Introduction

The work presented in this chapter was designed for the purposes of a specific European project, named “CrossJustice”, or CJ from now on. The first part of the chapter focuses on the modeling of definitions within the six european directives of the CJ domain, with the aim of illustrating the steps and the challenges towards the construction of a lightweight ontology to represent this information. The ontology creation process revealed a number of aspects that present opportunities for further research.

In the second part, an automatic experimentation concerning the harmonization of EU directives is presented, built on top of an automatic analyisis of national implementations for each pair of states to investigate the similarity of the corresponding texts. The applied text mining and natural language processing techniques then culminate in the computation of the cosine similarity between vectors associated with legal texts. On that basis, an aggregated index approximates the degree of harmonization within a certain EU directive.

Analogical lightweight ontology building

In this section, a description of a lightweight ontology creation is presented, focused on the domain of the European project “CrossJustice” (or CJ, from now on), i.e. criminal procedural rights in judicial cooperation. The ontology is intended to help legal practitioners understand the precise contextual meaning of terms, as well as helping to inform a rule ontology for modelling rules. While the focus is on a specific domain, this chapter may be used as a reference to guide the creation of lightweight ontologies in other domains.

The sources for the definitions are European directives, legal sources referenced by the directives, and judgments from the Court of Justice of the European Union. Applying the law necessarily involves applying abstract rules and concepts to specific scenarios. A term-based legal ontology can provide a useful reference to find the meaning of terms and their interrelationship with other terms, which can in turn help with search functionalities and rule ontologies.

Background

Ontologies can vary in their level of specificity. For instance, the LKIF Core Legal Ontology [1] is jurisdiction neutral, the LOIS ontology framework [2] has separate entries for EU and legal national terms, while the ELTS ontology framework [3] uses a bottom up

approach which allows multiple definitions of terms with each definition explicitly linked to the source of the definition. The classical definitions used for the ELTS ontology often derive from a specific article dedicated to definitions. They often follow formulaic wording such as "*X means Y*", "*X has the meaning of Y*" or "*X refers to Y*".

A general study of the nature of definitions [4] found that most classic definitions contain hypernyms (usually general rather than direct hypernyms), meronyms, synonyms and purpose-related information. In the framework of the CrossJustice project, we faced the unusual problem that in the six relevant directives, only two of them contain an Article dedicated to definitions. Article 3 in Directive 2016/800 contains 3 definitions, for the terms "child", "holder of parental responsibility" and "parental responsibility". Article 3 in Directive 2016/1919 contains only one definition, for the term "legal aid". There are some classical definitions to be found elsewhere (and we do use them in the ontology). For instance, Recital 15 of Directive states that "[t]he term 'lawyer' in this Directive refers to any person who, in accordance with national law, is qualified and entitled, including by means of accreditation by an authorised body, to provide legal advice and assistance to suspects or accused persons." However, there are not many of these, and apart from not being in the expected place, also have different connecting keywords to those usually used for classical definitions.

This is an opportunity to explore a phenomenon typically neglected in the construction of domain specific legal ontologies. Whether classical definitions are present or absent, laws or, more in general, legal sources, are typically peppered with a number of hidden (in the sense that they are not clearly marked out as definitions) and incomplete definitions, which may nevertheless help legal practitioners (and legal reasoning systems) to reason on the basis of analogy or teleology. Such definitions can be found not only in articles but also recitals, which play an important role in the legal interpretation of the Court of Justice of the European Union (CJEU).

In [5], different types of such definitions were identified and described as follows:

- **Example Definitions**: A concept is explained in terms of typical examples. This class of definition in particular invites reasoning by analogy. There is a sense of completeness, that the instances must belong either to the examples or something similar.
- **Include or Exclude Definitions**: Include/example definitions are often used to emphasise the inclusion or exclusion of certain items where this would otherwise be uncertain or even surprising. Include/exclude definitions are incomplete as there may (or may not) be other items that are included or excluded.
- **Definitions by Reference**: Some legislation explicitly refer to other legislation for definitions of certain concepts. Those definitions apply also to the referring legislation by virtue of the explicit reference. The scope of a definition may be expanded to cover another legislation where there is explicit reference to the definition from that other legislation.

In this chapter, a report on the work of collecting and representing such definitions in the domain of criminal procedural rights is illustrated.

