Adding Discovery to Scholarly Search: Enhancing Institutional Repositories with OpenID

Ian Mulvany, Product Development Manager, Nature Publishing *i.mulvany@nature.com*
David Kane, Systems Librarian, Waterford Institute of Technology *dkane@wit.ie*

Abstract

In this paper we suggest that the challenges presented by the vastly increased amounts of scholarly information available today can be dealt with using techniques employed elsewhere in the web in 'Social' web applications. In addition to the usual metadata, which is static, social data puts scholarly information in the context of the dynamically changing state of human knowledge - the dynamic forming and dissolution of links between ideas.

Where large amounts of data reside in an existing 'social platform' (in this case, Connotea.org), it can be used as a means of providing 'suggested links' as well as tags. These data can be inserted dynamically into Open Access Repositories, adding an element of discovery. We describe what has been done to that end in terms of integrating Connotea with Repositories, using OpenID and the TaggingTool as well as indicate future directions that could be taken to deepen and make more effective that integration.

Introduction

Our society is both the producer and the custodian of a great deal of knowledge. It is incumbent on us not only to manage the storage of that data effectively, but also to provide efficient ways to access this data.

The key challenge of the information age is to decide how to meet the information needs of individuals, when the overall amount of information available is huge and is growing very rapidly. There are 1.2 billion email addresses worldwide (Radicati.com 2007). Flickr has now got 2.9 billion pictures and that number is increasing at a rate of 100,000 per hour. Social networking sites, such as facebook, myspace, beebo etc. have shown tremendous growth and now constitute a significant fraction of overall website visits. This reflects a change of usage of the web from a passive medium of dissemination to one of social interaction. This moment co-incides with a huge growth in amount of

overall information, which presents a greater challenge in terms of information management, access and evaluation.

This huge growth in information is just as evident in the context of scholarly communication, with Pubmed growing by at least 7% per year it now contains over 1,000,000 items, including articles, editorials, letters, and so on. There are about 20,000 new academic book titles published each year. Springer, which has a 25% share of that market, publishes 5,500 new titles each year and publishes 1,900 journal titles. Elsevier publishes some 1,800 journal titles. This means that there are 250,000 articles per year with 200,000 referees, 1 million referee reports, 70,000 editorial board members and 500,000 submissions. Elsevier's ScienceDirect now holds over 8 million articles.

Not only is there more scholarly information, there has been an increase in the number and proportion of articles that are open access, through open access repositories, as illustrated by the growth in the number of entries in the OpenDOAR database (http://www.opendoar.org/), and through the increase in the number of open access journals. These are listed in the Directory of Open Access Journals, (http://doaj.org/).

In the past, the information available to academic specialists was of such a size as to be easily dealt with through social interaction and the traditional scholarly publication system. Now, with the ever-increasing volume of scholarly publication, the information available to scholars exceeds the capacity for these means of managing and staying current with developments in a field, and something else is required. This is where new technologies can help enhance our scholarly communication system.

Web 2.0, The 'Social' Web

Web 2 marks a shift in perception where the Web is now starting to be seen on its own merits for the first time. An unprecedented and defining characteristic of this new medium is its ability to capture ephemeral nuances of human behaviour through the passive (and active) collection of usage data on web servers.

The social web has an ability to 'turbocharge' the kinds of things we do naturally as humans, capturing the ephemeral data that we leave behind as we interact with each other and with information sources online, and drawing inferences from that data to better furnish us with our information needs. It holds the potential to allow us to interact with data with an augmented intelligence, though we should be careful to point out that the first steps towards these kinds of applications are only just being built out now.

Good Web 2 applications must not only allow users to easily follow the trails left by those who have gone before them, but also to be able to easily deposit their own information 'trail', whether consciously or passively.

Such applications can be called 'Social Software', a term defined by Farkas as meeting two out of three conditions (Farkas, 2007):

- It allows people to communicate, collaborate, and build community online.
- It can be syndicated, shared, reused or remixed, or it facilitates syndication.
- It lets people learn easily from and capitalize on the behavior or knowledge of others.

In each of these three instances the activity concerned is something that happens naturally to ideas (communicate, share/remix, build upon) in any non-Web community. Like all of the most successful technologies, the Web as defined by Web 2, simply

improves our ability to do what comes to us naturally, which is to spread, connect and evolve ideas.

