# Legal & Finances

##### Session 2.2 (SEMANTiCS)

#### Time: Thursday, September 21, 2023 - 16:45 to 18:00

#### Chair: Christian Dirschl, Chief Content Architect, Wolters Kluwer Germany

## **Talks**

### A Nucleus of a European Legal Data Space

We will present a nucleus of a European Legal Data Space, so a Data Space that aims to cover legal information both for business and research.

Currently there does not exist a European Legal Data Space. This means that people working with legal data have no platform to execute proper NLP or AI. Existing models etc. do usually not work, since the legal language is very specific. Also the legal situation is different in the European countries and therefore also language specific.

However legal startups, public administrations and governments are more and more depending on tools and models that are specifically tailored to legal data. We want to make a first contribution to this growing need with making a platform available, making initial content sets available and an integrated toolset.

We will go through the general architecture, the initial use cases and specific requirements and limitations that come from the domain.

We will show how this Data Space fits into the general objectives of the EU funded Databri-X project, which is focusing on Data Governance.

We will make a short demo and talk about the next steps for enhancing the prototype.

| Christian DirschlWolters Kluwer Deutschland GmbH | Anke LoschWolters Kluwer Deutschland GmbH |
| --- | --- |

### A New Approach to Taxonomy Creation: Combining Human Expertise with NLP to Extract

To mitigate the risk of subjectivity and time-loss in the process of extracting new taxonomy concepts from regulations and law articles, we present a solution used by the Dutch National Police. The application uses entity recognition to mark existing concepts and generates a subset of new possible concepts. To generate this, Large Language Models are enriched with police-specific knowledge from taxonomies defined and curated by our modellers. These suggested concepts are then verified by these same modellers. This human touch, in combination with the Natural Language Processing tooling, increases the accuracy, efficiency and objectivity of the extraction process.

Description:

The Dutch National Police faces the challenge of defining a vast array of concepts for their taxonomies and ontologies, aiming to create a unified semantic model that can be effectively utilized across the entire police organization. These concepts need to be manually extracted from regulations and law articles. Given the extensive nature of these documents and the subjective nature of the process, the manual selection of the concepts can introduce subjectivity and takes a substantial amount of time.

To counteract these problems, we present a solution using entity extraction and supervised Machine Learning to further populate the taxonomies by learning from the existing concepts. The solution enhances accuracy, efficiency, and objectivity, minimizing the potential for human error while keeping the human touch.

With this solution, the Dutch National Police, in partnership with Ordina, uses Natural Language Processing to thoroughly analyse the textual documents. It marks the concepts already present in our police taxonomies and suggests new possible concepts. These concepts are selected by building on existing Large Language Models (LLM) and enriching them with police-specific taxonomies. The suggestions are generated by considering a variety of factors, including the semantic context surrounding police-specific concepts. This is achieved through the implementation of pre-processing techniques such as tokenization and lemmatization, as well as leveraging the power of transfer learning. As a result, the model can be trained effectively to produce meaningful results, even with a smaller dataset. The reason for this being that the LLM provides the base, while the taxonomies add another, more subject-specific layer.

The outcome of using this tooling is two-fold:

- An annotated document that marks the existing concepts within the taxonomy and an overview of those concepts.

- An overview showing all the concepts detected within the document, including newly identified concepts to expand the taxonomy. These are generated in triples, using the SKOS vocabulary.

Both will enable the Dutch National Police to decrease the amount of time required to find and extract new concepts.

At this date of submission, the tooling has already been used by our modellers during the analysation of new regulation documents, further accelerating the selection of new concepts. In further development, the solution will be enriched so that not only regulations and law articles can be analysed, but also other types of documents. The long-term aspiration is to enhance this solution by not only extracting new entities but also automating the construction of new taxonomies.

| Kike FranssenData Scientist / Data Modeller at ORDINA | Eveline SchmidData Scientist / Data Modeller at ORDINA |
| --- | --- |

### 