

Entity Extraction from EU acquis

Jakub Piskorski, Dominik Skotarczak
Publications Office
of the European Union
firstname.lastname@ec.europa.eu

Océane Abrard
Thaleria
abrard.oceane@icloud.com

Abstract

We present a new annotated corpus for the task of Nested Entity Recognition from legal acts in the *EU acquis*, the collection of common rights and obligations that constitute the body of EU law and is incorporated into the legal systems of EU Member States. The corpus is subdivided into two subcorpora: GOLD - human annotated (41 documents) and SILVER - created semi-automatically (537 documents). They contain 28.5K and 275K entity mentions labelled with fine-grained entity types using a dedicated taxonomy of 45 fine-grained entity types. We describe the corpus creation process and present corpus statistics. We use knowledge-, transition-, transformer-, and generative AI-based models to provide baselines and benchmark results.

1 Introduction

Named Entities (NE) constitute a significant part of natural language texts and their proper recognition is crucial for building higher-level Information Extraction (IE) applications. The task of Named Entity Recognition (NER) has been studied since the 1990's [Keraghel et al. \(2024\)](#). Recently, NER has also gained attention in the legal domain ([Ariai et al., 2025](#)), which exhibits several challenges, particularly due to its linguistic complexity and scarcity of publicly accessible corpora.

In this paper, we present a new annotated corpus for the task of Nested Entity Recognition from legal acts in the *EU acquis*, the body of EU law¹. We have developed a two-tier entity taxonomy for the *EU acquis*. It includes 9 main categories subdivided into 45 fine-grained entity types. Unlike other NER corpora, we go beyond the classical NE annotation and also focus on the annotations of nominal mentions of certain entity types, namely, persons (e.g., *citizens*), organizations (e.g., *EU*

agencies), geo-political entities (GPE) (e.g., *EU Member States*), and some temporal expressions. Extraction of such nominal mentions is paramount for higher-level IE tasks in the legal domain, e.g., for determining the type of actor, whether a person, organization or GPE. Entity extraction from the *EU acquis* is challenging due certain characteristics, including, i.a.: (a) legal acts can be extremely long, (b) use of complex syntactical constructions distinct from general domains ([Pereira et al., 2024](#)), (c) presence of less-studied entity types, e.g., products, legislation mentions, (d) inclusion of non-English text fragments, e.g., references to national legislation, names of national organizations, (e) presence of entity mentions (specifically nominal ones), whose type is often difficult to infer even by humans, (f) use of numerous formats to refer to some entities, e.g., references to standards.

The corpus consists of 41 documents (some very long) with 258K tokens, and 28.5K text spans annotated with fine-grained entity labels, out of which ca. 25% and 14% are nominal and nested mentions resp. We present the process of corpus creation and its statistics. We report on the evaluation of various knowledge-, classical ML-, transformer-, and LLM-based baseline models for the coarse/fine-grained entity extraction using the corpus to demonstrate the complexity of the task. We also present a SILVER corpus of 537 documents and 275K annotated text spans created semi-automatically using a combination of these models. To the best of our knowledge, our work is the first one to report on nested fine-grained entity extraction from the *EU acquis*. Our contribution can be summarized as follows:

- we introduce a new task on the extraction of nested fine-grained entities from the *EU acquis*,
- we release a GOLD human-annotated corpus for this task,
- we use the GOLD corpus for comparative

¹<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=legissum:acquis>

evaluation of knowledge-, classical ML-, transformer- and generative AI-based models,

- we release a SILVER corpus automatically created using a combination of the models.

First, in Section 2 we present related work. In Section 3 we introduce the task of Entity Extraction from the *EU acquis*. Next, in Section 4 we report on the creation of the GOLD corpus and provide its statistics. Subsequently, in Section 5 we present the evaluation of the performance of the different baseline models. In Section 6 we describe the creation of the SILVER corpus. We end with some conclusions and future work outlook in Section 7. Supplementary material is provided in appendices.

2 Related Work

Named Entity Recognition (NER) focuses on identifying and classifying named entities in unstructured text into predefined categories. NER has been studied since the 1990’s and was initially fostered by the shared tasks of the Message Understanding Conferences (MUCs) (Chinchor, 1998) and the ACE Programme (Doddington et al., 2004). A vast bulk of research on NER, covering knowledge-, classical machine learning-, transformer- and generative AI-based methods and related resources has been reported (Nadeau and Sekine, 2007; Nouvel et al., 2016; Yadav and Bethard, 2018; Jehangir et al., 2023; Keraghel et al., 2024).

Nested NER focuses on identifying and classifying named entities that are embedded within one another², which is common in many application-oriented scenarios. The Nested NER task is more complex and less studied than classical NER. Various approaches to Nested NER have been reported, including, i.a., rule-based, classical ML and LLMs (Shen et al., 2003; Finkel and Manning, 2009; Kim et al., 2024). An overview of advances on Nested NER can be found in Wang et al. (2022) and Ji et al. (2025).

While NER research in the legal domain is active, it predominantly targets case law and contracts, with national or supranational legislation receiving comparatively little attention (Premasiri et al., 2025). MARCELL Legislative Corpus is a collection of linguistically annotated national legal texts from seven EU Member States (Bulgaria,

Croatia, Hungary, Poland, Romania, Slovakia, and Slovenia), which apart from generic morphosyntactic information also contains automatically extracted NE mentions for the generic categories such as persons, organizations, locations, etc., although covering only 3 of the language (Váradi et al., 2020). LegalNERo extends MARCELL for the Romanian sub-corpus through manually adding gold annotations for legal document references and identifying nested mentions of organizations and dates contained in them (Păiş et al., 2021).

Krasadakis et al. (2022) presented work on detection of NEs that go beyond the general domain (e.g., Law, Euro Case Law, Book, Foreign Law, etc.) from Greek legislative texts, however, the data set has not been made available to the community (Premasiri et al., 2025). Albuquerque et al. (2022) presented UlyssesNER-Br, a dataset composed of Brazilian Portuguese legislative texts with annotated NEs using an inventory of 18 legislation-specific NE types, e.g., entities for law foundations such as bills, legal norms, legislative consultations; entities for law products such as programs or systems created by legislation. The corpus was extended to UlyssesLegalNER-Br (Albuquerque et al., 2026) with broadened NE taxonomy to a larger set of legal categories, with 9 semantic categories and 23 fine-grained types. GerPS-NER described in Feddoul et al. (2024), is a dataset of 24k sentences from German national legislation documents, manually annotated with 10 domain-specific categories for legal norm analysis (e.g., main actor, result receiver, contributor, action, document, condition), which goes beyond NEs and focuses on legal semantic role labelling, but interestingly, encompasses nested annotations. We summarize the information on the NE-related legislation datasets in Annex A in Table 5.

Shared tasks have also advanced legal NER, such as SemEval-2023 Task 6 (LegalEval), although it centred around case law (court judgements) as most of available resources (Modi et al., 2023).

Given the scarcity of resources for NER from legislation texts the work and corpus presented in this paper closes certain gaps in three ways: (a) it focuses on supranational legislation embraced in the *EU acquis*, (b) it introduces fine-grained NE taxonomy tailored to EU legislative structure, and (c) it addresses nested entity extraction.

²Three cases of nesting exist: (a) entities contain one or more shorter embedded entities, (b) entities with more than one entity type (multi-category), and (c) coordinated entities

3 Entity Extraction from *EU Acquis*

The task of Entity Extraction from the *EU acquis* focuses on the recognition of 9 main categories of entities, namely, persons, organizations, locations, products, identifiers, legal references, temporal expressions, events and numerical expressions. This includes the recognition of nominal mentions of persons (e.g., *citizens*), organizations (e.g., *EU agencies*), geo-political entities (e.g., *EU Member States*), and some temporal expressions, including, e.g., frequencies (e.g. *in regular intervals*) as well as the recognition of nested mentions of entities, properly embedded within the top-level entity mentions. We use a two-tier taxonomy, where the aforementioned main categories are subdivided into 45 fine-grained entity types. The entity taxonomy is depicted in Figure 1, whereas Table 6 in Annex B provides the full list of entity type definitions (short versions) accompanied by some examples. We provide an example of text fragment from *EU acquis* with annotated entity mentions in Figure 2.

4 Dataset Description

4.1 Article Selection

For the sake of creating the corpus we have selected 41 legislative documents from the *EU acquis*, the collection of common rights and obligations that constitute the body of EU law and is incorporated into the legal systems of EU Member States. It comprises all the EU's treaties and laws, declarations and resolutions, international agreements, etc. We have sampled the documents in such a manner to cover all main Directorates-General (DGs)³, which are the administrative departments of the European Commission responsible for developing and implementing EU policy, law, and funding. Each DG covers a specific policy area (e.g., agriculture, customs, security, health, etc.). The *EU acquis* documents can be extremely large, the longest one in our dataset consists of 120K characters. A short description of how *EU acquis* documents are structured is provided in Annex C.

4.2 Annotation Process

The annotation process consisted of 3 phases: (a) text annotation, (b) text curation, and (c) consistency assurance. The annotation/curation was done using the INCEPTION platform (Klie et al., 2018).

³https://commission.europa.eu/about/departments-and-executive-agencies_en

4.2.1 Text Annotation

A team of 12 annotators was set up, with expertise in fields such as computational linguistics, linguistics, computer science, statistics, law, out of which 8 had prior experience in annotation tasks. The annotators underwent thorough training, including studying detailed annotation guidelines (see Annex I), participating in live demonstrations and trial annotation exercises. Each document was annotated by at least 2 different annotators.

