Interview with Kevin Emamy

Martin Fenner, Gobbledygook

January 30, 2009

One interesting session at ScienceOnline09 was Social networking for scientists, moderated by Cameron Neylon and Deepak Singh. We now have so many of these social networking sites, that it becomes difficult to differentiate between them and to see how they can interact with each other. One important category is social bookmarking sites for scientists. I spoke with Kevin Emamy from CiteULike to find out more.

Most of the CiteULike team. Kevin is the second person from the right.

1. Can you describe what CiteULike is and does?

CiteULike is social bookmarking for research papers. It enables you to easily store references online, share them in groups and discover new ones.

It's run as a web service, so everything is always stored online. You can access it from any computer (most of us use more than one these days) and if you're lazy and stupid like me you're much less likely to lose anything. Sharing and discovery obviously work better this way too.

The social side of it is simply that by default everyones posts are public and you can see who has bookmarked the same papers as you, what tags they are giving them and from there you can go on to browse what else they are bookmarking.

And then we have groups which are shared libraries of posts, newsfeeds for every user action, RSS feeds, watchlists, CiteGeist (which is a hotlist of posts) and obviously search as ways to encourage this social discovery.

2. Do your users basically just store their own bookmarks or do you also see a lot of use of the social networking features?

In the 30 days up to the Christmas Holidays 08, CiteULike received 116,477 posts and 6,189 of those were copies made directly from other user's librarys, which is the successful end result of the social stuff. That doesn't count people who find an article, go to the original source to read it, and then post it.

The other social stuff is hard to quantify without diving into the weblogs. Also, once you get into the realms of counting pageviews and linkouts etc. it becomes

easy to mislead yourself about what is really going on, which is why we use "posts" as a genuine metric (assuming you are on top of the spam).

But the social discovery stuff is certainly used. Groups for example; well over 50% are active, by which I mean people have posted to them in the last 90 days. Watchlists (following another user's posts) and RSS are very active. Many people search the site, either directly or through RSS feeds.

One thing that is definitely true is that many more people browse the site than actually register and post, by a factor of say 5:1, and I'm discounting the random Google traffic that just bounces on and off.

Of course, the service would never have got off the ground if it didn't provide individual value to a user without any social features, so yes, many people use it just like that and always will.

3. How is CiteULike different from other social bookmarking tools, e.g. Connotea, 2collab or delicious?

It differs from delicious because we extract and bookmark citation metadata along with the URL, so it's aimed at professional researchers and scientists. We have over 50 sites where we do this, covering most of the online sources of journal papers. Because of this specialization our userbase is very different and therefore much more relevant when you look at the community features. The tags, as one example, are very specialized.

There are lots of differences in detail with the other two 'scholarly' services but it seems that the users have voted with their feet (or should I say mice); CiteULike is far and away the most popular service. If you count the number of papers posted we estimate that CiteULike is currently 3-5 times the size of Connotea both in total posts and posts on a daily basis (2 million+ posts and very little spam for CiteULike vs. 650k posts including a significant proportion of spam for connotea, 3k-5k daily posts for CiteULike vs. 1k to 1.5k claimed for Connotea).

4. Does CiteULike integrate with other social networking sites and services for scientists, e.g. Connotea, Mendeley or FriendFeed? Does it integrate with desktop reference managers?

You can export and import files to and from pretty much all these services.

In the case of FriendFeed I have seen many people using RSS feeds to display their CiteULike posts there, which is great, it's a really good service.

Mendeley are based in London like us and we have begun to discuss ways where we can integrate more tightly, the point being to make the workflow better for users of both services.

5. What is your policy regarding personal PDF files uploaded to CiteULike?

Only the person who uploaded a PDF can download it, so CiteULike is acting as an online storage drive.

We have also allowed uploads to "private groups" which are invite only and otherwise invisible. In this case the user will commit that they have a right to distribute the document to the people in the group,
who they already know.

6. What are your responsibilities within CiteULike?

There are only five of us, so we don't really have roles. Unlike the rest of the team, I can't write code or design applications, so I have to leave that to my esteemed colleagues.

One of the things I try and do is to promote CiteULike and increase our userbase and traffic, which is a bit of a capricious art, but one way that has worked for us is to try to engage with, dare I say it,
the publishers.

Springer, who we recently agreed a sponsorship with, have been invaluable to us in this regard. They are one of the few major publishers who are really progressive and actually understand this stuff.

7. What did you do before starting to work for CiteULike?

We all worked in the software industry, sometimes together. Richard wrote CiteULike as a tool for his own use when he was back doing research at university.

8. Do you want to talk about future plans for CiteULike?

Well the team are continuously improving the service, making it perform better, fixing things that break etc. I wish I could count the number of improvements that have been made over the last 18 months. That is part of the secret to it's growth, the users can tell when the developers are making an effort. We have a very active newsgroup where users make feature and functionality requests or report problems. I have regularly seen my colleagues put stuff live in a matter of hours following a user request, which is one of the advantages of a service that is run by it's developers.

We have wanted to do some kind of recommendation system for a long time; automated collaborative filtering or whatever. We certainly have the data to do it. But it is, I'm told, a hard problem if you want to get useful results. It's possible we'll work with someone else on this.

It was really interesting to see PLoS announcing the variety of impact metrics they want to publish about their articles, a small part of which is going to be

social bookmarking posts. We are currently giving them this data on request, it would be nice to allow anyone to get this sort of data out of CiteULike themselves (we do already make the whole CiteULike dataset freely available for download).

Another example of that is that a lot of other sites want to display CiteULike tag data along with the relevant articles; there are 6.7 million user created tags on CiteULike now (that's total not distinct)
and those tags have been given by people who know the subjects. I suppose I am talking about some kind of API, which would also allow reference managers etc. to integrate with CiteULike more easily, as you asked about above.

Having said that, there is a great temptation to continually add more features or try to make it into something different to what it really is and I'm not sure if that is the right approach (though we have been guilty of it).

For example, one of the biggest issues we have is that CiteULike is good at matching articles that have been posted by hand but not articles imported by file upload. By "match" I mean identify that two articles from different sources are in fact the same, which is the basis of the social stuff; seeing who is reading what you're reading etc. Seeing as 40% or so of our posts come via file upload, it would be great to fix. Matching is done by DOI and PMID; however we ran some tests that showed only 5% of file uploaded articles contain these, so maybe we'll try a different approach.

So what we try to focus on is to make it better and more efficient at what it does well: social bookmarking for research papers. There is plenty of work to do in that regard.