**Entity and Relationship Extraction Plan**

From each document in our corpus, we can extract entities and the relationships between them to construct the knowledge graph. In each document, we will need to follow different strategies since the entities and relationships in each document may vary.

**Transcripts**

**Entities**

1. **Named – Entity Extraction**

The named entities include the people, organisations, and policies mentioned in the text. The text includes the statements by the speakers. In these statements, speakers have mentioned and referred to other speakers, some organisations, and policies. We can have ChatGPT identify these entities, and then perform manual string matching to further validate the extracted entities

Further we can use ChatGPT to get the attributes for these entities like their affiliation with organisations, their role / position within that organisation, and the regions they represent. However, since this information will be coming from ChatGPT, we may not get it entirely accurate or complete (specifically Organisations, Policies, and Entity Attributes).

The following table summarizes the entities we can extract from transcripts, the text source for extraction, and the attributes we can extract for each entity.

|  |  |  |
| --- | --- | --- |
| **Entities** | **Extraction Source** | **Entity Attributes** |
| People | * **Speakers** mentioned in transcripts. * **Speakers** mentioned in speaker’s statements. | Affiliation, Role / Position, Location. |
| Organisations | * List of Organisations from ChatGPT. * **Organisations** mentioned in speaker’s statements. |  |
| Policies  Initiative  Plans | * List of Policies from ChatGPT. * **Policies** mentioned in speaker’s statements. |  |

1. **Topic / Theme Extraction**

* Key Themes and Sub – Topics in the combined statements of each speaker. This may help us understand the themes that each speaker is concerned with and addresses the most.
* Representing them as named – entities or labels.

**Relationships**

The relationships are mostly extracted from the statements of speakers. The text in the statements can be used to identify the people (other speakers), organisations, and policies a speaker has mentioned in their statements. This can help us create a network revolving around speakers, and the references they have made in their statements. These extractions will primarily happen through ChatGPT, and later to valid these, we can perform a manual match with the pre-defined list of speakers, organisations, and policies.

Once we identify which Policies are mentioned in speaker’s statements, we can again leverage ChatGPT to identify the sentiments the speaker has when they mentioned these policies.

\*Note:

In some cases, the Organisations and Policies extracted by ChatGPT are a representative of what kind of topics the speaker is referring to, we may not have exact references to the organisations. For example, if the speaker is has mentioned ‘Enterprise Ireland’ we have this linked with the ‘Department of Business, Enterprise and Innovation (DBEI)’. Extracting and mapping Policies also shows similar results.

|  |  |  |
| --- | --- | --- |
| **Entity 1** | **Entity 2** | **What relationship can tell us?** |
| People | Other **People** (speakers) mentioned in their statements. | Can be used to create a social network and to identify people who are most referred. |
| People | **Organisations** mentioned in their statements. | Can be used to identify which organisations mentioned the most and potentially most influential. |
| People | **Policies / Plans / Initiatives** mentioned in speaker’s statements. | Can be used to identify which speaker talks about which policy. |
| People | **Sentiment** | Once we identify the policies referred by each speaker, we can extract their sentiment towards that policy. |

I think it would help if we can go through each of these documents one-by-one and decide which part exactly we would like to process to retrieve the information.

**Project Summary**

Please find a summary of our progress in extracting, processing, and analyzing information from unstructured documents. The work focuses on structured text extraction, entity recognition, stakeholder analysis, and knowledge graph development to enable effective document and knowledge graph search.

1. **Text Extraction**

Extracting text from various types of documents, such as transcripts, policy papers, and publications, each requiring a tailored processing approach.

* **Handling Different Document Structures:**
  + Since each document type follows a unique format, we designed separate pipelines for text extraction and processing.
  + Leveraged **NLP techniques and regex-based pattern matching** to clean and extract relevant content.
* **TF-IDF Based Document Ranking:**
  + Constructed a **TF-IDF term-document matrix** to facilitate keyword-based searches and ranking.
  + Applied keyword filtering to identify **highly relevant documents** related to key topics (**Dairy, Carbon, Water, Emissions, and Livelihood**).
  + Implemented an **elbow point analysis** to determine the most contextually significant documents.

1. **Entity Extraction from Transcripts**

Analysed statement text from transcripts to identify key entities—**policies, stakeholders, and topics**—and assessed the speakers' stance on each entity.

* **Entity Recognition Using ChatGPT API:**
  + Extracted **Policies, Stakeholder Organizations/Groups, and Topics** from individual speaker segments.
  + Evaluated each speaker’s **stance (positive, negative, neutral)** on these entities.
* **Entity Categorization and Sentiment Analysis:**
  + **Policies:** Identified **which policies were mentioned, by whom, and their stance on them**.
  + **Stakeholders:** Mapped speakers to **organizations or groups they referenced**, along with their sentiment toward them.
  + **Topics:** Determined **which key topics (Dairy, Carbon, Water, Emission, and Livelihood) were discussed by each speaker** and their sentiment.

1. **Entity Extraction from Publications**

Publications contain valuable insights into stakeholder interactions. We extracted and structured relationships between stakeholders, along with the context of their collaboration.

* **Stakeholder Relationship Mapping:**
  + **Primary Stakeholder:** The central organization or entity discussed in a given paragraph.
  + **Secondary Stakeholder:** Other stakeholders associated with the primary entity (e.g., funded bodies, partners).
  + **Contextual Relationship:** A **summarized explanation** of how the primary and secondary stakeholders are linked.
  + **Topic Tagging:** Each relationship is assigned a **concise topic label** (e.g., "Funding," "Collaboration").

1. **Entity Extraction from Action Plans and Strategy Documents**

Extracted structured action plans and linked them to relevant stakeholders, ensuring a clear understanding of **who is responsible for what actions and in what timeframe**.

* **Action Plans and Responsible Entities:**
  + Extracted **action items, responsible organizations, and timelines** from structured tables in PDFs.
  + Mapped action plans to relevant stakeholders to establish **who is responsible for implementing different policies and initiatives**.
* **Processing Strategy Documents (e.g., *Food Vision 2030*):**
  + Recognized and extracted **Missions, Goals, and Actions** based on the hierarchical structure of the document.
  + Used **ChatGPT API** to identify stakeholders linked to each action.
  + Established **connections between strategy documents and annual reports** to:
    - Identify **programs or initiatives** supporting each action.
    - Extract and summarize relevant reports linked to each initiative.

**5. Final Tool Features**

**Description:** We have built core functionalities that allow users to efficiently search, analyze, and visualize extracted insights.

* **Advanced Document Search Functionality:**
  + Enables **keyword-based, phrase-based, and partial-string searches** with logical operators (AND/OR).
  + Users can **search within entire documents** or **limit the search to specific speaker segments** in transcripts.
  + Implements a **custom phrase search technique** that improves accuracy by identifying multi-word expressions before vectorization.
* **Knowledge Graph Development:**
  + Work is underway to create **interactive knowledge graphs** that allow users to **explore query-specific connections** between policies, stakeholders, and key topics.
  + These graphs will provide an intuitive way to analyze **stakeholder relationships, policy discussions, and topic linkages**.