UK Company and Organisation Index for Genetic Resources and Traditional Knowledge

Paul Oldham & Stephen Hall
One World Analytics

Introduction:

This document provides a guide to the UK Companies and Organisations Index for Genetic Resources and Traditional Knowledge.

The Index is based on a review of international patent activity for genetic resources and traditional knowledge involving UK applicants between 1976 and the end of 2013. The Index consists of the following components.

1. An Excel workbook containing a series of worksheets containing different aspects of the Index.
2. An online Map showing the georeferenced locations of the companies and organisations in interactive form to identify organisations and clusters of activity. Additional search features by company name, sector and Standard Industrial Classification (SIC) code are provided. The map can be used online or can be used purely offline with the Firefox web browser.
3. A Tableau workbook with the same information for use with free Tableau Reader software available from here.

The main anticipated uses of the Index are as follows:

1. To identify clusters of UK research and development activity involving genetic resources and traditional knowledge;
2. To provide a tool for engagement with companies and organisations in debates on the implementation of the Nagoya Protocol on Access to Genetic Resources and Benefit Sharing and debates on intellectual property and genetic resources, traditional knowledge and folklore at the World Intellectual Property Organization;
3. To identify the range of economic sectors involved in research and development on genetic resources and traditional knowledge as a contribution to debates on sectoral approaches to the Nagoya Protocol;
4. To enable the identification of companies with Corporate Social Responsibility policies as a possible approach to incorporating the Nagoya Protocol into company policies.

Understanding the Index:

The index consists of all UK companies, research organisations, government agencies and hospitals that make reference to a species in an international patent publication between 1976-2013. Companies and research organisations make reference to species in patents for two main reasons:

1. Because the species is part of, or informs, the claimed invention.
2. Because the species is a target of an invention.

We include both categories of activity in the index because it can be difficult to distinguish between companies that are utilising a genetic resource and companies that are targeting

a genetic resource. We also assume that companies that target particular organisms may also use or test biological materials as part of their research and development process. As such the Nagoya Protocol could be of relevance to them. For this reason, at this stage, we assume that is better to use the index to inform the wider spectrum of companies and research organisations to whom the Nagoya Protocol could be relevant.
The practical significance of this is that users of the index will be communicating with companies whose research and development: a) utilises genetic resources; b) targets organisms, or; c) both.

Data Fields:

We initiated work with a list of over 3,020 patent applicant names from the EPO World Patent Statistical Database (PATSTAT) in the period 1976-2010. We then used a combination of computational and manual approaches to retrieve additional information about the companies and organisations focusing on the following information for inclusion in the worksheet Core Index.

1. The UK Companies House company number
2. Whether the company was listed as active, dissolved, dormant or merged
3. The Parent Company (in cases of mergers)
4. The Parent Company number (in cases of mergers)
5. The company or organisation website address
6. The address of the company
7. The main phone number
8. An email contact address (where available)
9. A Label Field for use for mail merging postal labels
10. A Reviewer Sector field based on manual review of the websites
11. The Standard Industrial Classification (SIC) codes for the company or organisation which links to UK national accounting system categories at the Office for National Statistics (ONS)
12. Georeferences for mapping companies based on Post Codes

The method originally focused on computational approaches to retrieving company and organisation information from Open Corporates (https://opencorporates.com) and later from Company Check (http://companycheck.co.uk) which aggregate information from Companies House. We also used the Company House application interface (API) where needed. This was followed by a combination of text mining of company and organisation websites and manual review. At all stages manual review was required to validate and clean the data. The data was then updated to identify new UK companies in the period 2011-2013.

1. UK Companies House Number

This is the key identifier for UK Companies and allows the company records and legal details to be accessed at Companies House. In a number of cases we were not able to identify a company number (suggesting the applicant could be using a trading name or was not formally registered). However, in most cases these were universities or non-commercial entities.

2. Company Status

The main issue encountered at this stage was establishing whether a company was active, dissolved (defunct) or merged. In practice, information on company status differed across sites. This possibly reflects data update cycles. The status of the company was therefore manually checked. The status of each company was then entered into the index manually.

One of the most striking features of the research was the number of companies listed as dissolved. We assume this partly reflects the historic dimension of the research from 1976 onwards. However, it also suggests that there may be a significant turnover of companies working with biological material. This has significant implications for implementation of the Nagoya Protocol because it raises the issue of what happens with an access and benefit-sharing agreement where a company ceases operations.

In some cases we identified companies that on further investigation proved to be foreign owned possibly as a result of a take over by a foreign company of the original UK company. Where an entry was identified for the UK subsidiary at Companies House this was included in the Index.

3. Parent Companies

In cases of mergers we identified the new Parent Company.

4. Parent Company Number

This field includes the Parent Company Number as provided by Companies House

5. *Company Websites*

For each company or organisation the company website was manually identified and checked for use in the index. Each website was searched using computational approaches to retrieve all instances of an address, telephone number and email address. Because various forms of address, numbers and emails were retrieved they were manually checked for each entry to ensure accuracy.

In some cases a company or organisation in the index lacked a company number and/or lacked a website. These are marked NA. In our experience companies that lacked a website had effectively ceased operations or had merged. However, where such companies were listed as active in the Companies House register they were retained in the Index.