The current problems in scholarly communication include, paradoxically, the problem of limited access and information overload. Full open access eliminates the first problem, but exacerbates the second and new ways have to be found to deal with this problem. In the world of scholarly search and discovery, the sheer amount of information available today makes searching a greater challenge. Something is needed to 'make the world smaller' and stimulate collaborative links between similar or cross-disciplinary researchers.

The necessary tools for handling a future information overload crisis in the scholarly communications world may already exist on the web. These are sites which can provide ways for users to evaluate information and their value in this capacity scales with the net amount of information that they contain about their user base, inventory and usage patterns.

Two good exampes of this are found in Amazon.com and LibraryThing.com. Amazon is concerned primarily with the selling of books and LibraryThing with the cataloguing of individual users' personal libraries. By comparing a given work with others that users have bought/catalogued overall, both sites can generate a list of titles, related to the title they have currently displayed. In LibraryThing, this has been taken a step further and an 'un-suggester' is provided, which indicates books a user is least likely to have (http://www.librarything.com/unsuggester).

In both cases above, high quality recommendations are generated, based on aggregrated user information. Both represent a single location where a large amount of data is concentrated, facilitating the discernment of statistical trends/usage patterns, that can be used to enhance the utility of the application.

In the context of open access repositories, there is no default central store of social data from which information or recommendations can be drawn. Some future candidates exist though, in the form of academic social bookmarking sites. Sites devoted to academic bookmarking include Connotea.org, CiteULike.org, Bibsonomy.org and 2collab.com.

What We Propose

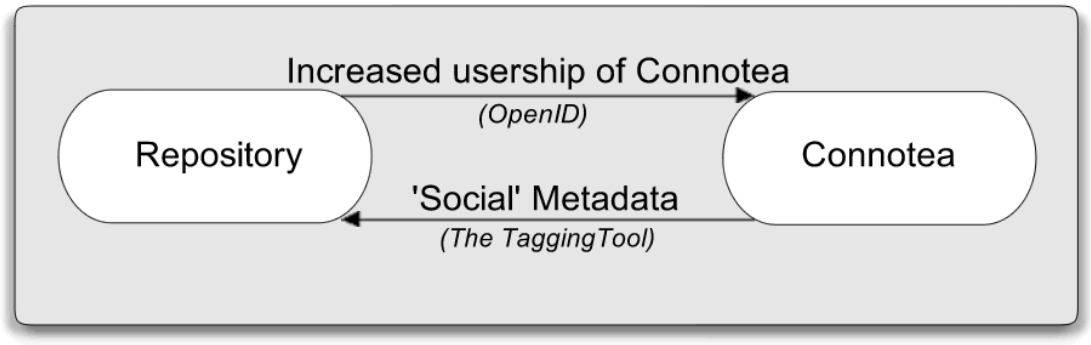
In order to fill this role of a central store of data, Connotea, like any of these sites faces the challenge of increasing the amount of social data that it has so that any trends observed are likely to be statistically significant. This can be done by increasing the number of regular users of Connotea and the amount of papers/tags in Connotea. The way of achieving these goals, that we propose, is to provide a model in which a data silo such as Connotea can be integrated more deeply with Open Access repositories. Such an integration would facilitate the adoption of such silos by academic authors, with repository accounts, and would make it very easy for them to 'fall into' using sites such as Connotea. Integration would have to be as seamless as possible and would involve a re-use of various kinds of technologies, such as OpenID and OAuth.

Repositories at the moment are very much in the paradigm of the library, with collection of static metadata for each item in the collection, being stored electronically along with the item itself – all being accessible via the web. In this picture, there is no room for the creation of new metadata to reflect changes in the state of human knowledge. What would be beneficial is the addition of a new, dynamic layer of 'social' data that reflects the 'neural' breaking and re-forming of semantic connections between ideas that the

contents of the repository represent. This is a big ambition and may possibly be solved in a number of ways and one way of doing this is to keep open access repositories very much as they are today, but to add an extra 'social' layer to them - fed by data from a central 'social platform'.

By linking Connotea with institutional repositories, value can be created for both parties. The value of the data on the repositories can be enriched by the insertion of social data from Connotea and registered users from linked repositories can be disposed towards automatic membership of Connotea, boosting the quantity and quality of Connotea's member base with academic authors. The relationship between Connotea, as a data 'silo' and the repository is reciprocal and self re-enforcing. In the first instance, useful data can be included in the repository record pages. This information might include tags and related articles, which enhance search and discovery. The challenge here is for Connotea to be able to serve up useful tags and apposite links to related articles. This reciprocal relationship is illustrated below in figure 1. Parts of this process have already been put in place (also illustrated). These are OpenID and the TaggingTool.