4.2.2 Text Curation

A team of 5 curators (more experienced annotators) was set up, each assigned to a subset of the annotated documents and tasked to merge complementary annotations and resolve any potential annotation conflicts. The curators held regular meetings to discuss and address ambiguities, to resolve disagreements, to ensure cross-document consistency across annotations, and to refine the annotation guidelines over time.

4.2.3 Consistency Assurance

In this phase, a list of potential inconsistencies was computed, i.e., text fragments labeled with different NE types or flagged with different mention types (nominal versus named) in different contexts. Subsequently, the curators checked and fixed these potential inconsistencies at document/corpus level. Finally, a list of all annotated text fragments was sorted by NE type, and then alphabetically. This format facilitated spotting outliers, i.e., entity mentions labeled with incorrect NE type and errors related to incorrect offsets, e.g., the annotated entity mention was longer than the annotation scope envisaged by the guidelines. This global consistency assurance step has contributed to spot many inconsistencies and errors that were difficult to identify when analyzing/curating single documents.

4.3 Annotation Quality

To assess Inter-Annotator Agreement (IAA), we calculated Krippendorff's α between annotators for each document (Krippendorff, 2004). On average, the IAA is .831 which is above the recommended value of .667 as a threshold for reliable data. The quality of the dataset was further improved using the procedure described in Section 4.2.2 and 4.2.3.

We noticed the following challenges in the annotation phase: (a) differentiating between organizations (ORG) and geo-political entities (LOC-GPE), (b) disambiguating certain nominal expressions,

PERSON (PER)	LEGAL-RESOURCE (LEG)	PRODUCT (PRO)	IDENTIFIER ("IDT")	- DOSSIER	TIMEX (TIM)
- INDIVIDUAL	- INTERNATIONAL	- SERVICE	- PUBLICATION	- PRODUCT	- TIME
- GROUP	- EU	- INITIATIVE	- POSTAL_ADDRESS	- ONLINE	- DATE
ORGANISATION (ORG)	- NATIONAL	- STANDARD	- GEO_COORDINATES	- CREDIT_CARD	- DURATION
- EU_BODY	- SUBDIVISION	- DOCUMENT	- PHONE	NUMEX (NUM)	- PERIOD
- INTERNATIONAL	LOCATION (LOC)	EVENT (EVT)	- EMAIL	- NUMERICAL	- FREQUENCY
- INSTITUTION	- GPE	- MEETING	- URL	- CURRENCY	- DEADLINE
- COMMERCIAL	- FACILITY	- CONFLICT	- IP	- MEASUREMENT	
	- LANDFORM	- INCIDENT-DISASTER	- VAT	- AGE	
		- OCCASSION	- BANK	- COUNTX	

Figure 1: Two-tier Entity Type Taxonomy.

[Switzerland]LOC-GPE also informed the [Commission]ORG-EU that its system for the issuance of [[COVID-19]EVT-IND certificates]PRO-DOC in accordance with the [Swiss [COVID-19]EVT-IND Certificate Order]LEG-NAT complies with the technical specifications laid down in [[Commission]ORG-EU] Implementing Decision([EU]LOC-GPE) 2021/1073]LEG-EU

Figure 2: A sample EU acquis document text fragment and entity annotations.

e.g., ‘user’, which could refer to both PER and ORG, depending on the context, (c) difficulty to assign the fine-grained label, e.g., facility vs. landmark, for locations, whose subtype was difficult to infer without the use of external knowledge, (d) distinguishing initiatives from services in the PRO category, and (e) discrepancies between annotators regarding entity mention boundaries, which was observed specifically in the context of longer text fragments containing nested entity mentions. The enumerated challenges were addressed during the annotation process by refining the annotation guidelines and by including additional examples and counter examples, and by introducing new type of disambiguation patterns (see Annex I).

4.4 Statistics

The GOLD corpus consists of 41 documents, with 1,693,103 characters and 258,102 tokens. The average length of a document is 6,295 tokens, where the shortest and longest document length is 1,866 and 120,500 characters respectively.

28,483 text spans are annotated with entity labels, out of which 7,797 (27.4%) are nominal mentions, and 3,906 (13.7%) are nested mentions (i.e., contained in other entities), and consequently, 86.3% are top-level entity mentions. The overall corpus statistics for coarse-grained entities are provided in Table 1, while the statistics for the fine-grained types are provided in Annex D in Table 7.

The corpus is highly imbalanced, where men-

TYPE	NUMBER (%)	NOMINAL (%)	NESTED (%)
PER	2849 (10.0%)	2349 (82.4%)	34 (1.2%)
ORG	7463 (26.3%)	3667 (49.1%)	966 (12.9%)
LOC	7798 (27.3%)	1733 (22.2%)	1473 (18.9%)
LEG	5590 (19.6%)	0 (0%)	32 (0.6%)
IDT	677 (2.4%)	0 (0%)	15 (2.2%)
PRO	1043 (3.7%)	0 (0%)	87 (8.3%)
EVT	78 (0.3%)	0 (0%)	45 (57.7%)
TIM	2343 (8.2%)	48 (0.02%)	1206 (53.8%)
NUM	642 (2.2%)	0 (0%)	48 (7.5%)
ALL	28483	7797 (27.4%)	3906 (13.7%)

Table 1: GOLD corpus statistics for coarse-grained entity types. NUMBER, NOMINAL, and NESTED columns refer to the total number of instances, nominal, and nested mentions within each entity class resp. The percentage figures in the NOMINAL and NESTED columns refer to the fraction of entity mentions within each entity class which are nominal and nested resp.

tions of organizations, locations and legislation constitute approx. 73% of all entity mentions. Significant fraction of mentions of persons (82.4%), organizations (49.1%) and approx. 31% mentions of geo-political entities (GPE) are nominal since legal acts frequently refer to specific type of persons, organizations and countries. The vast majority of nested mentions (approx. 93.3%) are mentions of organizations, locations and temporal expressions (e.g. dates or durations being part of deadlines or frequencies), where they constitute 12.9%, 18.9% and 53.8% of all instances in the respective entity classes. References to EU institutions and GPEs constitute 85.7% and 89.1% of all nested mentions of organizations and locations, respectively. Out of all nested temporal expressions the most frequent are dates (84%) and durations (14.1%). There are 7,857 unique surface forms annotated.

The GOLD corpus is available in offset-based format (see Annex D.1. Additional statistics and information (e.g., entity mention frequency distribution, most frequent entities, examples of ambiguous surface forms, etc.) are provided in Annex D.

5 Experiments

We compare the performance of various baseline models, including knowledge-based, transition-based system, transformer-, and LLM-based, whose results should serve as benchmarks (rather than as state-of-the-art) to demonstrate the complexity of the task. We split the GOLD corpus into 5 folds at random to allow for the fine-tuning of the transformer model. Document #40 was excluded from the split, since it was used to provide examples for the few-shot LLM. Within the training set (32 documents of each fold), we randomly split the pre-processed sentence-like text snippets (80-20) for training and validation and observed both losses. The remaining 8 documents of each fold were kept exclusively for evaluation.

5.1 Evaluation

We distinguish the following types of entity mentions. *Flat* mentions do not contain mentions of other entities. *Nested* mentions are mentions which are either embedded in other entity mentions or they contain mentions of other entities. Nested mentions are subdivided into *outer* mentions, i.e., the ones that contain other nested entity mentions, but are not nested within other entity mentions, and *inner* mentions, i.e., the ones that are contained within other entity mentions. An entity mention is called *top-level* if it is a flat or outer mention.

We use *recall*, *precision*, and F_1 micro/macro scores in 3 evaluation scenarios: (a) **top-level**: only top-level mentions are considered, (b) **all**: all (flat + nested) mentions are considered, where recognition of inner entities effectively compensate/penalize the recognition of the respective outer entities, and (c) **nested**: only nested mentions (inner + outer) are considered, with a compensation/penalization mechanism for correct/incorrect recognition of inner mentions as in the second scenario.

Regarding matching entity boundaries in text we distinguish between *Strict* and *Type* evaluation modes (Segura-Bedmar et al., 2013). In the former, both the exact boundaries and entity type have to be identified correctly, whereas, in the latter, the entity type has to be correct, but partial credit is given even if the boundaries are slightly off. Formal evaluation definitions are provided in Annex E.

5.2 Models

5.2.1 Knowledge-based (KB)

Given the existence of the Authority tables⁴, which contain, i.a., lists of organizations, locations, etc. that appear in *EU acquis* documents and the fact that, intuitively, many domain-specific entity types (e.g., identifiers, references to legislation, temporal references) follow certain strictly defined formats, we have developed a knowledge-based approach which combines lexical resources derived mainly from the aforementioned authority tables (including all EU language variants) and a cascade of approx. 100 patterns based on regular expressions over feature structures (see Annex F.1 for details). Inner entity mentions are extracted by applying the lexical resources and patterns recursively.

5.2.2 Transition-based

We also included spaCy’s *en_core_web_lg* to assess an (out-of-the box) transition-based NER system that utilizes CNN-encoded representations for its features (Honnibal et al., 2020). As spaCy’s model was trained for generic NER types, we only mapped spaCy’s entity types to the coarse-grained entity categories.

