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**
  - Experiment with Neo4j
  - Continue topic modelling development
  - Continue improving web app look

- **Progress made in the past week**
  - Unsure I could implement Neo4j into web app with current timeframe
  - Fixed DBPedia-Spacy API issue by using local version and hosting server
  - Improved site appearance
  - Changed topic modelling to entities and linked sensitivity
  - Created corpus analytics page and displayed topic modelling

- **Main questions for discussion**
  - What uses can be implemented for topic modelling? Most similar documents using cosine difference?
  - Additional corpus analytics features to complement topic modelling
  -

- **Feedback from meeting**
  -