Space Catalogue and describe the implementation of Web 2.0 technologies to the LSC together with relevant standards. In these sections we will describe some of the issues faced with the use of such technologies and how we intend to overcome these. The final section will conclude.

Web 2.0, Social Networking and Repositories: A Brief Overview

We will not provide a formal definition of Web 2.0 as many definitions exit. What we will describe here is a Web 2.0 concept which drives the ideas presented later in this paper.

Web 2.0 technology powered web sites are driven by the users that use them. Users generate the site's content. They allow users to generate metadata describing content and themselves. Users' browsing habits and metadata describing a user are used to link content and users together. This creates a social environment which is user driven and therefore compelling to use. In this paper we are primarily interested in the user generated metadata aspect of Web 2.0. We shall describe how user generated metadata can be used to enhance content discovery within the LSC.

This paper primarily concentrates on social bookmarking although techniques will be described that will link users based on certain criteria. Social bookmarking provides a method for internet users to store links to web content that is primarily of interest to them, but may well be of interest to others researching similar fields; unlike tools provided within browsers to store bookmarks which remain personal to the user, social bookmarking web sites allow users to share their bookmarks with other users.

The bookmarking tools built into browsers allow users to store their bookmarks in folders. Social Bookmarking web sites such as such as Del.icio.us (http://del.icio.us/) allow users to tag their bookmarks with textual labels which they think are descriptive of the resource being bookmarked. Tagging is a process whereby users associate keywords with the particular bookmark. They can then build up a set of tags which are words or phrases that have a particular meaning to them. All bookmarks with the same tags can be grouped together, allowing for easy and quick retrieval of bookmarks which are of interest to the user.

Once tags have been assigned to a bookmark these tags and the bookmark can be made available to other users as well as the details of the users who created the tags and bookmarks. The other users of the social bookmarking site can then find bookmarks using

the tags created and also, through these tags, users with similar interests. They can then provide tags of their own for these bookmarks. Through this tagging process a Folksonomy gradually develops that eventually produces a set of common terms that can be linked to the bookmarks. This Folksonomy improves the discovery of a resource to users of the same bookmarking site. Unlike traditional web search engines, this Folksonomy is created by humans who are more likely to understand the content of the web resource, which is an obvious advantage but also generates a number of issues, one of which is quality assurance. Privacy issues also occur when sharing user's data. These two issues will be focused on in this paper with our approaches to alleviating them.

There are a number of ways in which the issue of quality assurance has been dealt with by other organisations. For example, quality issues were of concern to the creators of Wikipedia, the online user generated encyclopaedia (http://www.wikipedia.org/). This led to them employing experts to act as editors to the content (Andrew Orlowski, 2005). They have also introduced a citation system, with items not cited eventually being removed.

Social tagging has become finer grained with web sites such as Diigo (http://www.diigo.com/). Diigo is a social annotation web site and allows users to tag/annotate particular text on a web site not just the entire document.

In 2004, with the introduction of Connotea (http://www.connotea.org) from the Nature Publishing Group and also Citeulike (http://www.citeulike.org/) social bookmarking was made available specifically for researchers. Both are online reference management tools allowing researchers to manage their references using tagging. Connotea also allows for the automatic extraction of metadata, such as creators and publishers, from online resources. Both of these bookmarking sites, however, are typically aimed at bookmarking papers.

Repositories and Web 2.0

A large wealth of knowledge and information in the UK and worldwide is gradually being built up in repositories. Many of these repositories are "Open Access", allowing any internet users to view the content within them. However for users to add content they generally need to obtain an account or be part of the particular institution hosting the repository. Some of these repositories are metadata only repositories such as the ReDReSS Learning Space Catalogue and others such as the Open Source ePrints (http://www.eprints.org/) and DSpace (http://www.dspace.org/) repositories allow files to be uploaded. Both ePrints and DSpace are currently popular repository software within the UK according to Opendoar (http://www.opendoar.org/), a registry of worldwide repositories. However these platforms are among many that are used within UK institutions.

