Dowling Review: visualisation of data on collaborations between universities and companies

*Cristina Rosemberg, Oliver Cassagneau-Francis, Xavier Potau*

Technopolis Ltd,June 2015

# Introduction

As part of the Dowling Review, the Royal Academy of Engineering (RAEng) has been exploring ways to illustrate the extent of collaboration between research and private organisations in the UK. To this purpose RAEng requested that all universities to send information on their collaborations with industrial partners in recent years. A total of 68 universities responded to this data request.

This note summarises the steps followed to analyse that data and produce several graphic representations of the collaboration intensity between academia and industry in the UK.

# Methodology

## Building a single dataset

The first task consisted of generating a single database containing all the information on collaborations provided by universities.

The data was very varied as universities had not all followed the same structure and format for their submissions and even the type of information submitted was inconsistent. The first step was to attempt to bring some level of consistency and achieve a similar format across the submissions.

Most submissions included data on:

* Business / organisation – the industrial partner
* Duration – the duration of the collaboration
* Department – the university department involved in the collaboration
* Number of Awards – this was only relevant for some universities who did not follow the format of a single collaboration per row

There were a total of 12,240 collaborations.

## Further cleaning and panel allocation

The “structured” data was then imported into Stata (a statistical and data management software) where the process of allocating collaborations to REF panels and sub-categories was implemented. We developed a programme that, based on whether certain “tokens” were present in the name of the department, allocated the collaborations to panels. This allocation was not mutually exclusive – i.e. one collaboration may be allocated to more than one panel, if the department involved was sufficiently cross-disciplinary.

The collaborations were allocated to one (or more) of the 4 REF main panels. A second round of allocations took place to identify sub-categories within the panels. The final classification is shown in Table 1.

There were a total of 10,933 collaborations classified into at least one panel / category (89% of the total number of collaborations).

Table : REF main panels and further disaggregation sub-categories

|  |  |
| --- | --- |
| REF Panels | Sub-categories |
| Panel A – Life Sciences | Clinical Medicine |
| Agriculture, Veterinary and Food Sciences |
| Health Professions and Services |
| Biological Sciences |
| Panel B – Engineering and Physical Sciences | Engineering |
| Chemistry |
| Physics |
| Mathematics |
| Earth and Environmental Sciences |
| Computer Science |
| Panel C – Social Sciences | Architecture and the Built Environment |
| Archaeology, Anthropology and Geography |
| Law |
| Education |
| Economics and Business |
| Sociology and Political Sciences |
| Sport and Exercise Studies |
| Miscellaneous / Combined |
| Panel D – Arts and Humanities | English, Area Studies, Modern Languages and Linguistics |
| History and Classics |
| Philosophy, Theology and Religious Studies |
| Art, Design and Performing Arts |
| Communication, Cultural and Media Studies, Library and Information Management |

Source: REF 2014 (main panels and basis for sub-categories)

## Cleaning of organisation names and analysis

### Fuzzy matching

In several cases, universities used slightly different names (and spellings) for the same industrial partners and a cleaning / standardisation process was followed in order to be able to estimate collaborations across individual industrial partners and identify the most important collaborators.

To this purpose we have used a tool called ‘Fuzzy matching’ (which is an add-in function in Excel). Fuzzy matching describes the process of matching similar strings (company names in this case) and is a valuable tool in data management.

Using this feature, we identified groups of similar company names, and could then make a decision about whether these did refer to the same entity or not. Then, we replaced the same entity with the same name. This has allowed us to “collapse” the data by organisation name, and count the number of collaborations involving each organisation.

### Counting collaborations per organisation (company)

Once we had cleaned the majority of names (focusing on the more important in terms of number of collaborations first), we used another Excel add-in, PowerPivot, which allows “distinct counts”. This meant we are not only able to count the number of collaborations that each university was involved in, but also the number of different universities with which each organisation collaborated (i.e. a distinct count). We were able to conduct this analysis at a number of levels, thanks to the allocation of collaborations to panels and sub-categories. In this way we can identify the most active collaborator companies, both overall and for each of the sub-categories and panels.

At this stage, we needed to identify which organisations were companies, and which were perhaps funders or other types of collaborator which we did not wish to include in our analysis. We “cleaned” the data in this fashion (marking as companies or other types of organisation) down to all organisations that were mentioned by at least two universities following fuzzy matching.

A total of 377 companies were named by more than one university. This represents 5% of the total number of organisations and covers 25% of the total number of collaborations. This means that there is a ‘long tail’ of companies, and this tail accounts for 75% of the total number of collaborations. Note that the existence of this ‘long tail’ should not affect the relevance the visualisations as they not affect the information on top collaborators (see Section 2.4.2) and does not affect the information on number of collaborations (which include the ‘long tail’).

