**For the research I in effect need a ‘big data’ source of research topics**

A lot of problems exist in accessing structured academic data from sources such as Google Scholar or Bing. Also, sources like the Directory of Open Access Journals, attractive at first sight tend to block when more than single browser-based queries are executed.

**The Open Archive Initiative**, supported by many academic organisations and institutions seems more accessible. The API is easy to use and appears to allow access to metadata for all the supporting institutions library contents.

Investigation of the Open Archive Initiative Data

Harvesting tools developed in Java have allowed the exploration of OAI Libraries.

**Statistics around the OAI**

OAI organisations are listed on the OAI.org page. There are 3589 organisations listed. For each organisation, there is a URL. Organisations are a mixture of Journals and academic institutions. There is no guaranteed way to distinguish them, but there are some approximations.

Initially simply performing the default ‘identify’ request on each of the URLs results in identifying that there are only 2529 of the organisations listed that have valid URLs (i.e do not time out).

Some of those valid URLs, despite returning a response, do not return a valid OAI protocol response in XML. Some simply return an HTML page and others a security challenge requiring a password.

There are around 1000 which return a valid OAI response.

**The OAI Identify Response Metadata**

This is metadata about the library. Of particular interest is the ‘Repository Name’ and the ‘Earliest Date Stamp’.

The repository names are quite restricted in value. The Nottingham one for example is:

‘Nottingham ePrints’

Doesn’t exactly give you a hint it is a university. There is sometimes more information in the free text description, but it is limited.

The earliest date field is interesting. These are quite variable, for the Nottingham example it is 2014. For Glasgow, 2018. For African Higher Education Research Online it is 1990 apparently. IT looks like information is available for at least a decade making the experiment planned viable.

The library metadata also includes usage policy for metadata and data. This will be examined before a harvest. When open the terms appear to be fairly standard (use for not for profit).

**Available Records**

Each Repository responds to metadata requests with around 500 items at a time. Each item is a paper. Nottingham eprints has around 50,000 papers. Glasgow around 200,000 papers in their library. From an initial scan there are at least 500 academic institutions (as opposed to journals). It is likely therefore that there are around 50,000,000 research paper records available.

The harvesting process takes between 2 and 24 hours per library (because if you query a library too quickly it blocks the requester. It would be possible though to execute the process in parallel. In principle I could harvest all of the libraries which are open (the 500) in around 2 weeks.

Language issues

Some institutions records are in foreign language. Chinese, Russian and some eastern European countries are the most problematic. Some libraries provide research in English in addition to the target language. It is possible to tell this from the query. Google does offer a translation service that could be used but there is enough data at present in English for the initial experimentation.

**Other Issues**

Harvesting Russian and Chinese libraries may not be that sensible a move in the current climate. There both are also problematic due to language.

**Characterising the Paper Topic Area/Innovation**

Ideally It would be helpful to filter the data based on topic area (exclude for example topics outside of natural and applied science).

THE OAI offers a wide range of papers, not just technology. The OAI Spec allows the use of sets or groups. These could provide a way to filter but are highly variable and sometimes irrelevant. For example some use groups of papers versus theses. Others departmental, e.g. Geography, and others still simply a building or location.

Explicit Keywords

Sometimes the paper does use explicit metadata to provide a topic. Terms such as ‘Radioactivity’ or ‘Graphine’.

Titles and Descriptions

Titles and descriptions can be harvested for terms. The simplest approach is to identify all individual words (excluding generic words such as a, the, it, with etc. It may also be possible to look for groups of words.

Another approach is to search text for specific word groups (e.g. ‘Kalman Filter’ or ‘Augmented Reality’. Since our search is historic this is valid for determining what correlation exists (as these terms are in the Economist, New Scientist, Gartner etc). The opposite approach (discover new word clusters) is more relevant for future prediction.

Dates

This is one area of strength. Dates are available for each paper. So if we do determine a subject we can assess its time profile either in an institution or globally. We can also look at a terms movement across organisations and regions over time.

Authors

Author information is fairly explicit. Not absolute but first/second name and initial is probably a relatively stable key. Tracking papers (associating them) between authors may also be useful. It to some degree shows how topics are related.

Geography

This is turning out to be the most challenging element. I can geo-code an institution and I am assuming (as a working assumption) an institution is likely to have papers primarily produced at that institution. This may be an invalid assumption for an institution but it is definitely invalid for journals hence the intention to use institutions initially.

Another method of establishing the geolocation of a paper might be the geographic location of the author. This is proving difficult to establish (Google has it but explicitly forbid you from executing large scale queries on scholar).