A Challenge to Research Libraries – Don’t Just Display Your Data, Use It![1]

David Lawrence, Department for Publishing Infrastructure (PI), Linköping University Library[2]

While the library world struggles with the complexities of open research data and is often immersed in an attempt at reaching a perfect, all-encompassing solution, one should remember that today there are already countless examples of online data sources freely accessible. At the same time, we are not overwhelmed with examples of how the library world makes use of this data, including traditional “library data” in order to enhance the services we offer our users. The challenge I issue in this article is “Let’s get going!”. To stimulate the imagination I will offer a presentation a couple of different examples of how Linköping University Library has made use of different data sources to create new services which are rather highly used.

Over a number of years, Swedish universities have been collecting the bibliographical information about their researchers’ publications in institutional repositories. These days, for many universities, this information is quite complete and gives a good picture of the research output. A large amount of work goes into collecting this information and it is used to some degree internally in performance analyses and in yearly reporting. Additionally, some use the data to generate webpages with lists of publications. Relative, however, to the amount and kind of information that is available, very little use is made of it so far. For many publications, abstracts are available; for perhaps 20% the full text is accessible.

Aggregated bibliographic information about research publications in SwePub

Each Swedish university has its own repository and these are harvested on a daily basis by the national database SwePub and so SwePub presents a good picture of Swedish research publications for at least the last five years. Relatively speaking, SwePub contains considerably more data than that delivered by individual contributing repositories but is used to only a fraction of its potential.

At Linköping University Library (LIUB) we use SwePub to generate publication lists for the Energy Systems Research Center. The Center is no longer active but was a partnership between Linköping University, Chalmers University of Technology, Uppsala University and KTH Royal Institute of Technology. LIUBs bibliometrics group has also used SwePub for assignments of a national nature.

SwePub has an API which is essential for allowing others to make use of the data. The latter is easy to use but no one format provides all the bibliographical data for a given publication and it occasionally stops working reliably, which makes it tricky to build services around. A new interface is under development, SwePub for analysis and bibliometrics, which offers more manual search

[2] Email: david.lawrence@liu.se

possibilities but the output has a high learning curve and often require some post-processing to get it into a form that can be analysed. If one really wants to have full control over one's searches then one can supply SPARQL-code, something which has proved baffling. There is also little in the way of an API for SwePub for analysis and bibliometrics. In any case, the main point is that SwePub contains a huge amount of data about Swedish research which is highly underexploited.

Challenge to subject classify all research publication output manually

In recent times we have been grappling with the challenge of classifying publications via Swedish Higher Education Authority (UKÄ) and Statistic Sweden's (SCB) classification system. The system has three levels, with the middle and the detailed level sometimes referred to as the 3 and 5-digit levels respectively, corresponding to the length of the numerical codes used to describe the levels. Indications are that it could very well be required that publications are classified according to the UKÄ/SCB system at least to the middle level and to the detailed level in medicine.

Experience has shown that manual classification by authors is unreliable (the system is somewhat confusing and spending a lot of time finding the optimal subject classification is not what most authors want to do). They also have a tendency to classify themselves as researchers rather than the relevant publication. Similarly, fully manual classification by librarians is also problematic, since it takes time to read enough of a publication to understand what it is about.

Automatic classification based on the affiliation of the authors fails because of the broad nature of most university departments (it is not possible to connect a single subject area to a given affiliation). Similarly, the journal in which an article is published usually publishes in quite a few of the UKÄ/SCB subject areas (at the detailed 5-digit level). The ideal, then, is a publication-based classification but in an automated or at least semi-automated way. It was this latter approach which appealed to us and started us off looking closer at it.

Developing an automated system for subject classification based on SwePub data

For an automated system to work one needs a definition of each of the subject areas and then a system that can "understand" what a publication that is to be classified is "about." Further, it needs to match that against the subject area definitions to get the best one. As far as I know there is no explicit definition of the circa 250 detailed-level subjects (or any of the higher level ones for that matter). This explains to some degree why author classification is not always consistent since different researchers have different perceptions of the fields.

