Dowling Review: visualisation of data from the REF case studies

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

Technopolis Ltd,May 2015

# Context

As part of the Dowling Review, the Royal Academy of Engineering (RAEng) has been exploring ways to illustrate the collaboration between research and private organisations in the UK. Recent work commissioned by HEFCE to Digital Science resulted in a searchable database of 6,642 impact case studies submitted in the latest REF exercise. The analysis carried out by Digital Science and the Policy Institute at King’s College was the first to use the whole body of REF impact case studies. In parallel, Technopolis piloted a new approach to obtain company mentions within unstructured text information as part of the study for the EPSRC and the RAEng, “Assessing the economic returns of engineering research and postgraduate training in the UK”. The study, which was published in March 2015, included a word cloud of the most mentioned 50 organisations in 514 engineering impact case studies from the same REF data source.

In the following work, we have extended that same analysis to all the REF impact case studies that have been published. The objective is to arrive at a final visualisation that gives a sense of the major companies and industrial players that are the main targets and beneficiaries of the impact of research undertaken in the UK.

# Methodology

In order to meet the stated objectives, we followed four steps, which are explained in detail in the subsections below.

## Systematising the REF case studies into a structured database

The full dataset of the published REF case studies is available from the HEFCE website[[1]](#footnote-2). A total of 6,975 case studies were submitted to the REF, and the public database contains information on 6,642 of them (95.2%). Some case studies were substantially redacted or marked as confidential as part of the REF process and as a result they have not been made available by the submitting institutions. This dataset can be downloaded in a variety of different formats and, for the purpose of this analysis we have used the JSON format.[[2]](#footnote-3)

For each of the available case studies, we extracted the following information:

* Case Study ID
* Unit of assessment
* Details of the impact

The units of assessment were linked to their respective REF Main Panels using the information provided on the REF website[[3]](#footnote-4).

## Identification of organisation names

The aim of the analysis was to identify organisation names (and more specifically companies) that are mentioned in the section ‘details of the impact’ of the REF case studies. The information contained in this section of the 6,642 case studies amounts to the equivalent of around 10,200 full pages of written text. As a result, one needs to use automated machine-reading techniques to sift through this amount of information using a reasonable amount of time and resources. In order to accomplish this, we have applied a state-of-the-art Name Entity Recognition (NER) engine to the texts. NER algorithms take text and tag it with specific metadata, locating and classifying expressions in different sets of predefined categories or entities. Some of these algorithms are pre-trained with large volumes of textual data and use advanced techniques to detect names in the text, and to classify them by type of entity[[4]](#footnote-5).

For this analysis we used the latest version of the Name Entity Recognition (NER) classifier published by the Stanford Natural Language Processing research group[[5]](#footnote-6). Using a 3-class model for English language, the classifier tags organisations, locations and person names present in a text. For example, the text

“Working with Synopsis Inc he developed interfaces for the Glasgow simulation tools. Asenov and Millar joined Gold Standard Simulations (GSS) as CEO and COO when it was spun out of the University of Glasgow”

is processed and annotated as follows,

Working with [Synopsis Inc]ORGANISATION he developed interfaces for the [Glasgow]LOCATION simulation tools. [Asenov]PERSON and [Millar]PERSON joined [Gold Standard Simulations]ORGANISATION (GSS) as CEO and COO when it was spun out of the [University of Glasgow]ORGANISATION

Our analysis focuses on organisation names only. While this type of analysis is quite experimental in nature, it has some interesting advantages. The main one is that we do not interrogate the text with a pre-assumed or expected list of companies and count the amount of mentions, but we let the data tell us who is mentioned in the text and to what extent (number of cases). This bottom-up text analysis (with a lack of predefined search targets) allows us to just see what emerges from the data, and we avoid missing unanticipated and important results. Of the 6,642 case studies analysed by the NER algorithm, 6,626 have at least a mention of an entity recognised as an organisation in its ‘details of the impact’ section, accounting for 99.8% of analysed REF case studies and 95% of submitted case studies.

## Data cleaning and post-processing

A total of 135,904 entities were identified by the NER algorithm as organisations, of which 45,582 are unique entities. This included spelling variations, acronyms and false positives, amongst others. We further cleaned the results using a two-step approach:

* In the first step, we used fuzzy matching techniques[[6]](#footnote-7) to group entities based on a similarity measure given by the fuzzy matching algorithm (candidate terms to be cleaned). We then carried out this cleaning step manually for any term initially mentioned in more than 5 case studies.
* In the second step, we took a list of 700 organisations by amount of case studies that mention them and we further cleaned the list manually, homogenising any names not initially detected by the fuzzy matching algorithm.

We tagged the organisations in this final clean list as ‘companies’ or ‘other organisations’, obtaining a final list of ‘companies’ by order of case studies that mention them. Note that Universities, Government Departments and other government organisations and networks were tagged as ‘other organisations’ and consequently omitted from our analysis.

Some media channels such as the BBC, Google and YouTube featured as company names mentioned frequently in the case studies. However, a quick scan through some REF case studies confirm that those organisations are named by the authors to show how their research was portrait in the media, rather than to explain how their research contributed to these companies. Consequently, names belonging to communication channels have been removed from the final results, as they do not match the original intent of the analysis.

Throughout the tagging and cleaning processes, we maintained the link to the case study ID for each record, so that we can generate lists of companies for the overall database as well as for the four REF Main Panels.

## Scoring and visualisation

Company mentions are scored as follows:

*A company is awarded one point for each mention in a case study impact descriptor (but not more if it is repeated several times within the same case study).*

This gives us information on which companies are mentioned in which case studies. Once we arrived at a clean dataset of companies mentioned in each of the case studies, we visualised the data using word clouds. **The word clouds depict the organisations mentioned in 5 or more REF Case studies, for the overall cloud as well as for the thematic ones.** This means that the word clouds for some panels are more populated than for others. The information on companies and frequencies have been provided in a separate file.

# Results

Figures 1 to 5 show the resulting word clouds. These figures are also provided at higher resolution for use in other reports.

Figure 1 Organisations mentioned in more than 5 impact REF case studies

|  |
| --- |

Figure 2 Organisations mentioned in more than 5 impact REF case studies of Panel A

|  |
| --- |

Figure 3 Organisations mentioned in more than 5 impact REF case studies of Panel B

|  |
| --- |

Figure 4 Organisations mentioned in more than 5 impact REF case studies of Panel C

|  |
| --- |

Figure 5 Organisations mentioned in more than 5 impact REF case studies of Panel D

|  |
| --- |

1. http://impact.ref.ac.uk/CaseStudies/About.aspx [↑](#footnote-ref-2)
2. http://impact.ref.ac.uk/CaseStudies/APIhelp.aspx [↑](#footnote-ref-3)
3. http://www.ref.ac.uk/panels/unitsofassessment/ [↑](#footnote-ref-4)
4. More information on different types of NER algorithms can be obtained in: Nadeau, David, and Satoshi Sekine. "A survey of named entity recognition and classification." Lingvisticae Investigationes 30.1 (2007): 3-26. [↑](#footnote-ref-5)
5. http://nlp.stanford.edu/index.shtml [↑](#footnote-ref-6)
6. See: http://research.microsoft.com/en-us/projects/datacleaning/ [↑](#footnote-ref-7)