Types of definitions

In the work of collecting and representing definitions for the purpose of the lightweight ontology, it is useful to start with and refine the classes described in [5]:

- The **Sense Definition** is what we typically imagine when we consider definitions and often have formulaic phrases to link the definiens with the 90 definiendum, such as X means Y, X is understood to mean Y. Di Caro [4] found that Sense Definitions typically contain synonyms, hypernyms, meronyms and/or purpose-related information. For our purposes, what distinguishes Sense Definitions from the other definition types described below is that they have a sense of completeness. As an example, Article 1(1), Dir. 2010/64 states "This Directive lays down rules concerning the right to interpretation and translation in criminal proceedings and proceedings for the execution of a European arrest warrant." From this we obtain the following Sense Definition for Directive 2010/64:"EU legal act providing rules concerning the right to interpretation and translation 100 in criminal proceedings and proceedings for the execution of a European arrest warrant."[1]
- The **Part Definition** consists of components or elements of a concept such as procedures or rights, where the meaning is best understood by the sum of its parts. For example, in Article 4(2) of Directive 2012/13, we can consider each numbered item as a component of the information to be provided in a Letter of Rights: "2. In addition to the information set out in Article 3, the Letter of Rights referred to in paragraph 1 of this Article shall contain information about the following rights as they apply under national law:
  (a) the right of access to the materials of the case;
  (b) the right to have consular authorities and one person informed;
  (c) the right of access to urgent medical assistance; and
  (d) the maximum number of hours or days suspects or accused persons may be deprived of liberty before being brought before a judicial authority."
- The **Essential Part Definition** consists of components or elements of a concept that are crucial for the existence of that concept. For example, in Recital 33 Dir. 2016/800, "Confidentiality of communication between children and their lawyer is key to ensuring the effective exercise of the rights of the defence and is an essential part of the right to a fair trial." The connecting key phrases "is key to" and "is an essential part of" are suggestive of an Essential Part Definition in this instance, but there are other such keywords.[2]
- The **Purpose Definition** seeks to explain a concept by its purpose. For example, there are two legitimate purposes for refusing access to certain materials in Article 7(4) Dir. 2012/13: "By way of derogation from paragraphs 2 and 3, provided that this does not prejudice the right to a fair trial, access to certain materials may be

refused if such access may lead to a serious threat to the life or the fundamental rights of another person or if such refusal is strictly necessary to safeguard an important public interest, such as in cases where access could prejudice an ongoing investigation or seriously harm the national security of the Member State in which the criminal proceedings are instituted". As such, we put as secondary concepts the following purposes: 1) to avoid prejudicing an ongoing investigation and 2) to avoid seriously harming the national security of the Member State in which the criminal proceedings are instituted.[3]

- The **Parameter Definition** contains one or more parameters that are taken into account in the application of a legal concept and that helps to understand that concept more clearly. Article 8(2) Dir. 2016/800 provides a good example of a Parameter Definition where we have a parameter for multiple legal concepts: "The results of the medical examination shall be taken into account when determining the capacity of the child to be subject to questioning, other investigative or evidence-gathering acts, or any measures taken or envisaged against the child".[4]
- The **Ratione Temporis Definition** is constituted by the timeframe of application of a legal concept, such as a principle, right, obligation or even the whole directive. For example, Article 2(1) Dir. 2016/800 enshrines two Ratione Temporis Definitions, in that it states that "This Directive applies to children who are suspects or accused persons in criminal proceedings. It applies until the final determination of the question whether the suspect or accused person has committed a criminal offence, including, where applicable, sentencing and the resolution of any appeal".[5]
- The **Ratione Persone Definition** identifies the subjects of a legal concept, such as a principle, right, obligation or even the whole directive. For instance, Article 2 Dir. 2016/343 enshrines that "This Directive applies to natural persons who are suspects or accused persons in criminal proceedings. It applies at all stages of the criminal proceedings, from the moment when a person is suspected or accused of having committed a criminal offence, or an alleged criminal offence, until the decision on the final determination of whether that person has committed the criminal offence concerned has become definitive".[6]
- The **Typical Example Definition** (a subclass of the Example Definition) is based on a typical example of a wider concept in order to provide the latter's definition. For instance, in Article 2(3) Directive 2010/64: "The right to interpretation under paragraphs 1 and 2 includes appropriate assistance for persons with hearing or speech impediments."
- The **Atypical Example Definition** (a subclass of the Example Definition) is based on a specific example of a wider concept that is not commonly included in the wider concept. For instance, in Recital 16, Directive 2016/800 the conclusion of the proceedings "is understood to mean the final determination of the question whether they have committed the offence, including, where applicable, sentencing and the resolution of any appeal". The legislature decided to clarify that the conclusion of the proceedings includes the resolution of any appeal, which is not commonly conceived as a stage of the proceedings and therefore represents an atypical example. Note that the connecting keyword "include" can be indicative of a Typical or Atypical Definition, depending on the context.