Figure 1



On one side, Connotea now supports sign-in using the OpenID protocol. Some initial work has been done on the WIT EPrints (http://repository.wit.ie/) server to facilitate the OpenID server end of this relationship. As OpenID is an open standard, it may more easily be deployed on all repository platforms. On the other side, Connotea can inserts tags and related items into the repository. This is facilitated in EPrints by the 'Tagging Tool' (Toth, 2006). The development of the TaggingTool was funded by the Joint Information Systems Committee, as part of their PALS Metadata and Interoperability Projects 2 programme. The purpose of the information provided by the Tagging Tool is to enhance the ability of a visitor to the repository to evaluate and discover information.

To date the uptake of this tool has been pretty low by repositories. Perhaps the biggest reason is that the tool requires users to log in to a Connotea account before seeing data pulled from Connotea. In addition the Tagging Tool is only as comprehensive as Connotea, and will not have tags and recommendations for more than a fraction of the papers that are likely to be found in any open access repository. The Tagging Tool serves as a model for the insertion of extra data into a repository record page. The hope with this inititave is that by adopting OpenID and OAuth the log-in barrier that existed with the Tagging Tool can be overcome and the ambient background trail of the user can flow between the repository and the data silo with minimal effort from the user.

The idea of linking is central to the use of OpenID. Although in a system where most of

the people who use the repository are not connected to that repository via a log in mechanism then OpenID will not be a candidate (indeed even where OpenID is an option not everyone will choose to use it), however where it is used OpenID and OAuth can accelerate the connection between data silos. In essesence what we seek to build in the long term is a system that can lower the friction in brining a users habits and history with them from one reading location to another in order to improve the recommendations that might be generated by the system. By connecting the reader to a society of other readers there are two types of substantive information that can be drawn on that would not be available without this connection. The first is only dependent upon the resource at hand and this is information such as extra meta-data, like tags supplied by other readers, or recommendations based on similarity of the item at hand. The second kind of information is a more personal recommendation, such as information like the number of people in the person's network who have rated the item at hand, or an Amazon-like recommendation for suggested reading. The Tagging Tool was supposed to offer the first kind of extra information, however it depended on a log in to operate. This was a fault in the design of the tool. OpenID could provide a quick fix for this, however beyond this, OpenID in used in conjunction with OAuth can drive a distributed recommendation system. This kind of integration couldprovide the basis for a new 'discovery' layer to be added to institutional repositories globally.

There are still some challenges. How do you increase the chances that people visiting your repository will have an OpenID? It may be that in a few years OpenIDs will become as pervasive as email addresses, in which case univerisity authentication systems may support OpenID, but at this point that remains How do you ensure that the data silo you are connecting to has a large enough overlap with the content of your repository to make any recommendations useful in any way? By looking at a technology that is based on an open standard we rather hope that it may be possible at some point to use such a mechanisim to aggregate social data from a variety of data silos, hopefully increasing the chances of a good match

As OpenID is an open standard, it may readily be deployed on other repository types.

What OpenID Is and How it Works

OpenID (Willison, 2007) is a technology that aims to allow an easy management of login details to multiple web sites. It is a single distributed, open standards based single sign-on system. An OpenID is a URI that is owned by the user. For instance one of the authors of this paper has the OpenID, http://www.mulvany.net/. This URI is registered with an OpenID provider. Instead of creating a new account with a site that supports the OpenID protocol, along with an associated user name and password, the OpenID is provided. Sites supporting the protocal are referred to as relaying parties, as they relay authentication through the OpenID provider. The relaying party being accessed sends the user to their OpenID provider for authentication. At this point the user can decide which information to share with the relaying party, and for how long. The relaying party takes information, such as username and password details, for the user from the OpenID provider. In this way access to any site that supports OpenID is managed with one set of credentials, those to the OpenID provider site.

