Clustering of EPSRC research topics and researchers: using a network analysis approach based on grant data
- Research Data Management Plan -

Sergiu Tripon
Department of Computer Science
University College London
Gower Street, London, WC1E
sergiu.tripon.15@ucl.ac.uk

1 Data Collection

1.1 What data will you collect or create?

The data used in this project will be collected entirely from EPSRCâĂŹs Grants on the Web (GoW) service. The service comprises of a collection of web pages storing EPSRCâĂŹs grant data categorised by a number of different entities: research areas and topics, industrial sector, scheme, socio-economic scheme, themes, region and so on. It also provides grant data from the past 30 years.

Both past and current data will be collected:

- Current data will be collected through the Grant Portfolio page;
- Past data will be collected through the Past Grant Portfolio page;

The Grant Portfolio and Past Grant Portfolio pages are very similar, with one major difference. The Past Grant Portfolio page provides the ability to select a start or end date or a start and end date, supplying grant data according to the period selected. This project uses the Past Grant Portfolio page to collect grant data between two periods: 1990-2000 and 2000-2010.

In contrast, the Grant Portfolio page follows a tree-like layout which presents the grants categorised by different entities such as research areas and topics, industrial sector, scheme etc. This Project uses the Grant Portfolio page to collect current grant data.

Each grant page consists of the researchers working on that project and the GoW service provides links to each individual researcher page. This project will also collect researcher data from each current and past grant.

Using the collected data from EPSRC, a number of topic and researcher-based networks will be created in order to be analysed and visualised. Files storing the analysis and visualisation of the networks will also be created.

I anticipate that the data produced will amount to approximately **360MB**.

1.2 How will the data be collected or created?

The data is provided in the form of web content on a web page. This means that a web scraper will have to be developed in order to scrape the data from the EPSRCâĂŹs GoW service. Usually, a web scraper is a script which extracts data that lies under various HTML tags on a web page. The web scraper will be developed in the Python programming language.

The collected data will be formatted into a suitable network format and a number of networks will be created using Gephi, a graph visualisation software. Other derivations of the networks will also be created programmatically in Python, using the Jetbrains PyCharm IDE. The networks will initially be stored as .tsv files and as the analysis progresses they will also be stored as .gephi and .graphml files. The statistics of each network will be stored as .txt files, while the visualisation of each network will be stored as .png files.

2 Documentation and Metadata

2.1 What documentation and metadata will accompany the data?

The data will be accompanied by the following contextual documentation:

- A GitHub Wiki documenting the entire project including the collected data;
- .txt files detailing the statistics of each network;
- .png images showcasing the process undertaken in creating the networks;

Files and folders will be named according to a pre-agreed naming convention. The final dataset as stored on GitHub will also be accompanied by a README file listing the contents of the other files and outlining the file-naming convention used.

3 Ethics and Legal Compliance

3.1 How will you manage any ethical issues?

There are no ethical issues regarding the generation of results in this project. There is no human subject or samples involved.

3.2 How will you manage copyright and Intellectual Property Rights (IPR) issues?

The research is not expected to lead to patents. IPR issues will be dealt with in line with the policy of University College London.

4 Storage and Backup

4.1 How will the data be stored and backed up during the research?

Storage and back up will occur in two places:

- On the laptop of the researcher;
- In a private GitHub repository owned by the researcher, which will be made public once the project is considered complete;

4.2 How will you manage access and security?

The laptop will be password protected. The risk is that the laptop will be hacked. The laptop has anti-virus software installed which is updated daily.

GitHubâĂŹs security information can be found at the following web address: `https://help.github.com/articles/github-security/`.

5 Selection and Preservation

5.1 Which data are of long-term value and should be retained, shared, and/or preserved?

All data from this research will be considered to be of long-term value and will be retained and preserved.

5.2 What is the long-term preservation plan for the dataset?

The data from this research will be stored in a private repository on GitHub owned by the researcher, which will be made public once the project is considered complete.

6 Data Sharing

6.1 How will you share the data?

The data will be shared with University College London and the supervisor of this research project when it will be submitted for marking. Additionally, once the project is considered complete, the entire project including the data collected will be made available within a public GitHub repository. Other users will be able to download and fork the projectâĂŹs source code and data and potentially contribute to it.

6.2 Are any restrictions on data sharing required?

There will be no restrictions on data sharing.

7 Responsibilities and Resources

7.1 Who will be responsible for data management?

The researcher will be responsible for:

- data capture;
- data management;
- metadata production;
- storage and backup;
- data archiving;
- data sharing;

7.2 What resources will you require to deliver your plan?

The researcher already has the required software to implement the data collection plan.