- The **Important Example Definition** (a subclass of the Example Definition), like the Typical Example Definition, provides an example of a wider concept in order to provide the latter's definition. However, in this case, while it invites wider analogy, it emphasises that at least the inclusion of this particular case must be respected. For instance in Recital 13, Directive 2013/48, the duty of care towards suspected or accused persons who are in a potentially weak position is emphasised "in particular" towards those who have "any physical impairments which affect their ability to communicate effectively".
- The **Parameter Example Definition** is both a subclass of the Example Definition and of the Parameter Definition. Just like the Parameter Definition, it adopts examples of parameters to clarify a concept. However, like the various Example Definitions described here, the list of parameters is not exhaustive and therefore invites reasoning by analogy. This can be seen in the following example from Recital 4, Dir. 2013/48: "The extent of the mutual recognition is very much dependent on a number of parameters, which include mechanisms for safeguarding the rights of suspects or accused persons and common minimum standards necessary to facilitate 220 the application of the principle of mutual recognition."
- The **Non Example Definition** (a subclass of Example Definitions) is based on an example that is excluded from a wider concept in order to provide the latter's definition. For instance, Recital 13, Directive 2013/48 excludes two specific proceedings for the wider concept of "criminal proceedings", and in so doing, provides a clearer definition of that concept. The norm states that "proceedings in relation to minor offending which take place within a prison and proceedings in relation to offences committed in a military context which are dealt with by a commanding officer should not be considered to be criminal proceedings for the purposes of this Directive".
- The **Definition By Reference** represents the fact that not every piece of legislation contains a definition for every concept, and some legislation explicitly refer to other legislation for definitions of certain concepts. For example, Recital 49, Dir. 2016/343 states that "the Union may adopt measures in accordance with the principle of subsidiarity as set out in Article 5 TEU [Treaty on the European Union]. In accordance with the principle of proportionality, as set out in that Article, this Directive does not go beyond what is necessary in order to achieve those objectives." The definitions for the principle of subsidiarity and the principle of proportionality apply explicitly to Directive 2016/343[7] .

The work described in this chapter is influenced by European Legal Taxonomy Syllabus [3] in the following ways:

- it is assumed that the scope of a definition is the legislative source itself, unless its scope has been explicitly restricted or expanded. In our work, restriction of scope is identified by phrases such as "for the purposes of paragraph X", while expansion of scope is identified by an explicit citation to the definition from another piece of legislation;

- it is assumed that definitions are specific to the jurisdiction of the legislation concerned. In the context of the EU, it is expected that transposition of legislation (and the concepts defined therein) may result in modification to the definition of the concepts, such that it is necessary to define the relationship between related concepts.

The implementation of the ontology, on the other hand, is based on the Linked Term Bank of Copyright-Related Terms [6]. This ontology is also domain-specific, multilingual and multi-jurisdictional. The Copyright Term Bank in turn is built on Lemon and SKOS classes. For the CJ lightweight ontology, the following classes from the Copyright Term Bank have been used:[8]

- *Concept*: the definiens
- *LexicalEntry*: the words or phrases used to represent the context.
- *LexicalSense*: represents the lexical meaning of a lexical entry, and when linked to a Concept, implies that the lexical entry can be used to refer to that Concept.
- *SenseDefinition*: the definiendum, along with the legal source of the definiendum.

To this, the following classes have been added:

- *PartDefinition*
- *EssentialPartDefinition*
- *PurposeDefinition*
- *ParameterDefinition*
- *RationeTemporisDefinition*
- *RationePersoneDefinition*
- *TypicalExampleDefinition*
- *AtypicalExampleDefinition*
- *ImportantExampleDefinition*
- *ParameterExampleDefinition*
- *NonExampleDefinition*

There is no specific class for a Definition by Reference. Instead, all definitions have one or more Scope fields. So in the example described above, the principle of proportionality defined in Article 5 of the Treaty on the European Union has as its scope not only the TEU itself but also Directive 2016/343, due to the Definition by Reference in Recital 49 of that directive.[9]

Since the above-mentioned new definition types necessarily involve a relationship between concepts, it is also important to model the relationship between concepts with the following properties:

- *IsPartOf*
- *HasPart*
- *IsPurposeOf*
- *HasPurpose*
- *IsParameterOf*
- *HasParameter*
- *IsRationeTemporis*

- *HasRationeTemporis*
- *IsRationePersone*
- *HasRationePersone*
- *IsTypicalExampleOf*
- *HasTypicalExample*
- *IsAtypicalExampleOf*
- *HasAtypicalExample*
- *IsImportantExampleOf*
- *HasImportantExample*
- *IsParameterExample*
- *HasParameterExample*
- *IsNonExampleOf*
- *HasNonExample*
- *IsEssentialPartOf*

This duplication has the following advantages:

- It enables the original source text to be easily accessed in the *Definition* instances.
- It enables users to visualise the relationship between different concepts (from the point of view of the relevant legal source).