Repositories can be though of as Web 2.0 applications according to the definition at the beginning of this section. Users can upload their own content and attach there own metadata, although often tags/keywords are provided from a controlled vocabulary. Repositories, however, generally have very simple methods for accessing their content. These include search interfaces and forms. They have limited means to promote content and only offer basic ways of linking related content e.g. by subject. By introducing current Web 2.0 techniques into repository software, the behaviour of users of the repository and the metadata they generate can be used to link and promote content and to promote collaboration, a key e-Science concept.

Although this paper describes work that will be carried out by the ReDReSS project involving our bespoke repository, there are currently a few initiatives attempting to combine

Web 2.0 techniques with repositories. Most of these efforts are concerned with the ePrints software. In 2005, The Nature Publishing Group, through funding from JISC produced a tagging tool for ePrints (see the JISC web site http://www.jisc.ac.uk/). Another JISC funded project led by University of London Computing Centre, SNEEP, has plans to produce a social networking extensions also for ePrints. Finally the PERSoNA project, led by Leeds Metropolitan University and funded by JISC aims, in their own words, to "focus on the individual's interaction with the Repository at every stage. It will investigate how social networking tools might facilitate the stakeholders' connection between the necessary institutional functions of the repository and the individual's own use and exploitation of it."

Since the submission of the abstract for this paper, two other projects have come to light. These were presented at the Third International Conference on Open Repositories, held at Southampton University on 1st to 4th April 2008. The first of these projects, the Faroes project, based at the University of Southampton, plans to develop a Web 2.0 style interface to the ePrints repository software, which will follow best-practice principles from other Web 2.0 sites (Millard et al., 2008). The second project, based at the University of Queensland Australia (http://www.itee.uq.edu.au/~eresearch/project/harvana/), Harvesting and Aggregating Networked Annotations (HarvANA), uses Open Standards to harvest and combine user generated metadata with more traditionally generated catalogue metadata. Their aim is to use such methods to enhance the discovery of content within catalogues and repositories.

A further look at some of the social bookmarking platforms mentioned here and some of the ideas behind social bookmarking can also be found in the article by Hammond et al. (2005) from the Nature Publishing Group. Although now over three years old, a number of today's Social Networking / Bookmarking sites and services are looked at including Connotea and CiteULike.

The ReDReSS Repository: The Learning Space Catalogue 1.0

For historical reasons, reasons that we will not go into here but in essence due to suitable repository software not being available at the time of the project's start in 2003, the ReDReSS project has not implemented an Open Source repository. Instead we have developed our own metadata repository called the Learning Space Catalogue. This has a simple database backend (PostgreSQL) with web based access provided by a bespoke interface (written in PHP).

The user interface started as a simple click and browse interface, allowing the user to view e-Social Science content placed within one of 17 e-Social Science sub-categories by members of the ReDReSS team. These categories were introduced to improve resource discovery, a resource being anything from a paper to a multimedia presentation of a workshop talk.

The interface has now progressed to provide additional functionality, including additional filtering and search options. This functionality is now provided by what we call the, "Content Viewer Tool." A number of additional tools have and are being developed that provide additional functionality to the repository and additional discoverability. These include tools to provide information and statistics on the repository content. One tool, that will form the basis of our Web 2.0 implementations, is the "Bookmarking Tool". This tool allows users to add items from the LSC to a separate list which can be printed or saved to file.

In developing the LSC we have kept to open standards. The Dublin Core attribute set is used (http://dublincore.org). Each record in our repository is assigned a description, which is generated by a member of the ReDReSS team. For each record there is subject metadata field, to which a set of keywords/tags are assigned. However these are not generated from a controlled vocabulary.

The catalogue has a RSS 2.0 feed, which is used to notify subscribers of new content added to the catalogue. This contains basic information linking to further information within the catalogue.

The LSC has basic functionality at present to aid content discovery. This functionality is fairly similar to most repository systems currently available. In the following section we will describe the additional Web 2.0 functionality which we plan to add to the LSC in order to enhance content discovery and collaboration.