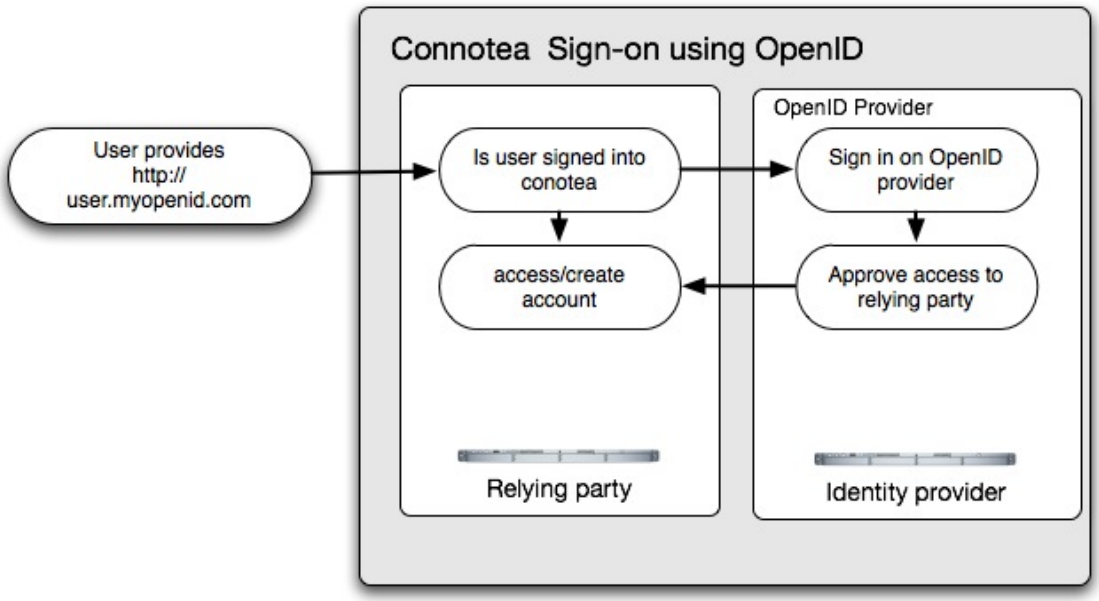
OpenID is not a unique identity, one person can have as many OpenIDs as they feel inclined to register. OpenID is not a definitive way of identifying a person, they can instruct the OpenID provider to return anonymous information to the relaying party. OpenID is a distributed open standards way of managing single sign on.

There are two issues to be aware of at the moment. The first is that in order to be a full solution to single sign on across the Internet every person would need to acquire an OpenID and every site would need to support the OpenID protocol. It is gaining traction with the early adopters and there is strong evidence to suggest that adoption is only a matter of time, (Powell, 2007). OpenID is increasingly gaining adoption among large sites with organizations such as AOL, Google, IBM, Microsoft, Orange, Verisign, Yandex and Yahoo! acting as providers. In effect by default many people now have OpenID's without realising it. For instance a Flickr account homepage can be used as an OpenID once the Yahoo! OpenID client has been activated for a given Yahoo! account. In principle then, the situation should be easy. Pick one OpenID and one provider and use this to log in everywhere. In reality things are probably going to take a little longer to be that simple. For instance, Yahoo! will allow you to use your Yahoo! ID as an OpenID, but at the moment you can't assign an OpenID from another OpenID provider as a login that Yahoo! will accept. Google is allowing commenting on Blogger blogs from any OpenID account, but not access to Google accounts, and related services such as Gmail. The second, and arguably more important issue is security. By having sign-on to multiple sites mediated through one site, the OpenID provider, phishing attacks on that transaction can do potentially greater damage.

In spite of these issues OpenID has some advantages. Installing the services to manage an OpenID server or an OpenID relaying party is easy and there are well supported libraries available in a host of languages. For this project we used the Perl Libraries available from http://janrain.com/. Of the many attempted solutions to single sign-on on the Internet this is one that has actually gained traction due to it's low-cost, easy implementation, and reliance on open standards.

The OpenID foundation was formed in June 2007 (http://openid.net/foundation/) with support form Google, IBM, Microsoft and Yahoo! to protect the IP and trademarks related to OpenID.

Figure 2



Bringing Connotea's Social data back into Repositories

Initially the hope of using OpenID was that it might provide a means by which users of a repository would be able to get an automatic account on Connotea, should they wish. By making Connotea an OpenID relaying party, and by making the repository an OpenID provider, people logging on the repository would have automatic access to Connotea. It became clear after beginning work on the project that most users of repositories access them either through institutional accounts, or through anonymously, and as a result we have modified our expectations somewhat. For the cases where a person accessing a repository is doing so through use of an OpenID we hope to implement our original idea. The hope now is that OpenID will lower the barrier to access to Connotea and in this way increase usage of the service. In addition we hope to develop a better version of the Tagging tool that does not require sign-on to Connotea in order to extract meta-data and recommendations. We also intend to look at the way in which Connotea provides recommendations. In book-related sites this can be 'users who bought this book also bought…' or 'also read'.