5.2.3 Transformers

To test a transformer model, we selected the span-based GLiNER-L v2.1 (**GLiNER**) with 0.3B parameters model with the underlying deberta-v3 backbone (Zaratiana et al., 2024; He et al., 2023). We assessed this model performance with and without fine-tuning on the GOLD dataset. The suggested hyperparameter combination yielded the best results on the training set for each fold.⁵ For additional training with the SILVER corpus (section 6), we executed a two-stage curriculum learning approach (pre-fine-tuning on SILVER data, followed by fine-tuning on the training set of each fold in the GOLD corpus), as suggested in ‘Simple Mix’ in Huang et al. (2024). See Annex F.2 for details.

5.2.4 Generative AI

We used Claude-Sonnet-4.6 (**Claude**). For cost efficiency, we modified the entity-category method suggested by Kim et al. (2024). Rather than prompting the LLM to generate plain text (which

⁴<https://op.europa.eu/en/web/eu-vocabularies/authority-tables>

⁵<https://github.com/urchade/GLiNER/blob/main/examples/fine-tune.ipynb>

requires complex post-processing and is prone to hallucinated categories), we utilized Claude’s function-calling capabilities to enforce structured data extraction. Specifically, we converted our fine-grained taxonomy into a formal JSON schema (the tool definition). We then instructed the model to prepare input for this tool by populating the schema with the entities it identified in the text. This forced the LLM to categorize entities strictly according to our taxonomy and guaranteed a structurally valid, easily parsable output.

We kept the temperature at 0 throughout the experiments. The tooling used in the entity extraction pipeline for **Claude** constrained us to the maximum of 50 text chunks per document. Initially, we set the number of lines per text chunk to 40 in order to meet this constraint for all documents in the corpus (zero-shot and few-shot). We later passed input text dynamically to minimize the amount of text passed at once (Dynamic Input - DI). It brought down the average number of characters passed at once from 5-6k to >1k. For each prediction made by **Claude**, we mapped the output dictionary to the character offsets in the benchmark corpus.

In the few-shot scenario, the context provided to the LLM comprised one entire document that included multiple examples of correct entities extracted from that single document. Specifically, the demonstration document contained 338 extracted entities. It provided a full textual context, while the multiple extracted entities served as the ‘shots’ to demonstrate the desired output format of the function-calling tool. This prompt structure also extended the few-shot prompt approach beyond the 4 entity types presented in Huang et al. (2024). Figure 6 in Annex F.3 shows the template of our few-shot prompt.

5.2.5 Hybrid

Given the different strengths the models showed with respect to the specific types of entities, we combined **KB** and **Claude** into a hybrid system.

When an entity is detected by only one system in a specific text span, it is retained only if that system’s corresponding F_1 score for that entity type is higher than the other system’s score. For example, if **KB** detects an entity of type X and **Claude** detects nothing, the entity is included in the final output only if the **KB**’s F_1 score for category X exceeds Claude’s F_1 score for that same category.

When both systems generate predictions in the same text region, resulting in overlapping or nested

entities, the resolution shifts to the comparison of the outermost entities. In these conflict scenarios, we isolate the longest entity boundaries predicted by each system within the contested region. To determine the winning system, we calculate a score for each model’s outermost span by weighting its precision for that specific entity type against the log-dampened length of the span. The system whose outermost span achieves the higher weighted score dictates the final output for that entire overlapping section. We exploit the hybrid system to create a SILVER corpus described in Section 6.

5.3 Results

The evaluation results for the coarse- and fine-grained entity recognition task for both *Strict* and *Type* evaluation scenarios are provided in Table 2. Table 3 provides the results for the *Strict* scenario for selected (best) models per coarse-grained entity type, whereas the results for the same models per fine-grained entity types are provided in Annex G in Table 11.

For the top-level entity mentions with fine-grained types, the average *micro* F_1 score ranged from .113 to .677 in the *Strict* evaluation regime. In all single-model scenarios for top-level entities, **Claude** surpassed the other approaches (in terms of *micro* F_1) after introducing examples from one of the annotated documents and lowering the average amount of input document text passed at once.

After including nested entities in the evaluation, we observed a slight decrease in nearly all performance metrics across **KB** and **Claude** models (all mentions). We found that within nested entities, outer entity mentions were extracted more precisely and exhaustively compared to their inner mentions. The final **Claude** approach remained the best performing model in terms of *micro* and *macro* F_1 scores, and the fine-tuned **GLiNER** reached the highest recall score across all evaluation scenarios for all entity mentions.

We raised **GLiNER**’s avg. *micro* F_1 score to .612 by in-domain fine-tuning on the respective training sets across the 5 folds (**GLiNER**(FT) in Table 2). We observed similar effects of fine-tuning in all top-level evaluations (*Strict*, *Type*) for the two granularity levels (fine-, coarse-grained). The improvement was higher for the coarse-grained evaluation. We reason that the decrease in entity type granularity simplified the task and allowed **GLiNER** to rely on the previously trained entity types (PER, LOC, ORG) that are more common in

NER tasks. It also provided **GLiNER** with more examples per entity type. Moreover, the finetuned **GLiNER** matched **Claude**’s performance (.721 vs .731 on *Strict*) in terms of the *micro* F_1 score for the coarse-grained types. The difference was below the standard deviation across the 5 folds (Table 2). While **GLiNER** achieved the highest recall, which became more evident after collapsing the entity types into the coarse-grained types, its precision remained lower relative to both **KB** and **Claude**.

The **GLiNER** model also underperformed relative to **KB** and **Claude** in terms *macro* F_1 score even after decreasing the granularity of entity types. We report on those metrics in Table 3.

Observing the difference between *micro* and *macro* F_1 scores, we inspected the performance metrics of interest across all coarse-grained entity types for the 3 best models. We observed large differences across the entity types. Among them, the extraction of identifiers, products, events and also numerical expressions was the most challenging (Table 3). Nominal mentions and complex annotation guidelines for mentions of PER increased the complexity for PER types recognition as well. We discuss those in detail in the error analysis section.

We also found that **KB** surpassed the other models for the mentions of legal resources. **Claude** was best at extracting the mentions of identifiers, products, events, and numerical expressions. **Claude** may, generally, be the preferred method for entity types with a high diversity, such as, products or events and surprisingly, identifiers. Then, **GLiNER** achieved the highest scores for mentions of persons, organizations, and time expressions. The latter two models tied for the mentions of locations.

The *macro* F_1 scores were noticeably lower for **GLiNER** throughout the experiments. To improve this metric, especially for the fine-grained taxonomy, we considered the results for the hybrid scenario. Our **Claude+KB** model achieved the highest F_1 scores (*micro* and *macro*) in nearly all evaluation scenarios and regimes.

Including data from the SILVER corpus increased **GLiNER**’s overall performance for the fine-grained taxonomy and specifically for 24 fine-grained types. That said, the improvement was unequal across the fine- and coarse-grained entity types. We found that while the model improved on the most domain-specific types (LEG, IDT, PRO), the training on the SILVER corpus slightly decreased its performance on more common coarse-grained entity types, e.g., PER, LOC,

ORG. Hence, we observed that introducing domain-specific entity subtypes comes at the risk of shifting **GLiNER**’s weights from previously learned labels for more common entity types. We report on the fine-grained types in Table 12 in Annex G.

We also observed that, not surprisingly, the detection of nominal entity mentions is harder than recognition of named entity mentions. We provide in Table 13 and 14 in Annex G a comparison of recall figures for the different models on nominal versus named entity mention recognition for both fine- and coarse-grained entity types.

5.4 Error Analysis

Claude detected numerical expressions but struggled to distinguish between NUM subtypes (e.g. plain numbers vs. measurements). It made errors when extracting tabular layout (context), domain-specific measurement conventions, and the corpus’ span annotation policy. For example, it detected too long sequences and erroneously recognized paragraph numbers (out of scope). The model also detected too many nominal mentions of entities which were out of scope, e.g., nominal mentions of events. As regards persons the model had difficulty to distinguish between individuals and groups, and had the tendency to overuse the generic PER label, and to label tokens being out of scope, e.g., honorifics.

KB approach which relies on lexical resources and local extraction patterns exhibited typical errors for such approach. For instance, **KB** failed to detect fine-grained entity type for some organisation mentions (e.g., ‘*competent authority*’) due to not considering broader context. It also failed to detect many mentions of persons due to “imperfect” lexical resources. Furthermore, many entity type errors and spurious extractions were due to ambiguous trigger words, e.g., ‘*EC*’ might refer to both European Commission (ORG-EU) and European Community LOC-GPE). Other errors were related to positional misalignment. **KB** approach exhibited high precision for recognizing of many types of identifiers, but often, due to the use of multitude of uncommon formats in the *EU acquis* it failed to recognize them, where the generative AI approach performed better.