The Learning Space Catalogue 2.0

Enhancing the Discoverability of Content

We will be implementing a number of Web 2.0 based techniques to enhance the discovery of content within the LSC. These techniques will utilise users' bookmarking behaviour to promote content within the LSC. Users' behaviour can help better organise content to suit their education and training needs. Basic aids to discovery such as search and our assignment of each resource to a category are provided but the new Web 2.0 enhanced bookmark tool will enable users to classify content to suit their own needs. This process of user classification may, with enough users of the LSC, lead to our own categories becoming obsolete.

Users at present can bookmark items in the catalogue. These are placed into a simple list which can be printed or saved with the editors of the catalogue assigned metadata. The bookmarks are stored in a cookie on the client machine. For our Web 2.0 additions to the LSC, we will need to track users' bookmarking behaviour, an authentication and registration system will be introduced to aid this. This system will be described later. The registration system will also enable us to provide additional social networking functionality through the use of the registered user's metadata and extra information they add to their profile. This system will not affect access to the LSC content and basic functionality of the bookmarking tool will be kept to provide non-registered users with a useful tool.

Registered users will, once registered, be able to generate their own metadata for a particular item in the catalogue. They will be able to add their own description/comment/note to a particular item as well as their own tags. However at present, we will not be offering a service that will allow users to annotate part of a web based resource.

The new enhanced bookmarking tool will offer a number of features. Users will be provided with the ability to sort their bookmarked content either by their own tags (or the tags that we assign to the item). Users will also be able to group their content under their own categories and send their bookmarks to colleagues.

Through user generated tags we can provide other users with an alternative way of classifying our content, which suits their needs. These tags will be shared, enabling others to utilise the tags. With widgets developed to work with these tags, we can promote content, for example,

with the use of tag clouds from our site or other web sites. Popular and most bookmarked items in the catalogue will also be tracked and displayed using widgets.

By storing user's content viewing habits, we will present other users with further content which may be of interest to them after viewing a particular item. This technique is widely used in the commercial web sites such as Amazon to promote useful content.

It should be pointed out that up until now most of the techniques discussed for promoting useful content to users are widely used by Social Networking/Bookmarking web sites but are rarely used within repositories.

The LSC is primarily aimed at educating Social Scientists in the benefits of e-Social Science and also providing supplementary training. We do not want to prescribe a particular learning path to our users as our content is aimed at users with varying competencies in the subject areas that we cover. We will therefore be utilising users' usage of the LSC content, in a novel way, to enhance the learning experience for others. We intend to achieve this by linking content together. As a registered user moves through viewing different items within the catalogue they will be asked a series of short questions aimed at accessing the usefulness of an item against its predecessor. This will then be related to their competencies, which will be accessed from their initial registration and a supplementary question. Care will be taken here not to ask for too much information from the user as this can be off putting. However, this process has a two way benefit. It will allow us to provide other users with the most effective learning paths for a particular topic and we will also be able to offer the user providing us with the data suggestion on items that may aid their learning.

All of the Web 2.0 applications described so far generate a lot of user data. This data can come from their initial registration, extra profile information and the use of the LSC bookmarking service. All of this data can be used to link users with similar interests. For example users with similar tags can be linked. This has the possibility of promoting research collaboration that otherwise may not have occurred. We will be utilising the data gathered on users and their content browsing behaviour to link users and provide suggestions to users concerning other users that have similar interests to themselves.

Another facility that we are investigating will involve making use of the JISC Information Environment (see http://www.jisc.ac.uk and JISC themes). In particular we plan to utilise services such as Jorum (http://www.jourm.ac.uk) to provide links to supplementary learning content on a particular e-Social Science subject. For example we could provide links to C++ programming tutorials when users are looking at content on Grid computing.

Web 2.0 applications present a number of issues. In the following section we will be touching on two of these issues, namely quality assurance and privacy. We will describe our methods to help assure quality in both content and metadata and to protect users' privacy.

Quality Assurance and Privacy

Quality Assurance

As mentioned at the beginning of this paper, quality in user generated data is a major issue. There are a number of ways in which we assure quality within the LSC.

