Building Social Networks from Institutional Repositories

David Tarrant[1] and Les Carr[1]

School of Electronics and Computer Science, University of Southampton
davetaz, lac @ecs.soton.ac.uk

Abstract. An Institutional Repository may offer a .set of services. to its local users, supporting the publication of research. More importantly, the repository also forms a key component in the global scholarly communications environment. In this presentation we investigate the role of the repository on a global scale by witnessing the effects on a changing economy and also show how worldwide collaboration networks can be predicted using the strong social links found in repository metadata.

A discipline represents an international community of scholars who consume each others research outputs and in turn contribute new material. This loose affiliation of institutions around particular kinds of material can be treated as a form of social network. The same flows of information in terms of knowledge transfer between countries can also be shown to tie with the growth, or shrinking, of the economy in that area.

As repository managers we have been encouraged to discover and expose broad statistics that show quantity as the ultimate good. Examples can be seen widely on the front pages of many repository systems, including dspace.mit.edu, who show a map of the world on their front page illustrating downloads from 149 different nations. In this paper we attempt to refine the stories a bit more to talk not about how hard we can punch, but the network of scholars that we have affected.

The Electronics and Computer Science EPrints publication repository (EPrints ECS) at the University of Southampton makes an ideal candidate for such studies. This repository, which has been running 10 years, and collecting access metrics for just under half of that, contains records relating to over 15,000 publications.

Before a scholar can consume an article via download, it has to be discovered, a process usually involving a search engine. Doing a quick study on the pathways people take to find content reveals no surprises in that Google dominates the top referrer spot with well over double the number of referrals of it.s closest rival. This figure is consistent over the past 4 years showing the trend for people to resort to search engines to give a spread of results even when they may know a direct source of information in their subject area.

Looking more closely at the data in EPrints ECS, specifically the top ten most downloaded items, reveals that this dominance by google does not apply to every publication evenly. The 2 publications shown in figure 1 are both taken from 2 publications in the top 10 in 2010. Here we can see that while the publication on the right is almost always downloaded as the result of a search (73%), while the top publication in the repository by downloads is in fact cited by a wikipedia article on the same subject "The Semantic Web". Because of this link the percentage of referrals from external websites is uncharacteristically high at 35%.

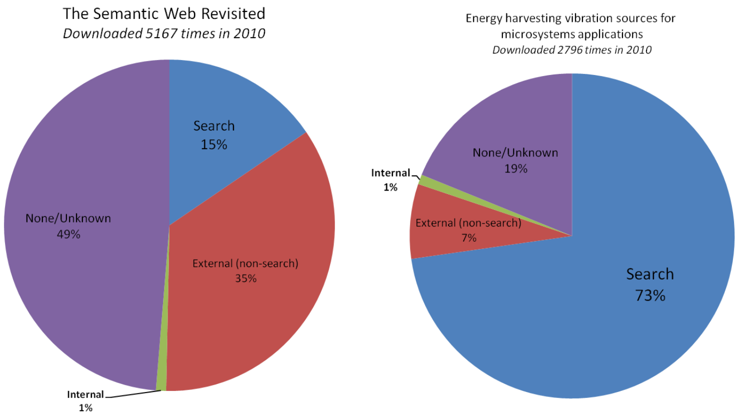

Fig. 1. Referrers for 2 (out of top 10) eprints, showing significant differences

It is also worth noting the high percentage of unknown referrers shown in figure 1. These have been growing in number over the past few years due to the growth in popularity of the google chrome and safari web browsers, both of which do not send referrer headers when opening content in a new window or tab.

Traditionally, statistical measurements of impact are used to help guide a reader in the direction of high impact and well respected articles. While citation count is regarded as the "gold standard.. impact metric, studies have shown that download metrics provide a good early indicator of subsequent impact (1). Thus exposing download metrics can assist and guide users when deciding weather to read a particular scholarly resource.

Although the evidence of an individual download, or referrer statistics, does not allow us to accurately track individual researchers. By realising the value in the links which people are publishing on web sites and how these are uses, we can begin to realise the shared interests of teams of people located in different institutions globally.

On a global scale, looking at the consumption of material within the repository allows the tracking of shifts in global trends affecting the institution. The EPrints ECS repository has received downloads from 215 different countries (according to the ISO 3166-1 country codes[1]) and figure 2 shows how the consumption of materials from the ECS repository has evolved over the last four years.