The majority (377 of 415 organisations, 91%) of these were companies, suggesting that it is reasonable to include the “long tail” of organisations (which we have not verified to be companies) in other analysis, as the majority are likely to be relevant (though of course companies may be more likely to appear in the high-collaborator group).

Below we present a summary of data contained in each of the visualisations:

* The top 50 companies account for 13.3% of all collaborations (1,622 out of all 12,240 collaborations)
* 377 companies (4.9% of all 7,551 organisations) are mentioned by more than one university, together these account for 24.4% of all collaborations (2,989 of 12,240 collaborations).
* 38 organisations were mentioned by more than one university but were excluded from the analysis at company level since they are not companies
* 146 collaborations did not list any collaborator (field was left blank)
* 7,135 organisations were involved in collaborations with a single university, covering some 8,878 collaborations (72.5%)
* 6,437 organisations were only involved in a single collaboration (52.6%)

We did not clean names in the “tail” (organisations named by a single university). This means that analysis across all collaborations (section 2.5.3) will include organisations that are not companies. Based on the information presented in bullets 2 and 3 we can roughly estimate that 10% of all collaborations refer to organisations that are not companies.

## Creation of treemaps

### Treemaps

The nature of this data suggested that treemaps were a natural choice to visualise this information. We produced a number of treemaps using a “package” in the open source / open licence statistical software, R[[1]](#footnote-2).

### Top 50 companies

In these treemaps we focus on presenting the companies, which had collaborated the most, using our list of the 377 companies identified who had collaborated with at least 2 of the 68 institutions who responded to the RAEng request (making up ~25% of total collaborations). These treemaps consisted of an overall top ~50 collaborator companies, and a top ~50 for each of the 4 REF main panels.

Please note that there was in each panel and overall, a tie around the 50th position, e.g. more than one company had the same number of collaborations for position 49th to position 52nd. In this case the graph will include the top 53 companies instead of top 50. Also note that companies are included in the graph only if they have been named by at least two universities.

In these treemaps the size of the rectangle is proportional to the total number of collaborations recorded for that company, and the colour / shade reflects the number of universities with which it collaborated (distinct count). These treemaps focus exclusively on companies, and the tail does not affect these treemaps[[2]](#footnote-3).

Figure Top 49 companies mentioned in all collaborations

|  |
| --- |

Note: includes 1,643 of 10,933 collaborations allocated to a panel.

Figure 2 Top 55 companies mentioned in collaborations in Life Sciences

|  |
| --- |

Note: includes 761 of 3,321 collaborations allocated to Panel A.

Figure 3 Top 48 companies mentioned in collaborations in Engineering & Physical Sciences

|  |
| --- |

Note: includes 916 of 6,276 collaborations allocated to Panel B.

Figure 4 Top 43 companies mentioned in collaborations in Social Sciences

|  |
| --- |

Note: includes 152 of 1,634 collaborations allocated to Panel C.

Figure 5 All 33 companies[[3]](#footnote-4) mentioned in collaborations in Arts & Humanities

|  |
| --- |

Note: includes 45 of 368 collaborations allocated to Panel D.

### Number of collaborations: overview across sub-categories and panels

In these treemaps we focus on showing the number of collaborations in each panel, and also further disaggregated into sub-categories within these panels. To create these treemaps we used all the collaborations that were allocated to a panel (10,933 ~90%) – not just those that we had identified as involving companies. Though we cannot make any real predictions about the make-up of this group (it may not reflect the major collaborators for a number of reasons), the analysis is still worthwhile as it shows which fields are collaborating the most (using the universities’ own individual definitions of a “collaboration”).

Figure All reported collaborations by sub-category

|  |
| --- |

Note: includes all 10,933 collaborations allocated to a panel (89% of total).

Figure All reported collaborations by REF main panel

|  |
| --- |

Note: includes all 10,933 collaborations allocated to a panel (89% of total).

1. The package, called “treemap”, is the work of Martijn Tennekes, a software developer in the Netherlands. It allows a number of options for changing colours, fonts and other variables that determine the look of the treemap, and then allows export of the image in high quality. [↑](#footnote-ref-2)
2. Here we are assuming that our fuzzy-matching was successful and picked up the same companies that may have been spelled differently – if this is not the case we may have missed some “important” companies, ie companies that collaborated with 2 or more universities. [↑](#footnote-ref-3)
3. From the population of 377 ‘cleaned’ companies that collaborated with at least 2 institutions. [↑](#footnote-ref-4)