The quality of each item in the catalogue is assured through restricting authority to add new items to the ReDReSS team. All items are sourced from quality assured sources. There are

however plans in the future to allow content to be added by other users but these users will be known and trusted. There are also plans to implement harvesting of content via using Open Standards. However this content will always be presented to users along with information concerning its source and quality.

Metadata will continue to be assigned by the ReDReSS team to items in the LSC to assure its quality. This metadata will always be viewable by the users as the primary metadata. User generated metadata will be presented with information stating that its quality can not be assured.

We will be trialling a method which can assure quality of user generated metadata. A ranking system will be applied to users to indicate the quality of their metadata. This rank will be clearly stated against their metadata. It will also be taken into account when generating for example, tag clouds. Users with a higher rank will have a more weighted tag when determining the tags within a cloud.

The rank of a user will be determined by a number of factors. This includes their profile information, which will provide a mapping of their competencies across the LSC categories but also from information inferred from the content of the catalogue. For example if the user has content for which they are an author in a particular category then their ranking will increase and they will be considered a more trustworthy source of metadata for items within that category. Also knowledge gains when following a learning path will be taken into consideration.

In order to prevent the misuse of the new services of the LSC, we initially propose that there will be a restricted user base. We propose to use the Athens Access Management system to restrict access to UK academics, the primary audience at which the ReDReSS project is aimed. This will not prevent viewing of content and metadata by other users. Since e-Social Science concerns collaboration we will be investigating methods for allowing access to academics from other countries.

Privacy

Privacy is another major issue with many Web 2.0 technologies. There are two areas where privacy becomes an issue in what we have described. The first is the privacy of users' profiles and the second is the privacy of users when relating them to their generated metadata. Strict controls on privacy will be implemented in the LSC.

Unlike many commercial social networking web sites a user's profile will not be publically displayed in this planned implementation. We will aim to link users from the content of their profile plus their bookmarking habits. When a link has been determined the users concerned will be informed. They will then be asked if they want to make themselves known to the other users. If they agree then limited details with be passed to the other user, such as academic interests. At this stage no information will be given that will identify the users to each other. Based on this information, each user can sever the link, accept the link, at which point e-mail addresses will be exchanged or they can use an inbuilt chat system to determine suitability of the proposed link and then accept or decline.

Users will be assigned an ID upon first accessing the service. This ID will be shown against any metadata that the user generates along with their quality assurance rank. At no point will their name be displayed publically. Users will also be able to select which of their content is displayed publically and which is kept private.

The quality assurance and privacy issue solutions that we have described here provide reasonable quality assurance and privacy protection to our users. We don't assume that we have solved all the issues and a number of issues may arise during the development of the service. We will report on these issues when they occur and the solutions which we have devised. However we intend to offer good quality assurance and privacy safeguards in the production service.

Standards and Sharing of Metadata

In the development of the Web 2.0 enhanced Learning Space Catalogue (and ReDReSS project site) we will use Open Standards wherever possible. To speed up production time we will be utilising Open Source code wherever possible.

At present, we are not intending to share user details or user generated metadata. This area presents a number of issues which are still to be investigated, for example IPR issues and further privacy issues. However there are many benefits to the sharing of metadata between other repositories and Social Bookmarking/Social Networking sites. These benefits will lead to further enhancement of the discovery of content within the LSC. They will also enable established users of other Social Bookmarking web sites to utilise the LSC and its services but still retain content from these other sites. We will therefore develop the new LSC with the possibility of sharing data in mind.

The LSC was developed with standards in mind in order to fit in with the JISC Information Environment (Powell, 2005). The Dublin Core metadata attribute set has been used to describe each item within the LSC. This provides a simple method for describing each item, which is necessary for quick addition of resources and their associated metadata to the LSC. Dublin Core is a required standard for the JISC IE.

We are planning to share and harvest content from other repositories, which are part of the JISC IE. The standard used within the JISC IE, which will be use for providing content to other repositories and harvesting content is OAI-PMH. Harvesting content is essential to providing a rich content base for our users. However as mentioned in the section on quality assurance, users will always be made aware that this content is harvested and the quality cannot be assured.