Unsurprisingly, **GLiNER** (fine-tuned on GOLD only) performed poorly on the recognition of low-populated entity classes, e.g., events, some sub-classes of identifiers. In general, **GLiNER** consistently assigned coarse-grained labels to mentions of

Model	STRICT				TYPE			
	P	R	$micro F_1$	$macro F_1$	P	R	$micro F_1$	$macro F_1$
top-level mentions (w/o comp/penalty)				top-level mentions (w/o comp/penalty)				
KB	.705(.061)	.614(.085)	.655(.073)	.567(.053)	.727(.062)	.634(.087)	.677(.075)	.600(.064)
en-core-web-ig	.370(.065)	.213(.050)	.269(.054)	.265(.042)	.509(.075)	.292(.044)	.370(.050)	.343(.048)
gliner(ZS)	.319(.055)	.273(.056)	.294(.056)	.238(.018)	.350(.057)	.302(.059)	.324(.058)	.253(.010)
gliner(FT)	.674(.023)	.775(.027)	.721(.023)	.545(.027)	.688(.020)	.801(.021)	.740(.018)	.557(.038)
gliner(FTS)	.651(.039)	.758(.023)	.700(.030)	.525(.045)	.660(.037)	.777(.021)	.714(.029)	.547(.046)
Claude(ZS)	.753(.024)	.626(.027)	.684(.025)	.616(.040)	.769(.021)	.643(.025)	.700(.022)	.635(.039)
Claude(FS)	.794(.017)	.656(.022)	.718(.017)	.643(.037)	.808(.014)	.671(.020)	.733(.014)	.654(.039)
Claude(FS+DI)	.761(.015)	.704(.021)	.731(.014)	.650(.035)	.774(.015)	.719(.020)	.746(.013)	.661(.030)
Claude+KB	.810(.022)	.706(.021)	.754(.017)	.649(.020)	.824(.022)	.720(.021)	.768(.016)	.698(.024)
all mentions (w/ compensation/penalty)				all mentions (w/ compensation/penalty)				
KB	.693(.057)	.606(.076)	.646(.066)	.556(.053)	.717(.053)	.631(.072)	.670(.062)	.587(.062)
en-core-web-ig	.370(.066)	.224(.046)	.278(.051)	.277(.045)	.504(.074)	.307(.043)	.380(.050)	.322(.043)
gliner(ZS)	.319(.055)	.273(.059)	.294(.056)	.238(.018)	.352(.058)	.305(.059)	.326(.059)	.247(.014)
gliner(FT)	.683(.024)	.765(.024)	.732(.021)	.542(.040)	.697(.020)	.789(.021)	.740(.016)	.560(.033)
gliner(FTS)	.664(.035)	.758(.022)	.708(.028)	.546(.046)	.673(.034)	.776(.021)	.721(.027)	.653(.045)
Claude(ZS)	.753(.019)	.607(.028)	.672(.023)	.601(.023)	.770(.017)	.625(.028)	.690(.022)	.618(.024)
Claude(FS)	.788(.008)	.648(.022)	.705(.014)	.626(.018)	.804(.005)	.655(.021)	.722(.013)	.635(.019)
Claude(FS+DI)	.763(.011)	.683(.020)	.720(.011)	.638(.026)	.778(.013)	.701(.019)	.737(.014)	.649(.020)
Claude+KB	.807(.015)	.691(.013)	.744(.012)	.644(.021)	.822(.015)	.707(.013)	.760(.012)	.689(.030)
only nested (w/ compensation/penalty)				only nested (w/ compensation/penalty)				
KB	.720(.082)	.536(.070)	.614(.075)	.532(.101)	.746(.073)	.588(.062)	.657(.065)	.582(.085)
en-core-web-ig	.363(.092)	.185(.071)	.244(.081)	.238(.033)	.520(.079)	.284(.073)	.366(.078)	.301(.065)
gliner(ZS)	.575(.058)	.219(.041)	.316(.047)	.302(.033)	.623(.051)	.262(.047)	.367(.051)	.320(.030)
gliner(FT)	.785(.053)	.530(.028)	.632(.031)	.489(.071)	.812(.048)	.597(.029)	.687(.026)	.566(.015)
gliner(FTS)	.793(.024)	.541(.025)	.643(.023)	.521(.032)	.813(.019)	.596(.022)	.688(.017)	.553(.035)
Claude(ZS)	.828(.026)	.520(.039)	.628(.035)	.575(.063)	.851(.028)	.560(.039)	.675(.035)	.593(.027)
Claude(FS)	.826(.035)	.553(.037)	.661(.035)	.590(.051)	.849(.037)	.595(.035)	.699(.035)	.628(.049)
Claude(FS+DI)	.854(.035)	.590(.034)	.697(.032)	.616(.049)	.874(.032)	.634(.030)	.735(.028)	.652(.038)
Claude+KB	.862(.042)	.626(.007)	.725(.015)	.624(.021)	.877(.040)	.663(.009)	.755(.013)	.668(.015)

Model	STRICT				TYPE			
	P	R	$micro F_1$	$macro F_1$	P	R	$micro F_1$	$macro F_1$
top-level mentions (w/o comp/penalty)				top-level mentions (w/o comp/penalty)				
KB	.666(.051)	.580(.079)	.619(.066)	.543(.042)	.686(.052)	.599(.081)	.639(.068)	.570(.040)
gliner(ZS)	.123(.024)	.105(.021)	.113(.022)	.224(.033)	.141(.026)	.121(.022)	.131(.023)	.238(.035)
gliner(FT)	.573(.052)	.660(.067)	.612(.058)	.492(.056)	.579(.052)	.668(.067)	.621(.058)	.499(.048)
gliner(FTS)	.600(.037)	.698(.025)	.645(.030)	.515(.019)	.605(.037)	.705(.035)	.651(.030)	.517(.022)
Claude(ZS)	.688(.025)	.605(.019)	.643(.020)	.531(.017)	.715(.021)	.595(.018)	.649(.018)	.547(.018)
Claude(FS)	.740(.034)	.611(.022)	.670(.026)	.546(.021)	.752(.033)	.622(.020)	.681(.024)	.556(.024)
Claude(FS+DI)	.707(.017)	.654(.028)	.679(.021)	.561(.027)	.715(.016)	.663(.027)	.688(.020)	.575(.025)
Claude+KB	.764(.026)	.666(.033)	.712(.028)	.581(.044)	.776(.026)	.678(.034)	.723(.029)	.600(.044)
all mentions (w/ compensation/penalty)				all mentions (w/ compensation/penalty)				
KB	.659(.049)	.578(.071)	.614(.061)	.531(.050)	.680(.046)	.599(.069)	.637(.058)	.562(.042)
gliner(ZS)	.146(.028)	.117(.021)	.130(.024)	.227(.032)	.168(.030)	.135(.022)	.150(.025)	.249(.039)
gliner(FT)	.586(.058)	.657(.067)	.620(.061)	.476(.053)	.597(.058)	.676(.069)	.634(.062)	.490(.047)
gliner(FTS)	.618(.034)	.706(.024)	.659(.029)	.526(.028)	.623(.033)	.722(.023)	.671(.027)	.527(.031)
Claude(ZS)	.693(.020)	.591(.012)	.638(.011)	.532(.018)	.720(.016)	.583(.015)	.644(.012)	.552(.014)
Claude(FS)	.739(.027)	.600(.014)	.660(.017)	.534(.017)	.752(.025)	.611(.013)	.674(.015)	.560(.020)
Claude(FS+DI)	.713(.014)	.638(.021)	.674(.016)	.557(.028)	.724(.012)	.651(.020)	.685(.015)	.577(.034)
Claude+KB	.766(.018)	.656(.025)	.706(.021)	.584(.047)	.778(.018)	.700(.025)	.720(.022)	.605(.051)
only nested (w/ compensation/penalty)				only nested (w/ compensation/penalty)				
KB	.710(.086)	.529(.073)	.606(.078)	.546(.091)	.736(.076)	.580(.066)	.648(.069)	.588(.078)
gliner(ZS)	.381(.091)	.140(.019)	.203(.033)	.333(.042)	.431(.094)	.168(.021)	.241(.036)	.343(.050)
gliner(FT)	.723(.111)	.487(.067)	.581(.081)	.453(.046)	.750(.109)	.544(.075)	.630(.085)	.522(.055)
gliner(FTS)	.781(.030)	.533(.030)	.634(.029)	.500(.019)	.800(.024)	.585(.027)	.676(.023)	.535(.041)
Claude(ZS)	.792(.019)	.511(.051)	.620(.043)	.521(.034)	.820(.024)	.535(.030)	.648(.026)	.568(.030)
Claude(FS)	.796(.027)	.533(.023)	.638(.022)	.519(.025)	.818(.028)	.568(.022)	.670(.021)	.578(.017)
Claude(FS+DI)	.833(.033)	.575(.026)	.680(.024)	.551(.024)	.852(.030)	.612(.023)	.712(.021)	.608(.017)
Claude+KB	.843(.041)	.613(.009)	.709(.015)	.581(.016)	.858(.037)	.648(.014)	.738(.014)	.629(.022)

Table 2: Evaluation for coarse- (TOP TABLE) and fine-grained types (BOTTOM TABLE): *Strict* (LEFT) and *Type* (RIGHT) using precision (P), recall (R), $micro/macro F_1$. In the *Strict* evaluation, we do not calculate composite scores for the top-level entities that are outer entities. Standard deviations are given in brackets. ZS - zero shot, FS - few shot, FT - fine-tuned, FTS - fine-tuned on SILVER, DI - dynamic input.

