British Library PhD placement scheme – project profile:

# Data Mining of Doctoral Theses

Reference: 2017-6-GOU

| | |
|---|---|
| **Supervisor** | Sara Gould, Development Manager for Researcher Services |
| **Department, Location** | Either at **St Pancras**, London or at **Boston Spa**, Yorkshire |
| **Start date/timeframe** | Anytime in the period May 2017-May 2018 |
| **Duration** | **3 months** (or part-time equivalent – see below)<br>*For a part-time placement, the student would be required to spend a minimum of 3 days/week on site at the Library* |
| **Remote-working** | This project does not offer scope for remote working – the placement must be conducted on site at the British Library. |

## Context for placement

The British Library manages the national database of UK doctoral theses, called EThOS (http://ethos.bl.uk). We work with UK universities to aggregate information about all theses produced by PhD students in the UK, and provide full access to as many as possible to support researchers everywhere.

The main purpose of EThOS is to support researchers to search for, discover and access theses for use in their own research. Where possible, users can download the full thesis or order a scan of older print theses, and EThOS is one of the British Library's most well used research resources.

But the almost complete aggregation of the metadata records of all UK PhD theses – some 450,000 – also has enormous value. For example, users can compare one university's outputs against another, make connections between thesis authors and their supervisors, or analyse trends in funding for doctoral research. The value of this large dataset is increasingly recognised, and we now want to extend and improve the data held within it as much as we possibly can. To analyse funding trends, we need consistent, comprehensive funding body information; to make connections between thesis authors and their supervisors, we need supervisor names to be present in the metadata records; to analyse trends in the research itself, we need at least the thesis abstract so that people can mine and re-use the information.

This placement project focuses on three metadata elements – supervisor names, research funding organisations, and the thesis abstracts. Very often, these data elements are described within the 'front matter' pages of the full theses themselves but not (yet) in the re-usable metadata records. The aim of the project is to use content mining methods to extract the missing data from the full theses.

## Expected tasks and outcomes

*This is an experimental project and we do not know how easy it will be to extract the required data. One aim of the project is to highlight the potential value of the EThOS dataset; a project that tries to extract re-usable data but fails for valid reasons will not be deemed a waste of time so long as the Library and future data miners learn from the work, and we are able to communicate the results to interested stakeholders in future*.

The text from the front pages of c. 150,000 theses will be made available by the Library for this project, having been processed through digital OCR (Optical Character Recognition) software. A typical thesis will have been scanned from the original print copy or produced as an e-thesis in the first place. The thesis abstract will be contained within the front pages under the heading of Abstract; very often supervisors and (less often) funding bodies are named within the Acknowledgements pages of the thesis.

The placement student will be asked to use Text & Data Mining (TDM) techniques to mine and extract the three metadata elements mentioned above from the full theses so that we can add them to our existing EThOS records. They will need to create or maintain a link between the existing EThOS metadata record, the full thesis and the extracted data using the EThOS ID identifier, so the new data can be re-used and matched with existing content. They will also analyse the extracted data to see if it is accurate enough to re-use in EThOS.

The student will write an end-of-project report for the Library to describe the project's results, successes, challenges and give recommendations for future similar activities. They will also deliver a presentation to British Library staff to describe the results of the project, as well as presenting project findings at a meeting of the EThOS Advisory Board.

The data extracted (if good enough) will be used first to enhance the EThOS records, but could also form the basis of further research projects. Depending on their interests and the time available, the placement student will be encouraged to liaise with stakeholders to explore opportunities to undertake preliminary research drawing on the extracted data – for example an analysis of the funder information on behalf of research bodies.

This placement could be undertaken at either the Library's Boston Spa site in Yorkshire at St Pancras, London. We would also be happy for the student to spend some time at both sites.

### Training and experience expected to be gained by student through the placement

This is an opportunity for a PhD student to apply their TDM skills and learning to meet a real service requirement – and in so doing to gain first-hand experience of an important area of the British Library's operations. There will be opportunities to shadow Library colleagues working in similar areas, e.g. BL Labs, Digital Scholarship and Metadata Services.

During the placement, the student will have access to all appropriate staff training and public events offered by the Library. This may include, for example, events in the Library's 'Digital Scholarship' and '21st Century Curators' staff training programmes, and events organised at the Library by the Alan Turing Institute.

The student will be able to take ownership of the project from start to finish. The Library has pockets of TDM knowledge but the student is likely to be the expert in the use of TDM tools and applying them to this project.

Delivering internal staff talks and presenting to the EThOS Advisory Group will enable the student to develop their communications skills and expertise in engagement with different audiences and stakeholders. The placement is also likely to bring additional opportunities for conference presentations, internal and external talks, and written communications e.g. blogs, especially once the project is completed, to share results.

### Required knowledge and skills

It is **essential** that the placement student has experience in, and ability to use, TDM tools in order to extract pieces of text from digital papers, e.g. programming/text processing (NLP/OCR) experience. The student will need to have the ability to select the best TDM approach to get the best results, as well as the ability to communicate the results of the project to British Library colleagues in accessible, non-technical language.