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 DBpedia spotlight
  - Experiment with Spacy entity linker
  - Read the links that have been shared
  - Come up with some ideas based on reading

- **Progress made in the past week**
  - Set up Spacy entity linker & DBPedia spotlight examples in Jupyter notebooks
  - Set up basic PDF file reading in
  - Unsuccessful in setting up neuralcoref
  - Read chapter three of thesis
  - Currently struggling to comprehend the *exact* aim of the project

- **Main questions for discussion**
  - What is the main project aim?
    - What would an ideal implementation do?
  - How to identify attributes that link to entities?
    - Confident in recognising entities, unsure how to link information to the entities
  - Methods of determining if attributes are in public domain
    - Web scraping DBpedia URLs of recognised entities?