Special Collections & Galleries, University of Leeds Libraries, Tim Procter, T.J.Procter@leeds.ac.uk

Within Special Collections & Galleries' catalogues, there are instances of historically offensive terminology that reflect historical prejudices. These often exist in legacy descriptions imported from old systems, or in data fields such as titles. A good example is the Liddle Collection of World War 1 and World War 2, where soldiers' letters and diaries contain some instances of outmoded or outright racist language about Asian and African soldiers, as well as the nations being fought against. Some of this language was included in old index summaries. Amongst the 1.5 million or so records, instances of outright offensive language are proportionally few, and therein lies a problem – actually identifying them to deal with.

We applied to The National Archives Testbed fund, and worked with the Leeds Arts & Humanities Research Institute (LAHRI) to fund a trial of applying corpus linguistic analysis software to the catalogues to identify instances of offensive language.

The trial has given us a proof of concept that this software, developed for academic research, can be applied to catalogue data. Through LAHRI we engaged two PhD researchers, Kevin Jones and Vic Clarke, who had experience of archival research and of applying linguistic software to large data sets. Kevin and Vic also gave pointers about thesauruses and terms lists which can help build searches, for example OfCom's offensive terms list for the broadcasting industries. They were also able to link our project with other archives sector studies and initiatives, and the project was informed by Caroline Bolton's RLUK / TNA Fellowship. It also fed into the development of the Cultural Collections Sensitivity Policy. The project has given us a good idea not just of the technical processes involved in running analytical software across catalogue data, but also the resources required. The next step is to resource the actual application.

The sheer scale of records is a challenge, so the trial focused on a few selected collections. The software cannot be run within the EMu Collections Management System, so the data had to be extracted into spreadsheets. Therefore there was a challenge in identifying where in a record (which specific data fields) offensive language might be found, as to extract all fields would create an overwhelmingly complex spreadsheet. There was also the challenge of exactly what we put into the software to search for, and this is where previous work in the sector, coupled with Caroline, Kevin and Vic's research into archive-adjacent practice such as that of OfCom, as well as established archives such as the Ahmed Iqbal Race Relations Centre, was invaluable. Perhaps the first challenge was to find people with the right experience of archive research and linguistic analysis, and this is where our partnership with LAHRI proved invaluable.

Kevin and Vic produced various qualitative and quantitative outputs from their work, showing what the software had indentified in the various catalogue samples, and showing how it could be expanded out. They also did an open workshop demonstrating their findings, which was attended by representatives from TNA's Archives Sector Development Scheme, who funded the project, archivists from across the sector as well as Leeds staff, academics from LAHRI and other PhD researchers.

The idea of TNA's Testbed is that it funds projects that are testing ideas that might have potential benefit across the archives and wider heritage sector. Use of analytical software on catalogue data is of potential benefit to any repository with historic catalogue data - archive, museum, gallery, library. Some of the packages explored by the project are open source, so can be used by anyone. Dissemination was also a mandatory part of the funding, hence the open invitation workshop, and the project has also featured on TNA's website and in the CILIP Newsletter.

Revised 2nd March 2023