For LIUB, however, it meant that we had to start by defining the fields. This was accomplished using the data in SwePub, where there are records for publications that contain an abstract and a detailed-level subject category. While it can be expected that the data in SwePub are not perfect (i.e. some of the existing classifications are not accurate), as long as there is enough data overall, the effect of incorrect posts is not seen. Not being an expert in the query language SPARQL, I got indispensable help of systems developer Theodor Tolstoy at the National Library of Sweden (perhaps not a scalable approach, but we only need new data every couple of years) and received 250 000 publication posts with an English abstract and a detailed-level subject code and 25 000 posts with a Swedish abstract and a detailed-level subject code. English and Swedish data were processed separately to create bilingual definitions for the subject areas.

The definitions of subject areas are created by parsing the SwePub data and collecting the abstract and titles for each of the subject areas together. The number of abstracts collected for a given subject area varies from 7 300 to 1, with some 90% of the fields having over 50 abstracts, i.e. enough data to work with. For a small number of subject areas the reliability of the definition is questionable but they correspond to fields where there does not appear to be much research in Sweden.

The text mass for each subject area is then processed by removing stop words, i.e. words which are so common that they do not assist in distinguishing subject areas. As we developed the technique we built up a database of some 550 stop words (English and Swedish).

One needs then to combine different forms of the same word to a common stem, e.g. run, runs, running, ran are all related. We used a porter stemming algorithm for this. Then it is just a case of counting frequency of unique words and then sorting that from most frequent to least. The most frequent 150 words and their frequency are retained and become the definition of the subject area. As an example, the beginning part of the definition for Nursing looks like this with stemmed words and frequency in pairs:

|care|11202|patient|10986|nurs|9371|health|5412|
experi|4572|life|4061|support|3666|woman|3459|interview|3278|group|3193|particip|3069|person|3001|result|2748|relate|2660|famili|2554|analysi|2336|qualiti|2321|

The frequencies must be normalised against the total number of words connected with a given subject area.

How to use the automated system for the purpose of subject classification

An abstract for an unclassified publication can be analysed in a similar fashion to generate a keyword profile (word and frequency pairs). This profile is then compared to each of the definitions above to find the best fit, which can then be suggested as the subject area with which to classify the publication. LIUB has set the system up as a simple web page where one can upload an abstract, indicate the language (there is some attempt to guess the language which can be accepted or changed) and get back the top five matches.

Quality control of and feedback about the system for classification

The system is used fairly extensively by quite a few university libraries in their daily work with their institutional repositories. The majority use the manual web interface, but a couple use the very simple API for a more automated approach. An important question is how well it works. We have applied a grass-roots approach to testing: collecting feedback from those that use the system on a daily basis. The feedback indicates that much of the time the suggested classification is good. If one wants to be entirely conservative, then having a person check the suggestion at some point in the workflow is safest (this can even be authors who check their posts a few times a year).

Further examples of building new services using openly accessible data

Another example of using accessible data to build new services revolves around Swedish Government Official Reports (SOU). There existed no way to search all SOUs from a single site.

The Swedish government have digitized reports from 1997 to present. Reports from 1922 to 1996 have instead been digitized by the National Library of Sweden where one can browse or search titles and other basic bibliographical data. In order to make it possible to search all SOUs from a single site, LIUB indexed the full texts of all 8300 plus reports from1922 to the present, making them all full text searchable.

While searchability is a step forward, the power in accessible data is to link together different sources. The Government of Sweden also makes Government bills (regeringens propositioner) freely available as full text pdfs. In many cases, an SOU forms the background for a Government bill and by retrieving information from both datasets in real time we can link SOUs to Government bills, i.e. we cross-link different sources of information.

*

There are many more examples of new ways of using freely accessible online data sources but the underlying point for librarians is to lead the way with more sophisticated use of the information sources that we are experts in: by parsing, collating, analysing, extracting and crosslinking.