# Hansard Data Mining Project - Decisions Made

|  |  |  |
| --- | --- | --- |
| **Date** | **Decision** | **Reason** |
| 28/08/2019 | Not scraping requested Committee Information Hansard records | Records are in a different format (PDF) which means that additional custom code will need to be written to scrape and store the record in the appropriate format for analysis. This would take additional time that is not available in this project with its short timeframe. |
| 9/09/2019 | Scraping two years of data rather than one | Was ahead of schedule for scraping data and therefore had time to scrape additional records (Bills and Answers to Questions). Initially only one year of Hansard records were scraped. |
| 11/09/2019 | Document decision made and the reason to be handed up with deliverables | Useful information for AGD to identify why the project includes and doesn’t include specific functionality |
| 12/09/2019 | Scraping last two years of Question Time records | Received feedback that Question Time records are more useful to audit teams than Answers to Questions records |
| 2/09/2019 | Not using Azure | AGD do not want to use Azure due to the complications with current processes and policies. |
| 7/10/2019 | Using regular expressions in Tableau to allow multiple term and wildcard search | This appeared was the easiest way to include more complicated string search in Tableau as the software is not designed for text search. The project team will document some common regular expressions in a user guide. |
| 7/10/2019 | Not all text analytics being investigated and implemented will be included in the final dashboards given to AGD. Instead they will be submitted and documented as potential future work for the project. | Time constraints mean that not all text analytics could be properly investigated, implemented and put in a format that can be integrated into the Tableau dashboard. Therefore, work not integrated into the dashboards will be documented appropriately in the final report as potential future work for the project. |
| 7/10/2019 | User guide included in deliverables | A short user guide will be produced for the dashboards to better explain some of the more complex functionality such as the use of regular expressions in text search. |
| 11/10/2019 | Will not be including Gazette information in the project | This was identified as being another useful data source to be analysed. However, given time constraints it is not possible to include this additional data source. This potential data source will be documented as future work for the project. |
| 12/10/2019 | Will not use names of only one-word length when searching for clients in Hansard records | Search terms consisting of only one word, in the Client names spreadsheet, are too broad and would return too many irrelevant records. |
| 16/10/2019 | Subject TreeMap in Subject Summary Dashboard will only drill-down two levels of information. | Allowing the TreeMap to drill-down more than two levels results in an unreadable chart because the text is too small. |