The difference between books and academic papers, such as those that might be bookmarked on Connotea, is that each paper is going to acquire far less 'social metadata' than might a book, which has a much lager readership. There is therefore, much greater overlap between readers 'libraries' of books [librarything] than 'libraries' of references [connotea]. That is to say; where an article link occurrs in more than one Connotea user's library, there is likely to be virtually very little else in common on their 'bookshelves' from which interest profiles may be inferred for any particular article.

So, this calls for a different way of extracting meaningful inferences about what a Connotea reader finds relevant. Perhaps the granularity could be at a journal title level. Given that there are over 200,000 journal titles in the world, it is possible that this could give meaningful data to a searcher. For instance, by informing them of the most popular articles from a given journal that have been bookmarked in Connotea, or all of the tags associated with that journal rather than with that paper. In any case, it is important and likely that Connotea will come up with other ways of 'profiling' the concepts that are represented by its bookmarks.

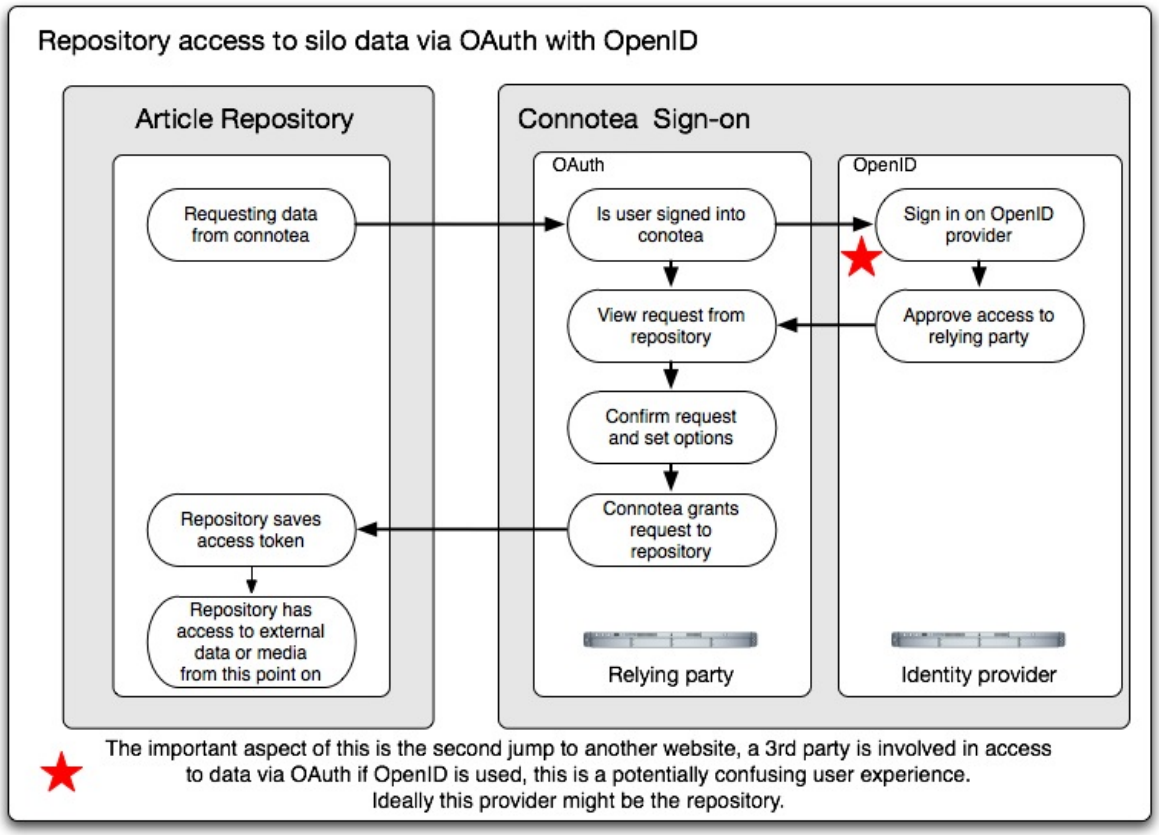
Future Directions

Although the project is still at a relatively early stage there are some definite future directions that suggest themselves. As pointed out earlier a limit to the usefulness of this approach is the likely number of limited institutional readers that will come to the repository with an OpenID. If they are signed in at all it is probably going to be through an LDAP server. One future avenue of investigation is to implement an OpenID server that works with an existing LDAP setup (http://www.openid-ldap.org/).