Ontology creation

Below, the structure of a possible ontology is reported, together with some examples.

The first step is represented by the import of the Linked Term Bank of Copyright-Related Terms into WebProtégé. The Term Bank was available as an N-Triples file (http://www.cosasbuenas.es/blog/copyright-term-bank), which can be then converted into the Resource Description Framework (RDF) format using an online conversion tool (https://www.easyrdf.org/converter). The RDF file can be then imported into WebProtégé.

Here is a summary of the structure of the Linked Term Bank of Copyright-Related Terms:

- *Owl:Thing* has 4 direct subclasses: *Concept*, *Lexical Entry*, *Lexical Sense* and *SenseDefinition*
- Concepts have one or more of the following AnnotationProperties:
    - *rdfs:label*: the most common term for this *Concept* as a *plainLiteral* value
    - *skos:definition*: a link to an instance of *SenseDefinition*, which provides the definition, source and other relevant data
    - *isSenseOf*: a link to one or more *LexicalEntry* instances, which provide the terms used to express the *Concept*
    - *jurisdiction*: a link to a DBPedia entry which provides information about the jurisdiction

    - *reference* (value: *link [dbpedia entry]*)
    - *closeMatch*: a link to a similar concept in the IATE EU terminology database
    - *narrower*: a link to an instance of a narrower *Concept*
    - *rdfs:comment*: as a *plainLiteral* value

The AnnotationProperties *rdfs:label*, *skos:definition* and *isSenseOf* appear in all Concepts.

- *LexicalEntries* have the following Annotation properties:
    - *rdfs:label*: a term used to express a *Concept* in a *plainLiteral* value
    - *denotes*: a link to one or more *Concept* instances denoted by the term in the *LexicalEntry*
    - *language*: the language of the term, in a *plainLiteral* value
    - *sense* (value: *owl:NamedIndividual* of class *LexicalSense*)

The AnnotationProperties *rdfs:label*, *skos:denotes* and *sense* appear in all LexicalEntries.

- LexicalSenses have the following Annotation properties:
    - *reference*: to one or more instances of the class Concept)
- SenseDefinitions have the following properties:
    - *source*: the name of the glossary of terms from where the definition came from, as well as an URI for the glossary
    - *value*: a definition as a plainLiteral value, with the value property itself having a "lang" property

For the ontology, all the above classes and properties have been kept, being interested in representing (classical) sense definitions. However, in addition to SenseDefinitions, a number of other definitions have been created, so that the overall class structure is as follows:

- *Concept*
- *LexicalSense*
- *LexicalEntry*
- *Definition*
    - *SenseDefinition*
        - *PartDefinition*
        - *EssentialPartDefinition*
        - *PurposeDefinition*
        - *ParameterDefinition*
        - *RationeTemporisDefinition*
        - *RationePersoneDefinition*
    - *AnalogicalDefinition*
        - *TypicalExampleDefinition*
        - *AtypicalExampleDefinition*
        - *ImportantExampleDefinition*

- *ParameterExampleDefinition*
- *NonExampleDefinition*

Here is an example of a Typical Example Definition from Article 2(3) Directive 2010/64:

> *The right to interpretation under paragraphs 1 and 2 includes appropriate assistance for persons with hearing or speech impediments.*

In the ontology, the Concept "*the right to interpretation*" is linked to a Typical Example Definition, which has a field for the definition, as well as a comment field providing the original article for reference. There is another Concept for "*appropriate assistance for persons with hearing or speech impediments*".

The Copyright TermBank relies entirely on Annotation Properties for showing links between concepts, their lexical senses, lexical entries and sense definitions. On the other hand, in our ontology we also have definitions that are defined in terms of their relationship between other concepts. As such, we use Relationship Properties to define such relationships. This has the benefit of enabling the viewer to visualise the relationship between different concepts, as can be seen in the screenshot below.

Here is an Important Example Definition from EU Directive 2010-64, Recital 27.