Model	PER			ORG			LOC			LEG			IDT			PRO			EVT			TIM			NUM		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
KB	.372	.277	.317	.605	.573	.589	.800	.654	.719	.974	.858	.913	.648	.491	.559	.804	.253	.385	.032	.143	.053	.744	.561	.640	.237	.533	.328
gliner(FT)	.645	.678	.661	.800	.844	.820	.795	.835	.814	.802	.793	.800	.085	.100	.092	.163	.501	.246	.055	.286	.091	.741	.703	.721	.300	.358	.325
Claude(FS+DI)	.600	.513	.553	.804	.644	.715	.890	.751	.814	.812	.822	.814	.739	.858	.794	.394	.578	.469	.040	.286	.070	.807	.590	.681	.421	.533	.470
gliner(FTS)	.619	.699	.657	.821	.784	.802	.769	.820	.793	.856	.802	.828	.242	.186	.210	.215	.489	.299	.015	.051	.023	.717	.713	.715	.137	.564	.220

Table 3: Evaluation (all mentions - STRICT only) per coarse-grained type in terms of precision (P), recall (R), $micro/macro F_1$ with FS - few shot, FT - fine-tuned, FTS - fine-tuned on SILVER corpus, DI - dynamic input.

fine-grained entities. We reasoned the model continued to rely on the general labels used in previous training due to an insufficient number of training examples, given the complexity of the taxonomy. **GLiNER** produced many false positive mentions of products. It often assigned the incorrect sub-type to the recognized numerical expressions and detected spurious nominal mentions of numbers, currencies and units.

6 Silver corpus

Given that the **GOLD** corpus is of medium size and some classes are scarcely populated we created a **SILVER** corpus in the following manner. We selected a set of 537 *EU acquis* documents using the procedure described in Annex H.1 and run the hybrid system described in 5.2.5 on this document selection. Next, since all models struggled with the **NUM** category ($F_1 < .5$) we manually corrected the hybrid system responses for this category. This cleaned version of the system response constitutes the **SILVER** corpus.

The **SILVER** corpus statistics for the coarse- and fine-grained entity types are provided in Table 4 and Table 15 in Annex H resp. The **SILVER** corpus consists of 537 documents with 2,675,01 tokens. The average length of a document is 4,982 tokens, where the shortest and longest document length is 666 and 678,761 characters resp.

During **SILVER** corpus creation, we observed 159,110 overlapping instances between **KB** and **Claude**'s outputs. The majority 105,115 (66.1%) were exact span and type matches, while 53,995 were genuine conflicts. 16,423 (30.4%) of those conflicts concerned different entity types. 13,515 (25.0%) were character offset mismatches with the same labels, and the remaining 24,057 (44.6%) instances included different character spans and entity types. Our approach to resolving those conflicts has been described in Section 5.2.5.

275,146 mentions of entities are labeled, out of which 34,388 (12.5%) are *inner* nested entity mentions, and, consequently, 87.5% are top-level mentions. Analogously to **GOLD** corpus, the **SILVER** corpus is highly imbalanced, where mentions of organizations, locations, legislation and numerical expressions constitute approx. 77% of all entity mentions. Noteworthy, contrary to the **GOLD** corpus, the **SILVER** corpus contains significantly more numerical expressions, i.e., 28.7%, which is due to inclusion of some documents with large

TYPE	NUMBER (%)	NESTED (%)
PER	19821 (7.2%)	3189 (16.1%)
ORG	44063 (16.0%)	9506 (21.6%)
LOC	38392 (13.9%)	6372 (16.6%)
LEG	47162 (17.1%)	458 (1.0%)
IDT	6821 (2.5%)	63 (0.9%)
PRO	9940 (3.6%)	850 (8.6%)
EVT	769 (0.3%)	165 (21.6%)
TIM	26453 (9.6%)	9913 (37.5%)
NUM	81725 (29.7%)	3872 (4.7%)
ALL	275146	34388 (12.5%)

Table 4: **SILVER** corpus statistics for the coarse-grained entity types. **NUMBER** and **NESTED** columns refer to the total number of instances and nested mentions within each entity class resp. The percentage figure in the **NESTED** column refers to the fraction of entity mentions within each entity class which are nested.

tables containing numerical values.

The majority of nested mentions (approx. 75%) are mentions of organizations, locations and time references, where they constitute 21.6%, 16.6% and 37.5% of all instances in the resp. entity classes. References to EU institutions and GPEs constitute 84.5% and 96.5% of all nested mentions of organizations and locations resp. Out of all nested time references the most frequent are dates (76.7%).

7 Conclusions and Outlook

We presented a new corpus for the task of Nested Entity Recognition from legal acts in the *EU acquis*, the body of EU law⁶, subdivided into two sub-corpora: **GOLD** - human annotated (41 documents), and **SILVER** - semi-automatically created (537 documents), which contain 28.5K and 275K entity mentions labeled using a fine-grained entity taxonomy. We described the corpus creation process, provided corpus statistics, and evaluated various types of baseline models. The corpus is freely available for research purposes⁷, and we believe it constitutes a useful resource for researchers and practitioners working on IE solutions for the legal domain.

In the future, we plan to: (a) enhance the size of both **GOLD** and **SILVER** corpora, (b) extend the corpus to cover EU Case law, which consists primarily of the landmark judgments, preliminary rulings, and orders delivered by the Court of Justice of the EU, and (c) work on modeling relation extraction for the *EU acquis*, focusing initially on extraction of information on legal obligations of different stakeholders within the EU landscape.

⁶<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=legisum:acquis>

⁷<https://github.com/jpiskorski/eu-acquis-ner>

Ethics and broader impact

Intended Use The corpora created in the context of the presented work were specifically designed to advance research on entity extraction from legal acts from the *EU acquis*. Although they represent key domains covered by EU law, they should not be considered perfectly balanced in any particular way. Consequently, we recommend responsible use thereof.

Please note that the collection of legal acts annotated with entity mentions do contain named mentions of persons and organizations. It is important to emphasize in this context that the entire collection of the legal acts that form the *EU acquis* is freely accessible to the public in 24 languages of the European Union.

Environmental Impact The deployment of LLMs might have a large carbon footprint, in particular, when training new models. In our endeavor, we have only exploited existing LLMs and we did not train them, which had a relatively low impact in terms of computing.

Fairness The creation of the entity corpus involved the annotation of legal documents, which was done by the staff of the institutions of the authors of this paper. They were fairly remunerated as part of their job.

Limitations

Corpora Representativeness The selection of the legal acts used for the creation of the GOLD and SILVER corpus was done solely using the *EU acquis*, the EU law. Therefore, it might be considered as representative only for this specific subdomain of legal texts. The legal documents contained in the corpora do share certain common vocabulary and syntactic constructions in terms of expressing legal provisions (e.g., permissions, obligations, exemptions, etc.) imposed on or attributed to the entities mentioned in the text. Consequently, we believe, the corpus is representative in terms of the aforementioned linguistic phenomena which are of relevance for developing IE solutions for the *EU acquis* domain, e.g., for the extraction of structured information on legal obligations, where the extraction of entities is of paramount importance.

However, we emphasize that these corpora, although representing key domains covered by EU law, should not be considered as perfectly balanced in any particular way.

While the corpora introduced in the manuscript are of medium size and the class distribution is heavily imbalanced we believe corpora of this size (although suboptimal) constitute a useful resource to support training entity extraction models and provide meaningful insights on IE from the *EU acquis*.

Biases Although a comprehensive annotation guidelines were used for the purpose of creating the human-annotated corpus and multiple rounds of curating the annotations were carried out we are aware that some degree of intrinsic subjectivity might be present in this dataset.

Models While our experiments covered various approaches for the entity extraction task, namely, knowledge-, transformer machine learning-, and generative AI-based ones they should not be considered as exhaustive, but rather to provide an insight on the complexity of the task at hand.

Disclaimer The information and views set out in this publication are those of the authors and do not necessarily reflect the official opinion of the European Commission. The Commission does not guarantee the accuracy of the data included in this study. Neither the Commission nor any person on the Commission’s behalf may be held responsible for the use that may be made of the information contained therein. This document should not be considered as representative of the European Commission’s official position.

Acknowledgments

We are greatly indebted to Guillaume Jacquet and Xavier Chapelant, who contributed significantly to the work reported in this paper. We also thank the entire team of annotators, without whom the work presented in this paper would not have been possible.

References

Hidemberg O. Albuquerque, Rosimeire Costa, Gabriel Silvestre, Ellen Souza, Nádia F. F. Da Silva, Douglas Vitório, Gyovana Moriyama, Lucas Martins, Luiza Soezima, Augusto Nunes, Felipe Siqueira, João P. Tarrega, Joao V. Beinotti, Marcio Dias, Matheus Silva, Miguel Gardini, Vinicius Silva, André C. P. L. F. De Carvalho, and Adriano L. I. Oliveira. 2022. [UlyssesNER-Br: A Corpus of Brazilian Legislative Documents for Named Entity Recognition](#). In Vlória Pinheiro, Pablo Gamallo, Raquel Amaro, Carolina