One way of moving information from Connotea to the repository is through a javascript query. As we mentioned before this information might include extra meta data about the resource at hand, or it might include recommendations based on a user profile. An interesting possibility for people who have a Connotea account and who are using the

repository is to make their Connotea library accessible from the repository. The best way to do this is through an OAuth (http://www.oauth.net), not to be confused with OpenAuth. OAuth is a protocal for allowing two services (desktop application to web-application or web-applicaiton to web-application) to share user data without the user having to share login information between the services. OAuth is a very complimentary technology to OpenID, but is independent from it. Through implimenting OAuth on both sides a user could set up their repository account to pull in data from Connotea. The user is sent to Connotea to confirm the data request from the repository and upon authentication Connotea and the repository share an API key which allows the repository to access user data from Connotea without the repository having to be aware of the users Connotea login credentials. The authentication step can be meditated with OpenID, though this is optional. Once the API key has been exchanged the user can access data from Connotea while logged in to the repository without having the undergo this authentication process each time they log in to the repository. This process is described in figure 3, below.

Figure 3



They key theme in what we are engaged in is lowering the barriers of communication between a repository and a data silo. We are also interested in experimenting with ideas such as auto bookmarking items visited in the repository into the silo as a way of capturing reading trails of the user, though the best approach to this idea remains to be determined.

Ultimately, Connotea and similar sites, do not live in isolation, but can be connected to other services or repositories. We envisage an evolution of these kinds of sites as meta-data mediating services. The hope is that a standard set of API calls will emerge

that may be data-silo independent so that people wishing to build on top of such half-way houses will be able to do so without having to implement different calls for each service.

Conclusion

So, in addition the traditional metadata associated with each record, like author, abstract, date etc. There could be a new level of data that includes information about who else has the article and what other articles are similar as well as perhaps direct links to the full-text of those articles where institutional subscriptions permit. Such 'social' metadata immensely quickens the rate at which relevant pieces of information are sifted from irrelevant information relative to a particular context, enhances the discovery aspect of search and could constitute a new kind of 'grass roots' peer review.

A scholarly communications system that adopts the features that distinguish many of these social web applications will be far more compelling and useful than a passive search portal. Adding a 'social' dimension to scholarly communication will have a synergistic effect – increasing the ease with which relevant materials can be sought and discovered by researchers. By the same token, potential for collaboration may also be more easily identified.

A key foundation for this could be the integration of Connotea with the most commonly used repository software using OpenID, complemented by the introduction of social metadata from Connotea into open access repositories.

References

FARKAS, M. G. (2007)
Social software in libraries: building collaboration, communication, and community online., Medford, NJ, Information Today Inc.

POWELL, A. RECORDON, D. (2007)
Main Articles: 'OpenID: Decentralised Single Sign-on for the Web', Ariadne Issue 51. *Ariadne*.

RADICATI.COM (2007)
Business Social Software Market, 2007-2011. London, Radicati Group Inc.

TOTH, F. & MCALEXANDER, S. (2006)
TaggingTool for Connotea. PALS Metadata and Interoperability Projects 2 programme. Joint Information Systems Committee.

WILLISON, S. (2007)
The Implications of OpenID. Google Tech Talks. video.google.com. (http://video.google.com/videoplay?docid=2288395847791059857)

Links

OpenDOAR database - (http://www.opendoar.org/)
Directory of Open Access Journals, - (http://doaj.org/)
LibraryThing.com UnSuggester - (http://www.librarything.com/unsuggester)
Social Bookmarking Sites: - (http://Connotea.org/, http://CiteULike.org/,

http://Bibsonomy.org/, http://2collab.com/)
The OpenID foundation - (http://openid.net/foundation/)
OpenID-LDAP Project - (http://www.openid-ldap.org/)
OAuth - (http://www.oauth.net/)

Acknowledgements

We would like to thank Gavin Bell for assistance with the figures.