> *The duty of care towards* ***suspected or accused persons who are in a potentially weak position, in particular because of any physical impairments which affect their ability to communicate effectively,*** *underpins a fair administration of justice. The prosecution, law enforcement and judicial authorities should therefore ensure that such persons are able to exercise effectively the rights provided for in this Directive, for example by taking into account any potential vulnerability that affects their ability to follow the proceedings and to make themselves understood, and by taking appropriate steps to ensure those rights are guaranteed.*

Observations for semi-automated ontology population

The analogical ontology described in this chapter has been created entirely manually. Due to the novel features described, there is a lack of suitable data available to attempt automated extraction of definitions and ontology population. However, the data collected in preparation for creating the ontology contains identifiable factors that may be useful for developing such systems in the future. One such factor is the source of the definition. For example, judgments normally provide real examples of what counts as a particular legal concept. The fact that a CJEU hearing is needed to establish such a relation suggests that the definitions extracted are Atypical Example Definitions.

More often, we envisage that a definition classification system might put significant weight on certain keywords (or key phrases) connecting a definiendum with its definiens. This factor has been used extensively in the manual identification and classification of

definitions carried out for this project. For example, in Article 4 Directive 2013/48, the connecting keyword "include" precedes typical examples:

> *Member States shall respect the confidentiality of communication between suspects or accused persons and their lawyer in the exercise of the right of access to a lawyer provided for under this Directive. Such communication shall include meetings, correspondence, telephone conversations and other forms of communication permitted under national law.*

Here are some examples of connecting keywords (and key phrases) and the definition class they normally indicate:

- "*based on*" can indicate a Part Definition (but not always as it can also mean "*reasons for*")
- "*cornerstone*", "*essential part*" or "*presupposes*" indicate an Essential Part Definition
- "*for example*" or "*such as*" indicate a Typical Example Definition
- "*in particular*" and "*at least*" indicate an Important Example Definition
- "*include*" or "*including*" indicate a Part Definition or a Typical, Atypical or Important Example Definition. "*Could include*" indicates a Typical Example Definition. "*Should include*" and "*including as a minimum*" indicate an Important Example Definition.
- "*excludes*" or "*does not include*" indicate a Non Example Definition

However, the relationship is not always straightforward, and real-world knowledge or contextual analysis can help to resolve classification difficulties.

Recital 55 Dir. 2016/800 provides a difficult example with the connecting keyword "including".

> *Children should be treated in a manner appropriate to their age, maturity and level of understanding, taking into account any special needs, including any communication difficulties, that they may have.*

Our rules of thumb for the keyword "*including*" are as follows:

1. if "*including*" is followed by a list of items, these are, by default, typical examples
2. if "*including*" is followed by just item, that item is an atypical example that might otherwise be excluded from inclusion in the broader concept
3. if the item is clearly not an atypical example, then it must be an important example.

According to this reasoning, "*communication difficulties*" should be classed as an important example of "*special needs*". But our real-world knowledge strongly suggests that what we have here is in fact a very typical example, and nothing else in the recital suggests that particular attention must be given to this example. And so, we conclude that in this case, we have a Typical Example Definition. Our real-world knowledge reasoning overrides syntactic reasoning.

There are hard cases that require not only real-world knowledge, but also contextual understanding. Recital 44, Dir. 2013/48 contains the text:

> *Requested persons should have the right to communicate with the lawyer representing them in the executing Member State. It should be possible for such communication to take place at any stage, including before any exercise of the right to meet with the lawyer. Member States may make practical arrangements concerning the duration, frequency and means of communication between requested persons and their lawyer, including concerning the use of videoconferencing and other communication technology [emphasis added] in order to allow such communications to take place.*

Normally, videoconferencing would be considered an atypical means of communication between persons subject to proceedings and their lawyer, but the norm concerns European Arrest Warrant proceedings that often imply geographical distance between the accused person and their lawyer. In this context, videoconferencing is not considered atypical but rather important.

There are also cases where alternative classifications are possible. For example, Recital 55 Dir. 2013/48 contains the following text:

> *This Directive promotes the rights of children and takes into account the Guidelines of the Council of Europe on child friendly justice, in particular its provisions on information and advice to be given to children. This Directive ensures that suspects and accused persons, including children [emphasis added], are provided with adequate information to understand the consequences of waiving a right under this Directive and that any such waiver is made voluntarily and unequivocally.*

The extracted definition of "*suspects and accused persons*" could reasonably be classified as either an Atypical or Important Example Definition. Children are atypical examples of suspects, but they are also an important example in the context of this particular norm, because it is especially important for this group to receive information. An option to represent this as an Atypical Example Definition.

The precise ratio of classification that requires deeper analysis than syntactic rules is unknown at this point. However, given the challenges described above, we envisage that any system for extracting and classifying definitions will require manual post-editing for the foreseeable future.