We intend to enhance the searchability of the LSC by applying the tried and tested combination of RDF (Resource Description Framework, http://www.w3.org/RDF/) and XPath queries. We will utilise our current pool of Dublin Core annotated resources to initially build and then adapt an RDF graph representation of the LSC resources.

RDF has the advantage of being able to describe non-web based items too. This means that it can also be used to describe the users of the LSC. It can be used to describe user's bookmarks and the structure of their bookmarks as well as the annotations that users apply to them. All of this information can be output in an XML based syntax, which can be read by a variety of applications. RDF is a well established framework.

There are a many ways to represent information about people and their relationships in machine and human readable ways. The Friend-of-a-Friend (FOAF, http://www.foaf-project.org/) schema is once such way. This is an RDF schema that can be used to supply machine readable information about a person and their relationships.

Another way of representing information about a person and their relationships in a more human readable way, is XHTML Friends Network (XFN, http://gmpg.org/xfn/). This microformat however only provides information about relationships between people. It does not provide information on a particular person. However this format is useful for displaying information about relationships between users, which can be used within Blogs, for example. Each person a user is related to in some way is represented by a URI and a "rel" attribute in the link defines the relation. In the case of the LSC, the relation attribute will be taken from a controlled vocabulary, which can be updated by the user depending on the status of their relationship. For example if a result of their linking with another user through methods described above results in them collaborating, the "rel" attribute could state "collaborating". This will also provide us with a way to monitor the effectiveness of our linking methods.

We intend to support popular standards for sharing user information and will adopt these as and when necessary.

Conclusions

In this paper we have described planned implementations of Web 2.0 techniques to existing repository software developed by the ReDReSS project. This technology will enhance content discovery within repositories beyond that of existing commonly used search techniques. These techniques can also be used to link together researchers with common interests, sparking collaborations that may not have occurred otherwise.

We have described a number of common Social Bookmarking techniques that will be implemented, including some novel techniques that will aid learning through other users' usage of the LSC. These are only a sample of possible techniques that we will be using.

A novel way of helping to assure the quality of user generated metadata, using a ranking system is described; this will be trialled to test its effectiveness.

The service described here will be delivered as a pilot service and we will assess the usage of the Web 2.0 features. We hope that this pilot service will provide an insight for other repository developers into how Web 2.0 techniques can be used to enhance content discovery within their repositories.

Open Standards are being utilised where ever possible to allow integration with the JISC Information environment, other social services and other repositories, although the sharing of information about the users of the LSC will initially be limited due to privacy issues.

Acknowledgments

We would like to acknowledge the JISC and ESRC for their funding of the ReDReSS project. The ReDReSS project is carried out in partnership with the STFC Daresbury. We would also like to thank NCeSS for their continuing collaboration and support.

References

Hunter, J., Khan, I., Chernich, R. and Gerber, A. (2008): 'Open Repositories 2.0: Harvesting Community Annotations to Enhance Discovery services.', *Proceedings of the Third International Conference on Open Repositories 2008*, 1-4 April 2008.

Millard, D., Howard, Y., Wills, G., Watson, J. and Arrebola, M. (2008): 'Towards an Open Repository of Teaching Resources.', *Proceedings of the Third International Conference on Open Repositories 2008*, 1-4 April 2008.

Dolby, P., Pearce, N., Fish, A., Van Ark, T., Crouchley, R. and Allan, R. (2007): 'Supporting the Uptake of Cyberinfrastructure in the Social Sciences and the Challenges Faced.', *Proceeding of the Third International Conference on e-Social Science*, October 2007.

Orlowski, A. (2005): '$10m for a Wikipedia for grown-ups', *The Register*, December 2005.

Powell, A. (2005): 'JISC Information Environment: Technical Standards', *UKOLN*, May 2005.

Hammond, T., Hannay, T., Lund, B. and Scott, J. (2005): 'Social Bookmarking Tools (I) A General Review.', *D-Lib Magazine*, vol, 11, no. 4, April 2005.