Many government documents contain sensitive information that must be identified and protected before the documents can be released to the public. While manually reviewing such documents for sensitive information it can be important to determine contextual information about specific entities that are mentioned in the documents and whether the information that is discussed about these entities is already in the public domain. In this project, you will develop a system that can automatically identify external information about specific entities from publicly available knowledge graphs (e.g. Wikidata or DBpedia). The system should be able to assist human sensitivity reviewers by identifying entities that are referenced by different names in the collection (based on the entity's attributes) and whether personal information about named entities is in the public domain.

You will work with named entity recognition tools (e.g. spacy https://spacy.io/) along with entity linking tool such as ReFinED (https://github.com/amazon-research/ReFinED) or DBpedia Spotlight (https://www.dbpedia.org/resources/spotlight/). A graph databases such as Neo4j (https://neo4j.com/) will likely also be used to dynamically build a definitive view of the entities within the document collection.

- **Summary of what was agreed last week**
  - Classification model to judge entity sensitivity

- **Progress made in the past week**
  - Referred back to my previous classification models from Text as Data
  - Attempted to implement model
  - Realised text processing was too time consuming and will use existing information in database
  - Needed to complete processing of all documents in SQL DB to continue

- **Main questions for discussion**
  - LDA topic modelling for corpus analytics page?
  - No major questions, just stuck at a current chokepoint of processing documents

- **Feedback from meeting**
  -