- Scarton, Fernando Batista, Diego Silva, Catarina Magro, and Hugo Pinto, editors, *Computational Processing of the Portuguese Language*, volume 13208, pages 3–14. Springer International Publishing.
- Hidemberg O. Albuquerque, Ellen Souza, Danilo C. G. Lucena, Héldon J. O. Albuquerque, Nádia F. F. Da Silva, Márcio Dias, Rafael O. Nunes, Adriano L. I. Oliveira, and André C. P. L. F. De Carvalho. 2026. [UlyssesLegalNER-br: from legislative to legal, a comprehensive corpus of Brazilian legal documents for named entity recognition](#). In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 331–341, Salvador, Brazil. Association for Computational Linguistics.
- Farid Ariai, Joel Mackenzie, and Gianluca Demartini. 2025. [Natural language processing for the legal domain: a survey of tasks, datasets, models, and challenges](#). *ACM Computing Surveys*, 58(6):1–37.
- Nancy Chinchor. 1998. Overview of MUC-7/MET-2. In *Proceedings of Seventh Message Understanding Conference (MUC-7)*.
- George R. Doddington, Alexis Mitchell, Mark A. Przybicki, Lance A. Ramshaw, Stephanie Strassel, and Ralph M. Weischedel. 2004. [The Automatic Content Extraction \(ACE\) program—tasks, data, and evaluation](#). In *Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 837–840, Lisbon, Portugal.
- Leila Feddoul, Sarah T. Bachinger, Clara Lachenmaier, Sebastian Apel, Pirmin Karg, Norman Klewer, Denys Forshayt, Robin Erd, and Marianne Mauch. 2024. [GerPS-NER: A dataset for named entity recognition to support public service process creation in Germany](#). In *Proceedings of the Second International Workshop on Semantic Technologies and Deep Learning Models for Scientific, Technical and Legal Data (SemTech4STLD 2024)*, volume 3697, pages 1–10, Hersionissos, Greece. CEUR Workshop Proceedings.
- Jenny Rose Finkel and Christopher D. Manning. 2009. [Nested named entity recognition](#). In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 141–150, Singapore. Association for Computational Linguistics.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing](#). In *The Eleventh International Conference on Learning Representations*.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spacy: Industrial-strength natural language processing in python](#).
- Yining Huang, Keke Tang, and Meilian Chen. 2024. [Leveraging large language models for enhanced nlp task performance through knowledge distillation and optimized training strategies](#). *Preprint*, arXiv:2402.09282.
- Basra Jehangir, Saravanan Radhakrishnan, and Rahul Agarwal. 2023. [A survey on named entity recognition — datasets, tools, and methodologies](#). *Natural Language Processing Journal*, 3:100017.
- Lixia Ji, Yiping Dang, Yunlong Du, Wenzhao Gao, and Han Zhang. 2025. [Nested named entity recognition: A survey of latest research](#). *Expert Systems*, 42(7):e70052.
- Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. 2024. [Recent advances in named entity recognition: A comprehensive survey and comparative study](#). *Preprint*, arXiv:2401.10825.
- Hongjin Kim, Jai-Eun Kim, and Harksoo Kim. 2024. [Exploring nested named entity recognition with large language models: Methods, challenges, and insights](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8653–8670, Miami, Florida, USA. Association for Computational Linguistics.
- Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. [The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation](#). In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.
- Panteleimon Krasadakis, Evangelos Sinos, Vassilios S. Verykios, and Evangelos Sakkopoulos. 2022. [Efficient named entity recognition on Greek legislation](#). In *2022 13th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pages 1–8.
- Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–433.
- Ashutosh Modi, Prathamesh Kalamkar, Saurabh Karn, Aman Tiwari, Abhinav Joshi, Sai Kiran Tanikella, Shouvik Kumar Guha, Sachin Malhan, and Vivek Raghavan. 2023. [SemEval-2023 task 6: LegalEval - understanding legal texts](#). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2362–2374, Toronto, Canada. Association for Computational Linguistics.
- David Nadeau and Satoshi Sekine. 2007. [A survey of named entity recognition and classification](#). *Linguisticae Investigationes*, 30(1):3–26.
- Damien Nouvel, Maud Ehrmann, and Sophie Rosset. 2016. [Named entities for computational linguistics](#). John Wiley & Sons.
- Francielle Vasconcellos Pereira, Ana Frazão, and Viviane P. Moreira. 2024. [Automatic text simplification for the legal domain in Brazilian Portuguese](#). In *Intelligent Systems: 34th Brazilian Conference, BRACIS 2024, Belém Do Pará, Brazil, November 17–21, 2024*,

Proceedings, Part IV, page 31–45, Berlin, Heidelberg. Springer-Verlag.

Damith Premasiri, Tharindu Ranasinghe, Ruslan Mitkov, Mo El-Haj, and Ingo Frommholz. 2025. [Survey on legal information extraction: Current status and open challenges](#). *Knowledge and Information Systems*, 67(12):11287–11358.

Vasile Păiș, Maria Mitrofan, Carol Luca Gasan, Vlad Coneschi, and Alexandru Ianov. 2021. [Named entity recognition in the Romanian legal domain](#). In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 9–18, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matias Rojas, Felipe Bravo-Marquez, and Jocelyn Dunstan. 2022. [Simple yet powerful: An overlooked architecture for nested named entity recognition](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2108–2117, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. [SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts \(DDIExtraction 2013\)](#). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

Dan Shen, Jie Zhang, Guodong Zhou, Jian Su, and Chew-Lim Tan. 2003. [Effective adaptation of hidden Markov model-based named entity recognizer for biomedical domain](#). In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, pages 49–56, Sapporo, Japan. Association for Computational Linguistics.

Tamás Váradi, Svetla Koeva, Martin Yamalov, Marko Tadić, Bálint Sass, Bartłomiej Nitoń, Maciej Ogródniczuk, Piotr Pęzik, Verginica Barbu Mititelu, Radu Ion, Elena Irimia, Maria Mitrofan, Vasile Păiș, Dan Tufis, Radovan Garabík, Simon Krek, Andraz Repar, Matjaž Rihtar, and Janez Brank. 2020. [The MARCELL legislative corpus](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3761–3768, Marseille, France. European Language Resources Association.

Yu Wang, Hanghang Tong, Ziyi Zhu, and Yun Li. 2022. [Nested named entity recognition: A survey](#). *ACM Trans. Knowl. Discov. Data*, 16(6).

Vikas Yadav and Steven Bethard. 2018. [A survey on recent advances in named entity recognition from deep learning models](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. [GLiNER: Generalist model](#)

[for named entity recognition using bidirectional transformer](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

A Related corpora

Table 5 provides an overview of NE-annotated corpora with legislation documents. It distinguishes the corpora by language and jurisdiction, and also by the nature of the annotation task, since some resources provide classic NE annotations, whereas others encode broader legal-semantic categories that go beyond traditional NER. We also distinguish them by the main and broader categories they cover in the annotations and whether they allowed for nested entities or not.

B Entity taxonomy

Table 6 provides a list of all entity types with a short version of their definitions and accompanied by some examples.

C EU acquis legislation documents

EU legislation documents, such as regulations, directives, and decisions, adhere to a highly standardized structure to ensure clarity, legal precision, and consistency across the European Union. These documents typically begin with a preamble, which is divided into two distinct parts: the legal basis and the recitals. The legal basis explicitly cites the relevant articles of the EU Treaties (e.g., ‘[Article 294 of the Treaty on the Functioning of the European Union](#)’) that empower the EU institutions to adopt the act. The recitals, introduced by the term ‘*Whereas*’, provide the context, objectives, and justifications for the legislation, explaining the reasons for its adoption and its expected impact. Following the preamble, the operative part is structured into articles (numbered sequentially), which contain the substantive provisions, obligations, and implementation details. Articles may be further subdivided into paragraphs (e.g., ‘[Paragraph 1](#)’) or points (e.g., ‘*(a)*, ‘*(b)*’) for greater specificity. Annexes are often included to provide technical details, lists, or templates referenced in the main text. The document concludes with final provisions, such as the entry-into-force date, the addressees (e.g., Member States or specific entities), and the signatures of

Dataset	Annotation Type	Main categories	Nested	Language	Jurisdiction
(Váradí et al., 2020)	Classic NER	PER, LOC, ORG, TIME	No	Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak, Slovenian	Bulgaria, Croatia, Hungary, Poland, Romania, Slovakia, Slovenia
(Păiș et al., 2021)	Classic NER + legal ref	PER, LOC, ORG, TIME, LEGAL REF	No	Romanian	Romania
(Krasadakis et al., 2022)	Fine-grained legal NER	LAW, CASELAW, EURO LAW, BOOK	No	Greek	Greece
(Albuquerque et al., 2026)	Fine-grained legal NER	PER, ORG, LOC, LEGAL REF, CASELAW REF, PRODUCT, MONETARY VALUE	No	Brazilian Portuguese	Brazil
(Feddoul et al., 2024)	Legal norm labelling	Actors, actions, conditions, legal basis	Yes	German	Germany

Table 5: Overview of legislation domain datasets for NER and related annotation tasks.

the adopting institutions (e.g., ‘*the European Parliament and the Council*’). References to CELEX numbers (e.g., ‘*32018R0001*’ for ‘*Regulation (EU) 2018/1*’) and Official Journal (OJ) citations (e.g., ‘*OJ L 123, 12.5.2018, p. 1*’) are included for identification and traceability.