Similarity and harmonization

Harmonisation is the process of adopting regulations in a common way across the different states, aiming to have the same rules apply to each Member State. A directive comes into effect only after it has been transposed into national law via the so-called National Implementing Measures (NIMs). This section focuses on legal texts officially adopted by the Member States to transpose the provisions of EU directives.

Two main methods have been used for transposing EC law [9] into national law:

i. ‘Copy-out’: implementing legislation adopts the same, or mirrors as closely as possible the original wording of the directive.

ii.. ‘Elaboration’: choosing a particular meaning according to what the draftsperson believes the provision to mean, with the aim of working a provision into something clearer (this is an UK practice).

The typical method for transpositions is ‘copy-out’. In this case, texts of NIMs are expected to be similar. As European Directives provide national legislators of each Member State some discretion in the choice of methods and forms for implementation, the corresponding legal texts (transpositions) are different. Nevertheless, a certain degree of similarity is expected, by comparing the English versions of the six european directives implementations of the CJ domain. In particular, legal experts annotated in the Transposition Tables (TT) of the CJ project four types of annotations:

- *Explicitly transposed [Exp]:* either via new legislation or via amendments to existing legislation.
- *De facto/indirectly implemented [Ind]*: transposition unnecessary because the right already existed in previous legislation.
- *No national implementation [NoN]*: lack of transposing national norm or non-conformity of the national norm with the requirements of the EU provision
- *Specific transposition is not required [Not]*: transposition may be unnecessary because 1) the legal provision lacks deontic or constitutive value e.g. articles 1 and 2 of directives usually only define the scope of the directive, or 2) member states may derogate from a particular provision (e.g. Article 6(3) of Directive 2016/800).

To investigate the similarity of transpositions into the national law, a typical computational approach in Natural Language Processing and in Information Retrieval use to represent text as vectors of terms frequencies (Bag-of-words, BoW representation). Legal texts have to be processed, for instance, by considering fixed-length vectors of words (n-grams).

Related works on similarity and harmonization proposed various automatic techniques in legal domain: i) The analysis of legal texts by using both bag-of-n-grams and the frequency of terms with TF-IDF; ii) Word2Vec implementation as an improvement over traditional bag-of-words; iii) A network analysis to describe the dataset; iv) graph embedding approaches (e.g. Node2Vec) on the network, as a state-of-the-art algorithm for learning embeddings of nodes in a homogeneous network (a network having the same type of nodes). This library aims to map the vertices/nodes of the graph to a vector space such that nodes having similar neighbourhoods in the network have similar embeddings/representations; v) Similarity metrics computation, analysis of results and visualisation.

Methodological framework

In this chapter, a 'text analysis' pipeline as presented in a similar work [8] is reported. The text of NIMs is preprocessed to remove less useful information with the following four main steps:

- [Step 1] By adopting two *regular expressions* to remove some particular cases from the original JSON file, i.e. some "Camel cases" of terms (this kind of errors appears when a term is followed by a new line wrap and then a term with upper case) as well as cases of digits in the form of numbered list (1°, 2°, ecc).

- [Step 2] *Text preprocessing* according to the following passages:
  - lower case reduction
  - stop words and punctuation removal
  - pos-tagging (to consider only: nouns, verbs, and adjectives)
  - stemming (reduce terms in their root form according to Porter stems)

- [Step 3] *Bag-of-words model*. This step transforms each NIM/article to a fixed-length vector of terms (model). With *Bag-of-ngrams* models, by considering n-grams (bigrams and trigrams include respectively two and three words) and obtaining different representations of the same text . Most frequent features can be selected to reduce sparsity.

Here the focus is on the collection of six EU directives made by individual documents *d* (the TT parts/articles of an EU directive) in Step 3, assuming that the corpus is the entire set of NIMs (by considering 'Explicitly transposed' articles) for each EU directive.

Bag-of-words technique aims to represent the text of each document in numbers, based on a vocabulary from all the unique words. Bag-of-ngrams is a more sophisticated approach to create a vocabulary of grouped words of length *n* (i.e., *n*-grams). The corresponding vector of numbers counts the occurrences of terms in the document. A fine-grained measure considers different frequency metric, i.e. Tf-Idf.

Term frequency–inverse document frequency (Tf-Idf) is a numerical statistic for reflecting how important a word is to a document in our collection. The measure implies two parts: Term Frequent (TF) simply describes how frequently a term appears in each document. Inverse Document Frequency (Idf) computes the importance of the term in the complete collection. The Tf-Idf metric is the fraction of the total number of documents over the number of documents which includes the term of interest, by computing the logarithm of the whole fraction to obtain a more compact measure.