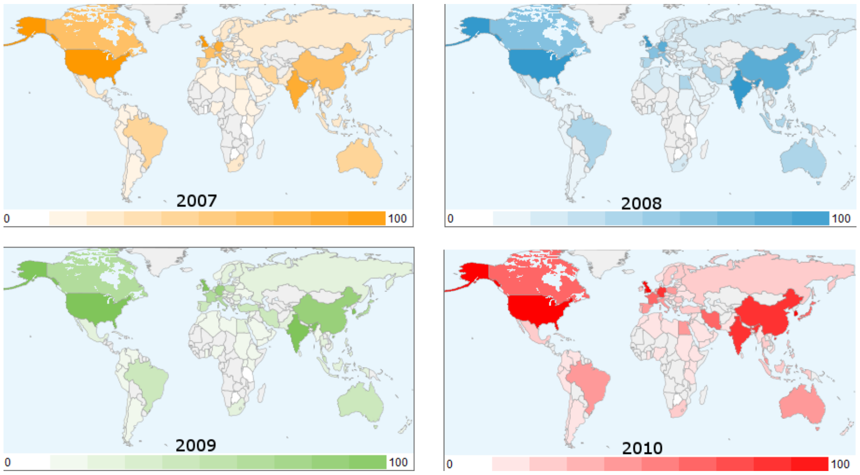

Fig. 2. Growth in global downloads from the EPrints ECS repository

While the western world, specifically the UK and US, lead the number of access to the ECS repository, over the last 4 years China, India and Korea have seen a huge growth in the usage of the repository. This can be attributed to the changing economies around the world, as at the same time the number of downloads from these countries grew, the US figure took a dip in 2010.

Download statistics can help guide users and repository manager alike when making decisions about an article or judgement of popularity. After discovery, users also have the opportunity to link to published articles from websites such as Wikipedia, giving a further mechanism of discovery. As shown such links could prove very beneficial to both the author and institution in terms of downloads, citations and later impact. Studies of such links have also shown that the location of links on the web can also be used to reflect an institutions standing (2).

Links can also be created between users and the content which they download, such information can also be used to create links between people. The download

[1] ISO 3166 Country Codes - http://www.iso.org/iso/country_codes.htm

data that we used to tell us about ourselves, can actually be used to construct a loose social network defined by a set of institutional participants who share an interest in similar research artifacts. This opens up a whole area of investigation around repositories identifying potential institutional collaborations in specific areas.

Figure 3 shows some early experimentation with such networks, in this case between academic institutions worldwide. This small selection of data represents a the interest in a number of popular records by many different institutions, here yellow nodes represent the institutions and the darker red nodes the scholarly works they share a strong common interest in.

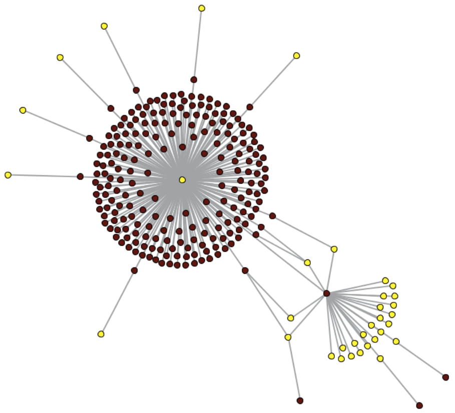

Fig. 3. Connecting institutions via scholarly publications

In figure 3 the node connected to the highest number of publications represents the University of Southampton. By limiting the data to instances where a publication has been downloaded a significant number of times, each arc represents a strong link to a publication. While there are many single links to institutions at this level of sampling, one publication clearly links a great number of institutions together. Shown in the bottom right of this diagram this publication represents the most downloaded item over the last year, referenced earlier in figure 1.

In addition to forming a social network of institutions, we have also demonstrated the potential impact that sites such as Wikipedia are having in forming these social networks. In conclusion, current techniques use repository citation metadata to identify existing collaborations. This work outlines a method allowing the identification of new collaborations before they happen.

Bibliography

[1] Brody, T., Harnad, S.: Earlier web usage statistics as predictors of later citation impact. Journal of the American Society for Information Science and Technology **57**(8) (2006) 1060–1072

[2] Thelwall, M., Harries, G.: The connection between the research of a university and counts of links to its web pages: An investigation based upon a classification of the relationships of pages to the research of the host university. Journal of the American Society for Information Science and Technology **54**(7) (2003) 594–602