TYPE	DEFINITION	EXAMPLES
PER		
PER-IND	Mentions of individual persons, disregarding whether living, dead, real or fictive. This category includes also nominal mentions of individuals	<i>John Smith, president of the European Parliament</i>
PER-GRP	Named and nominal mentions of groups of people. In particular, toponym-related references to groups of people linked to geopolitical entities or geographical zones are covered by this type	<i>Ukrainians, Members of the European Parliament</i>
ORG		
ORG-EU	Mentions of all EU institutions, including European Commission DGs, EU agencies, Council, EP, etc. It includes also nominal mentions of such entities.	<i>DG MARE, JRC, European Commission, EU agencies</i>
ORG-INT	Mentions of international organizations and non-profit institutions, including nominal mentions too.	<i>World Health Organization, European Patent Office, international organization</i>
ORG-INS	Mentions of government, national non-profit, public and political institutions. It includes also nominal mentions of such entities	<i>Polish government, Member States government</i>
ORG-COM	Mentions of profit organizations, companies, corporations, banks, etc.	<i>Bank of Scotland, car companies</i>
LOC		
LOC-GPE	Mentions of Geopolitical Entities (GPE). They are complex entities referring to a geographical area which is associated with some sort of political structure, i.e., consists of a population, a government or some administrative body, and a physical location. For example, countries, provinces, states, counties and cities, and suchlike entities are considered GPEs. Furthermore, groups of countries and supranational unions of countries are also considered GPEs, whereas international organisations (e.g., ones which govern such supranational unions) are not considered GPEs.	<i>Luxemburg, Province of Varese, European Union, rapporteur Member State</i>
LOC-FAC	Facilities are functional and human-made structures created for a particular purpose which have a geographical location and are recognizable from a distance. For instance, tourist attractions (e.g., museums, amusement parks, stadiums, monuments), urban and extra-urban infrastructure (e.g., streets, highways, tunnels, parkings, bridges, hospitals), transportation hubs (e.g., airports, sea ports, etc.), buildings (e.g., churches, municipalities), are considered facilities.	<i>Schiphol Airport, Kyiv City Duma</i>
LOC-LND	Landforms are natural land features (geological formations) on the solid surface of the Earth, which together make up a given terrain. Landforms include hills, mountains, canyons, valleys, plains, plateaus, peaks, mountain ranges, dunes, water bodies (e.g. oceans, seas, lakes, rivers, bays, waterfalls), islands, peninsulas, continents, etc. Human-made land features are not considered landforms (e.g., bridges).	<i>Baltic Sea, Niagara Falls, South and East France</i>
LEG		
LEG-INT	Mentions of international legislation. In this specific context, it includes only regulations, agreements, treaties and other legal resources where at least two different countries are involved AND where NONE of the European Union institutions is a party thereof. International legislation includes any bilateral/multilateral agreement between two/multiple countries, even if these countries are EU Member States always with the condition that NONE of the European Union institutions is a party thereof.	<i>North American Free Trade Agreement, Constitutive Act of the African Union</i>
LEG-EU	All mentions of legislation, including agreement, treaties, regulations where any of the European Union institutions is one of the involved parties. All founding treaties (EURATOM, EU, EEC, etc) are considered as EU Legal Resources. Otherwise, having one or more EU member states as a party of an agreement without any involvement of EU institutions does not fall under EU legal resource.	<i>Mercosur, Schengen Agreement, Council Regulation (EEC, Euratom, ECSC) No 259/68 of 29 February 1968</i>
LEG-NAT	Mentions of any kind of national legislation	<i>Constitution of the Fifth Republic of 4 October 1958</i>
LEG-SUB	Any mention of document/resource sub-section or part	<i>Article 2, Annex III, Paragraph 1.a, Article 2, 3 and 5</i>
IDT		

TYPE	DEFINITION	EXAMPLES
IDT-PUB	Mentions of publication identifiers	<i>ISSN 1977-091X, ISBN 978-951-45-9693-3, OJ L 53, 27.2.0008, p. 52</i>
IDT-POS	Mentions of postal addresses or elements of postal addresses	<i>Via Piccinni 20, 24061 Albano Sant'alessandro, Italy</i>
IDT-GEO	Mentions of geographical coordinates	<i>51°28' 38" N</i>
IDT-PHO	Mentions of phone numbers	<i>+49 60 1234576</i>
IDT-EMA	Mentions of emails	<i>abc@derf.com</i>
IDT-URL	Mentions of URLs	<i>http://www.google.com</i>
IDT-IP	Mentions of IP addresses	<i>255.000.000.000</i>
IDT-VAT	Mentions of VAT numbers	<i>ATU99999999</i>
IDT-BAN	Mentions of bank account numbers, IBANs	<i>DE44 5001 0517 5407 3249 31</i>
IDT-CCN	Mentions of credit card numbers	<i>2345 3523 2453 3453</i>
IDT-PRO	Mentions of product codes such as TARIC codes or CN (Combined Nomenclature) codes	<i>CN codes 0401</i>
IDT-DOS	Mention of a specific EU dossier, such as inter-institutional procedures, as cases from the Court of Justice, Merger cases, State aid cases, Digital Market Act cases, Antitrust and Cartel cases, Foreign Subsidies cases	<i>F-59/09, T-264/11 P, M.0009, SA.00058</i>
IDT-ONL	Mentions of online media-like identifiers	<i>@johnS</i>
PRO		
PRO-SER	Named mentions of services, where a service in this context refers to an instrument, tool, functionality put in place in order to help or facilitate accomplishing certain tasks and processes or to fulfil specific requirements in business of any kind. For instance, IT systems, online platforms, portals, human-operated assistance services, banking instruments, etc., are examples of services. Services are used to implement initiatives covered by a different entity category	<i>Official Journal of the European Union, VIS Information System</i>
PRO-INI	Named mentions of initiatives, i.e., actions and endeavours such as plans, agendas, projects, programmes, strategies, visions, etc., designed to achieve specific goals or to bring about change in the context of any domain	<i>Horizon 2020, Circular Economy Action Plan</i>
PRO-STD	Mentions of concrete standards, where standard is considered a set of agreed-upon and documented guidelines, specifications, accepted practices, technical requirements, or terminologies for diverse fields, which has been introduced by an organization (international or national) dealing with standards, e.g., ISO	<i>API Standard 2543, ASTM D1086-1964</i>
PRO-DOC	Named mention of any kind of documents which do not fall under the LEGAL RESOURCE entity type. This is the case also for some non-legal publications produced by an European institution like JRC Scientific Report or reports produced by European Agencies. General Publication also falls under this category.	<i>Annual Activity Report on Climate Action</i>
EVT		
EVT-MT	Mentions of any type of meetings and gatherings aims at discussions, presenting views and sharing information, including, for instance: parliamentary sessions, conferences, workshops, summits, fora, expositions, debates, seminars, etc.	<i>Member States meeting within the Council of 16 May 1989, ACL 2025 Conference</i>
EVT-CON	Mentions of any type of conflicts and disputes, including, for instance: wars, armed invasions, revolutions, disagreements between countries, coup d'etat, etc.	<i>World War II, Russia-Ukraine war</i>
EVT-IND	Mentions of any type of natural and man-made disasters, including, for instance: pandemics, earthquakes, floods, humanitarian crisis, cyber attacks, assassinations, car and plane crashes, etc.	<i>Chernobyl nuclear accident, COVID pandemic</i>
EVT-OCC	Mentions of an event happening in particular time marked by certain circumstances or occurrences, including, for instance: awards, sport events like world cups, festivals, religious periods or holidays, anniversaries, commemorations, parades, etc.	<i>Fatherland Day, Ramadan, Europe Day on 9 May</i>
TIM		

TYPE	DEFINITION	EXAMPLES
TIM-TM	Explicit mentions of time using a reference to concrete hour/minute	<i>10 o'clock, 20:30 CET</i>
TIM-DAT	Mentions of point in time referring to a concrete date, month or year	<i>1 January 2000, March 2023</i>
TIM-DUR	Mention of a concrete amount (quantity) of time expressed in, for instance, years, months, weeks, days, hours, minutes, etc. The mention can also refer to a specific time span, indicating minimum and maximum amount expressed using some time unit	<i>20 weeks, 1 year, two months</i>
TIM-PER	Mention of a time span with a start and end date	<i>between March and April 2023</i>
TIM-FR	Mentions of frequencies expressed in some time units, e.g., years, months, weeks, or expressions indicating periodicity, even if it is not specified exactly	<i>at least twice a year, at regular intervals, semi-annually</i>
TIM-DE	Mentions of deadlines which include references to concrete dates or point in time, or expressions and phrases indicating a deadline that provides the sense of urgency, even if it is somewhat vague	<i>by 1 May 2020, at the latest within six months, immediately</i>
NUM		
NUM-EX	Mentions of any numbers, including natural and rational numbers, percentages, written both using digits and words. Combinations of numbers and operators are not considered to fall under this category, although some elements they comprise of do so. Ordinal numbers are out of scope.	<i>12 000, two hundred, 30%, one third</i>
NUM-CU	Mentions of numerical expressions that refer to an amount of money in an explicitly mentioned currency. The numerical values can be expressed both using numbers and words. Mentions of the currency itself alone does not fall in this category	<i>2.50\$, two dollars, 23K Euro</i>
NUM-AG	Mentions of an age of a person, object or an artefact	<i>20 years old, 10-years old</i>
NUM-ME	Mentions of any type of measurements, including, but not limited to: physical extent, space, volume, weight, temperature, speed, intensity, pressure, lighting, intensity, electrical power, etc. References to durations fall under a different category, namely, TIM-DUR	<i>10 m3, thirty watts, 60 km/h</i>
NUM-CT	Mentions of quantities of persons, living organisms, locations, organizations, events and objects, like for instance, products, etc.	<i>20 persons, hundred computers, 5 meetings</i>

Table 6: Entity Type Definitions and Examples.