6. *Company Addresses*

In the course of the research it became clear that the company address held by Companies House is the address of legal registration. In a significant number of cases this is a solicitors or accountants offices rather than the actual company offices. Addresses were therefore checked against the addresses on company websites. Because of the variety of formats of company addresses on websites there may be a need to perform additional standardisation of the company address field.

7. Company Phone Number

Companies and organisations vary in providing telephone numbers. Wherever possible we sought to identify central contact details.

8. *Company Email Addresses*

Websites can contain a number of email addresses including for general enquiries i.e. info, sales, or individuals. We sought to identify emails for general enquiries. In some cases the only available email addresses were for sales contacts or named individuals at the Company or organisation. We include these as addresses for follow up. In the case of Universities or Hospitals we attempted to identify central contact points. The same is true of government departments. These were included in the index for follow up.

9. *Label Field*

This field joins the Company Name field with the address field with the aim of facilitating mail merging for mailing labels.

10. Reviewer Sector

One of the aims of the Index is to contribute to the analysis of different sectors of UK activity that may be affected by the Nagoya Protocol or the outcomes of debates on intellectual property, genetic resources and folklore at the World Intellectual Property Organization (WIPO).

In conducting the research (see below) it became clear that Standard Industrial Classification (SIC) Codes were too ambiguous in a significant number of cases to be able to accurately identify a sector. To address this the research focused on recording a sector from the dominant terms used on the company or organisations website to provide a one word or short phrase to describe the main area of activity (i.e. pharmaceuticals or biotechnology etc.).

Because the sector review concentrated on terms used by companies it is likely that a second summary scheme and tidying of the sectors will be desirable.

11. Standard Industrial Classification Codes (SIC)

Standard Industrial Classification Codes are used by organisations such as the Office of National Statistics (ONS) as part of national reporting on economic activity. However, our research revealed that these codes are frequently lacking in sufficiently fine detail to make a judgement about the economic sector. Thus the head office of a pharmaceutical or biotechnology company may well be given the SIC code for Activities of Head Offices (code 70100) or companies may lack a SIC code (325 cases).

In practice the SIC Codes are useful. For example the top ranking codes where 72190 - Other research and experimental development on natural sciences and engineering. The second ranking code was 72110 - Research and experimental development on biotechnology. However, the information needs to be complemented by the more detailed Reviewer Sector information.

12. Geocoded Addresses

The geocodes (latitude and longitude) are provided based on the coding of the Post Code for use in Google Maps or similar software. A known issue with Google maps is that it will place multiple markers on addresses with the same post code. This can give the impression that data is missing from the map. In practice it is hidden by the top marker. To control this problem the Google Map version of the Index contains code that separates the markers when the coordinates are identical.

A known example of this issue is for UK Research Councils that are based at Polaris House in Swindon. The Google Map now correctly shows the markers for the individual research councils.

Other Worksheets:

1. 1976-2010 Review

This includes the raw data that formed the basis for the Core Index and is provided to allow for tracking back and revision.

2. 2011-2013 Review

This worksheet contains the recent update to the index for companies appearing in patent data for the first time between 2011-2013. This includes 159 entries. It is reasonable to expect that future years will produce similar new levels of activity.

3. Patent TrackBack Review

This worksheet establishes a link between the underlying patent data and the company data including the species name appearing in a patent document. This worksheet represents work in progress but is important because it establishes a link between the company number and the permanent applicant identifies in the EPO World Patent Statistical Database (PATSTAT).

The 2013 edition of the EPO World Patent Statistical Database introduced a new feature in the form of a permanent numeric person identifier (company, organisation or individual). One of the significant issues with patent data is that there are multiple spellings of the applicant name. The addition of the person identifiers is designed to help address this. At present, as will be seen in the worksheet, some organisations have more than one person identifier. However, it should be readily possible to establish a “master” identifier to which the less frequent identifiers link.

The idea therefore would be that a master identifier would in future retrieve all patent data from the company or organisation (such as newer filings) and this could then be linked to the company data. This would, if it is desirable, facilitate cost effective monitoring of activity in the patent system involving genetic resources and traditional knowledge of relevance to the Access and Benefit Sharing and the Nagoya Protocol.

It should be noted that this worksheet includes the original raw data for companies that were excluded as false positives in the course of developing the Core Index. It is therefore presently only to be used for experimental purposes.

Known Issues:

The following issues need to be considered in using the index.

1. The address field and label field were developed using a combination of computational means and manual cut and paste from websites by researchers. Because websites are written in HTML this can introduce noisy characters and spaces if pasted into Excel. The data has been through a cleaning process but will require further cleaning.
2. There are a small number of duplicate entries in the core index that need to be removed.
2. Address formats are not yet fully regularised (i.e. commas after all address terms).
3. Reviewer Sector. This list requires some further work to cluster results by sector i.e. Agriculture rather than Agriculture and Agricultural Technology or Agriculture (Dairy). This will reduce the overall number of sector descriptors from over 500 to more general sector categories.
4. In the course of the research we removed companies or organisations that proved to be false positives from patent searches on species because the names were “species like” but not actual species upon review. In the case of the Core Index it will be possible to further clean or prune the data based on investigation of the relevant sector descriptor.