- [Step 4] *Document-Term matrix*. Further, the cleaned data need to be converted into a numerical format where each word is represented by a matrix (word vectors). In language processing, the assumption on

vectorisation is that similar text must result in closer vectors (i.e., the vectors derived from textual data to reflect various linguistic properties of the text).

In particular, for each individual NIMs the aim was to obtain a vector of the corresponding Document-Term Matrix (DTM). In the resulting matrix, every row is a NIM (here, a single TT part/article) and every column is a term/stem/n-gram. The values in the matrix are the frequencies of each term in a document.

The columns can be too many, so a Dimensionality Reduction strategy must be applied, e.g. Multi-Dimensional Scaling (MDS) and Principal Component Analysis (PCA).

Notes on computer programming. The *scikit-learn* python library is used, recording results on SQL database tables, and exploiting several methods of interest, e.g.:

- *count_vectorizer*: the method converts a collection of text documents to a matrix of token counts. Two parameters of interest are: *ngram_range*, able to consider n-grams in the text (for bigrams: ngram_range = (2,2); and *max_features* considers the top-ranked features (e.g., max_features = (100)).
- *fit_tranform*: the combination of *fit* and *transform* methods. While *fit* method is calculating the mean and variance of each of the features present in our data, the *transform* method is transforming all the features using the respective mean and variance.

Similarity

The 'similarity' of two legal documents can be assessed "manually" by legal experts, and the challenge here is to automate this similarity computation [10]. Existing automatic methodologies for finding similar legal documents can be classified into two categories:

- (i) network-based methods, which rely on citations to prior case documents;
- (ii)text-based methods, which use the content/textual information of the documents.

As mentioned, in this chapter we report an experimentation of the (ii) approach on 'similarity' in legal informatics concerning the comparison between the EU directive and the transposition into the national law. In a similar recent work, text mining and natural language processing (NLP) techniques have been explored to assist the Commission and legal professionals in studying and evaluating the transposition of directives at a fine-grained provision level [11].

The metric adopted is Cosine Similarity (CS), a measure to compare the vectors of two NIMs (V1 and V2). CS includes the dot product of the vectors V1 and V2. The denominator is the product of their lengths, given by the Euclidean distance. The effect of the document length is compensated by the denominator which normalizes the similarity value. The range of values that the CS can vary is {-1, 1}.

The CS values can be computed between the Member States implementations at the level of each NIM article in the TT (e.g., for a predefined EU directive, making a comparison of the corresponding "Explicit transpositions" of Article 1 both in Italy and in Bulgaria, and so on). Finally, the most similar NIMs for each EU directive is reported.

Notes on computer programming. We adopted the *cosine_similarity* method from the *scikit-learn* python library (*sklearn.metrics.pairwise*).

Finally, to synthesize the results, a visualization technique to describe the similarity between pairs of states on each directive item has been employed. As in Figure 1, a colored table (or heat map) can describe for each article the degree of similarity between pairs of member states. Dark colors (e.g., green or blue) imply no similarity, while lighter colors (e.g., white/yellow) indicate a certain degree of text similarity.

For instance, the following 'heat map' about "Art_5" of EU directive n. 2016/1919 clearly describes how three pairs of States (Italy and German, Germany and Spain, Poland and Germany) have more similar text (light color) than other States. On the contrary, Italy and Poland case seem very different. The diagonal is null (black color in our case), because the relationship between the text of a State and itself is not considered.

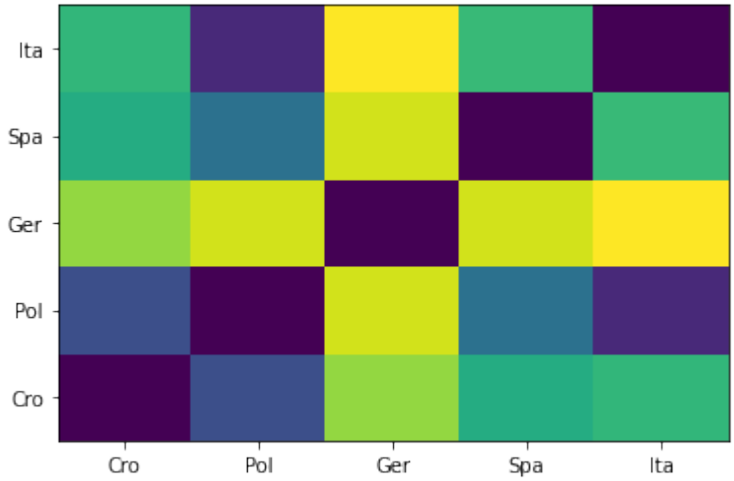

Figure 1. An heat map for similarity scores of Article 5 (2016/1919 EU directive).

