A lecturer in the Faculty of Arts (Language, Literature, and Linguistics) approached his Subject Librarian in March 2018 for assistance with a class assessment task for a Semester 2, 2018 course “ATS2672 Exploring texts with computers” which involved text and data mining of, at the time, undetermined data sets. (See Appendix 1 for Assessment Task details.) The tasks were to take place in September/October. The Subject Librarian approached the Collection Development Librarian (CDL) to seek assistance with: a) determining what sources the lecturer could mine, b) what sources the lecturer wanted to mine based on the assessment tasks, and c) how to acquire the data.

The lecturer is a leader in the Digital Humanities area and is well-versed in applying computational methods to datasets.

Findings:

After a few meetings with the Subject Librarians, Collection Development Librarian, and Lecturer, the Lecturer settled on the assessment tasks he would like his students to complete in Semester 2, 2018. The Lecturer had determined three assessment tasks, and the datasets he either had already, or would like to acquire. For these assessments:

1. The Lecturer already knew he wanted to use the Austlit “out of copyright” data set for one of three assessment tasks, and he already had this dataset. The CDL offered to see if it was possible to acquire the Austlit “in copyright” corpus, which the lecturer eagerly accepted.

After contacting The University of Queensland regarding the dataset of the in copyright corpus, and after discussing the conditions under which the material was to be used and that the proper acknowledgement would be made, they sent the dataset of 1,231 XML files within three weeks. The acknowledgement required was to state on the course material that the corpus was provided by AustLit at The University of Queensland and to add the homepage link: www.austlit.edu.au.

2. The Lecturer wanted to see if the Library could obtain the datasets of the 18th and 19th Century UK Parliamentary Papers, which the Library had acquired previously (on the platform) from Proquest.

The CDL contacted Proquest to inquire about obtaining the dataset for the 18th and 19th Century UK Parliamentary Papers. The cost was USD 1,250 and the initial estimate of shipping the data on a hard drive was six months. Due to time constraints, CDL asked if the shipment could occur earlier and was told no. The lecturer was disappointed but said he would construct a Plan B. However, after a
week, CDL was told by Proquest that the data was arriving the following week and the lecturer was thrilled. Because this was a pilot test, the Library agreed to pay for the data set.

3. For the third assessment task, the Lecturer again already had the dataset that he wanted to use, so no follow up from the Library was necessary. The dataset was from the ESRC Centre for Corpus Approaches to Social Science (CASS).

The Library’s mission is to make resources available as widely as possible, and the CDL ascertained at the pre-acquisition stage that this particular dataset could be made available to the entire Monash University community (with the usual authentication restrictions applied), so that the reuse of the data is viable.

An unknown at the pre-acquisition stage was how to store the data, so this was discussed between the CDL, Applications Librarian, and Systems team post-acquisition. Currently, this is a case by case assessment based on whether or not the data can be used by the entire Monash University community, and if there are other restrictions including size and format of data.

Discussion:

The conversations between the Subject Librarian (SL), Collection Development Librarian (CDL), and lecturer were “chicken and egg” at first, as the lecturer was unsure of which datasets he wanted to obtain for his class, and the CDL couldn’t present a clean list of specific resources to pick and choose from for text and data mining. This had to do largely with the notion that some whole resources (such as Web of Science, Scopus, The New York Times) are single entities with clear use permissions across the corpus. A majority of the primary source humanities databases (e.g. Adam Matthew, Gale, and Proquest products) are made up of individual collections, to which there are often rights restrictions around certain titles in a collection, making it difficult to purchase for text and data mining use.

After the CDL acquire the in-copyright content of the Austlit data, the lecturer mentioned that it would improve their prospective ARC grant application in the future.

Conclusions:

The lecturer went through the appropriate steps (i.e. contacting his Subject Librarian first) to try and ascertain the process and restrictions around data mining. The Collection Development Librarian (CDL) did have to work with the lecturer to understand specific content to be acquired and then proceeded to speak to the vendors/proprietors to ascertain licensing conditions, process, and cost.

The question of reuse will continue to be built into the pre-acquisition stage as well as questions around storage and format of data. Lastly, the question of who will pay for the data will have to be assessed on a case by case basis for now.
Appendix 1: Unit Guide for ATS2672, Exploring texts with computers

Assessment task title: Assignment 3 - Researching texts

Due date: 29/10

Details of task: The task is to research a question by examining a substantial body of text. Three suggestions are given below but you can also research a question you devise yourself after consulting the teaching staff. In writing your report, how you approached the question is more important than the answer you arrive at (if there is one). As well as a short account of your findings, you should give an account of the methods which you used, how they complemented each other (or did not) and how effective they were in trying to answer the question. An important question to keep in mind is at what point does interpretive reading by you become an essential part of the research process i.e. at what point do the limitations of automated text analysis become an impediment?

1. (TBA subject to library access to collections)
   What names are used to refer to what we now call Australia in UK policy discussions between 1770 and 1820? Did usage change over time and did it vary between e.g. government documents and parliamentary debates?

   Sources: UK Hansard, UK Government records.

2. Is there variation in language use in Australian newspapers depending on the state of publication?

   Sources: The corpus ‘Australia 2015/2016’ includes all articles from major Australian newspapers published from August 2015 to July 2016 that include the key term ‘Australia’ or ‘Australian(s)’ in the title. A version of this corpus will be made available on Moodle which has the material grouped by publication.

3. What kinds of differences in language are associated with difference in gender in Australian literature before 1935?

   Sources: The open part of the AustLit corpus is a collection of texts which are now out of copyright (i.e. published more than 70 years ago). The collection will be available on Moodle.

Word limit: 2000 words