Harmonization index

On the basis of similarity indices, the idea is to obtain an aggregated index to describe the degree of similarity of NIMs with respect to a certain EU directive. Such a 'harmonization index' can be compared across states and the six EU directives adopted in the project to investigate how the automatic analysis of legal text may help for identifying the different degree of implementations.

Finally, exploiting the similarity among pairs of states, a proposal of an Harmonization Index is obtained by the following three steps:

- i) select only the texts (provisions) with the label "Explicitly Transposed" for each Directive and for each article (grouping them in case there are more than one for each article)
- ii) on these processed texts, calculate the *cosine similarity* (so for each article we obtain the similarity of the texts between the pairs of states at textual level - represented with heat maps)
- iii) the average of cosine similarity values (AVG(CS)) of the normative texts gives an idea of how similar or not all these parts "Explicitly Transposed" are. These values are continuous, so in addition to the numerical value (multiplied by 100 to be more readable) one proposal is to use three classes: "*High*" if above 30, "*Low*" if below 20, and "*Medium"* if in between.

Table 1. Harmonization index (H-I) and the average values of similarity in EU 2013/48.

Article	AVG(CS)%	H-I
art_2	31,02	High
art_3	21,59	Medium
art_4	21,01	Medium
art_5	18,11	Low
art_6	16,59	Low
art_7	21,92	Medium
art_8	20,68	Medium
art_9	21,37	Medium
art_10	38,11	High

These measures provide an idea about how to address the issue of obtaining an automatic value of an Harmonization Index. It is not possible to calculate the index when the occurrences of the label "Explicitly Transposed" are not significant (for example, in the directive 0343 there are very few cases in the TT). When it is high, it indicates that - looking only at the legal texts - the States are likely to be "aligned" among themselves in applying a certain EU Directive.

References

[1] Hoekstra, R., Breuker, J., Di Bello, M., & Boer, A. (2007). The LKIF Core Ontology of Basic Legal Concepts. *LOAIT*,*321*, 43-63.

[2] Tiscornia, D. (2006, June). The LOIS project: Lexical ontologies for legal information sharing. In *Proceedings of the V Legislative XML Workshop* (pp. 189-204).

[3] Ajani, G., Boella, G., Di Caro, L., Robaldo, L., Humphreys, L., Praduroux, S., Rossi, P. & Violato, A. (2016). The European legal taxonomy syllabus: a multi-lingual, multi-level

ontology framework to untangle the web of European legal terminology. Applied Ontology, 11(4) (pp. 325-375).

[4] Di Caro, L. (2020). What's in a Definition? An Investigation of Semantic Features in Lexical Dictionaries. In *WEBIST*(pp. 225-232).

[5] Humphreys, L., Boella, G., van der Torre, L., Robaldo, L., Di Caro, L., Ghanavati, S., & Muthuri, R. (2021). Populating legal ontologies using semantic role labeling. *Artificial Intelligence and Law*, *29*(2) (pp. 171-211)

[6] Rodriguez-Doncel, V., Santos, C., Casanovas, P., & Gómez-Pérez, A. (2015). A Linked Term Bank of Copyright-Related Terms. In *JURIX* (pp. 91-100)

[7] van Kralingen, R. (1997, June). A conceptual frame-based ontology for the law. In *Proceedings of the first international workshop on legal ontologies* (pp. 6-17)

[8] Emilio Sulis, Llio Humphreys, Fabiana Vernero, Ilaria Angela Amantea, Davide Audrito, Luigi Di Caro, Exploiting *co-occurrence networks for classification of implicit inter-relationships in legal texts*, Information Systems, 2021, 101821, ISSN 0306-4379, https://doi.org/10.1016/j.is.2021.101821

[9] Bernard Steunenberg and Mark Rhinard. "The transposition of European law in EU member states: between process and politics." In: European Political Science Review 2.3 (2010), pp. 495– 520

[10] Bhattacharya, P., Ghosh, K., Pal, A., & Ghosh, S. (2020). Methods for computing legal document similarity: A comparative study. arXiv preprint arXiv:2004.12307.

[11] Nanda, R., Siragusa, G., Di Caro, L. et al. Unsupervised and supervised text similarity systems for automated identification of national implementing measures of European directives. Artif Intell Law 27, 199–225 (2019). https://doi.org/10.1007/s10506-018-9236-y