D GOLD corpus details

This Section provides additional statistical data about the GOLD corpus. The overall corpus statistics for fine-grained entities are provided in Table 7. The visualisation of the distribution of the fine-grained entity types is also provided in Figure 3. The distribution of entity mention frequency is provided in Figure 4, whereas the top-10 most frequent entities are provided in Table 8.

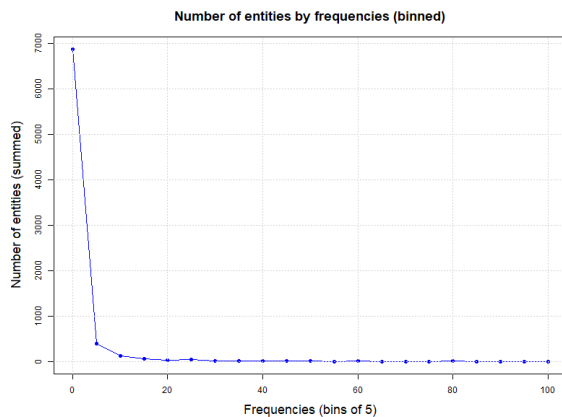
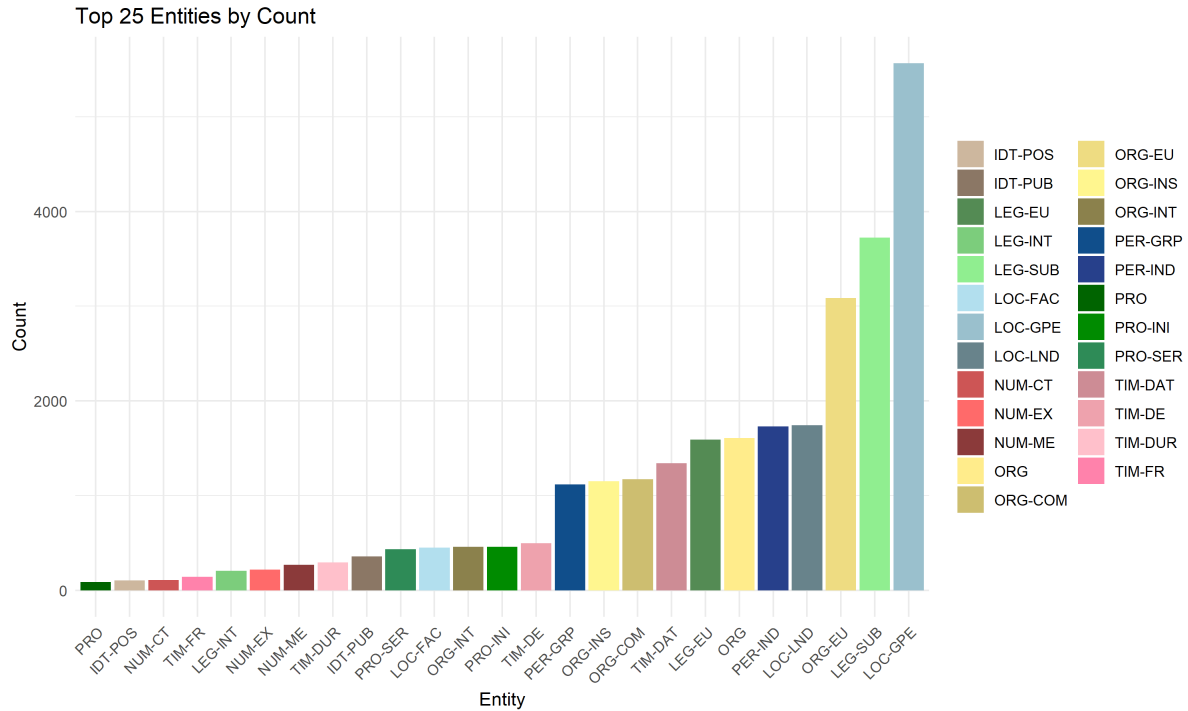
In Table 9 we provide examples of surface forms that have been labeled differently depending on the context. For each entity we provide its frequency and the nature of the ambiguity, distinguishing between *subtype* ambiguities (entities classified under subcategories of the same broader type) and *cross-type* ambiguities (entities spanning entirely different types, such as locations vs. organizations or persons vs. dates). Subtype ambiguities are more prevalent, particularly among nominal mentions of organizations, e.g., ‘*entity*’, ‘*competent authorities*’, and ‘*parties*’, where the distinction often hinges on granular classifications (e.g., generic organizations vs. commercial or institutional subtypes). For example, ‘*credit institutions*’ are overwhelmingly labelled as commercial organizations (97.5%), while ‘*Management Board*’ is almost entirely labelled as a generic organization (96.15%). In contrast, cross-type ambiguities, though less frequent, reveal more complex semantic challenges: ‘*EEA*’ can refer to a geopolitical entity (63.16%, acronym for ‘*European Economic Area*’) and an EU organization (36.84%, acronym for ‘*European Environment Agency*’), while ‘*March*’ oscillates between a location (66.67%) and a date (33.33%). Other notable cross-type cases include ‘*applicant*’, which is primarily an individual (73.08%) but sometimes can refer to an organization (26.92%), and ‘*candidates*’, where the majority of the mentions refer to geopolitical entities (88.89%) rather than individuals. These examples highlight, on one hand, the need of precise annotation rules, and, on the other hand, the complexity of the NE recognition task at hand.

D.1 Annotation Format

The entire corpus is available as a single UTF-8 encoded file, where each line represents a single entity mention and contains the following tab-separated elements: document (file) name, text span that contains the entity mention, start and end character offsets, entity type, and a boolean flag indicating

TYPE	NUMBER (%)	NOMINAL (%)
PER-IND	1732 (6.1%)	1265 (73.0%)
PER-GRP	1117 (3.9%)	1084 (97.1%)
ORG	1609 (5.6%)	1382 (85.9%)
ORG-COM	1175 (4.1%)	868 (73.9%)
ORG-EU	3067 (10.8%)	319 (10.4%)
ORG-INS	1153 (4.0%)	1020 (88.5%)
ORG-INT	459 (1.6%)	78 (17.0%)
LOC	16 (.1%)	-
LOC-FAC	453 (1.6%)	-
LOC-GPE	5584 (19.6%)	1733 (31.0%)
LOC-LND	1745 (6.1%)	-
LEG	5 (.02%)	-
LEG-EU	1591 (5.6%)	-
LEG-INT	208 (.7%)	-
LEG-NAT	61 (.2%)	-
LEG-SUB	3725 (13.1%)	-
IDT	31 (.1%)	-
IDT-DOS	34 (.1%)	-
IDT-EMA	3 (.01%)	-
IDT-GEO	29 (.1%)	-
IDT-PHO	53 (.2%)	-
IDT-POS	105 (.4%)	-
IDT-PRO	16 (.1%)	-
IDT-PUB	359 (1.3%)	-
IDT-URL	47 (.2%)	-
PRO	88 (.3%)	-
PRO-DOC	33 (.1%)	-
PRO-INI	462 (1.6%)	-
PRO-SER	435 (1.5%)	-
PRO-STD	25 (.1%)	-
EVT-CON	1 (.001%)	-
EVT-IND	68 (.2%)	-
EVT-MT	9 (.03%)	-
TIM-DAT	1342 (4.7%)	-
TIM-DE	497 (1.7%)	39 (7.8%)
TIM-DUR	294 (1.0%)	-
TIM-FR	144 (.5%)	9 (6.2%)
TIM-PER	61 (.2%)	-
TIM-TM	5 (.02%)	-
NUM-CT	110 (.4%)	-
NUM-CU	43 (.2%)	-
NUM-EX	220 (.8%)	-
NUM-ME	269 (.9%)	-
ALL	28483	7797 (27.4%)

Table 7: Overall GOLD corpus statistics for the fine-grained entity types. The columns NUMBER and NOMINAL provide the total number of instances per entity type and nominal mentions within each entity class. The percentage figures (in brackets) in the NOMINAL column refer to the fraction of entity mentions within a given entity class which are nominal.



ENTITY	FREQUENCY
Commission	997
Union	938
Member States	866
Council	567
Member State	528
European Parliament	369
EU	290
paragraph 1	237
Belarus	229
Agency	219

Table 8: Top 10 Most Frequent Entities.

whether the entity mention is nominal. Some examples are provided below.

```
eu_acquis_40.txt Regulation (EC) No 726/2004 7539 7565 LEG-EU false
eu_acquis_40.txt EC 7551 7552 LOC-GPE false
eu_acquis_40.txt Swiss COVID-19 Certificate Order 7690 7721 LEG-NAT false
eu_acquis_40.txt COVID-19 7696 7703 EVT-IND false
eu_acquis_40.txt Member States 7724 7736 LOC-GPE true
```

E Evaluation details

Apart from exact matches of entity spans and entity types (*Strict* evaluation), we transpose the criteria introduced by Rojas et al. (2022) to the taxonomy introduced by Segura-Bedmar et al. (2013) to allow for partial offset matches and compensation/penalty for recognition of nested entity mentions. To this

end, their *flat* and *nested* criteria should be considered *Strict* evaluation at a given nesting level. In our evaluation approach, we also consider partial offset matches (with a correct entity type) as in the *Type* evaluation regime (Segura-Bedmar et al., 2013). Moreover, we define a *composite* criterion modifying the additivity of the *nested* score by Rojas et al. (2022).

E.0.1 Basic Notions

We introduce some basic notions. Let $W = (w_1, w_2, \dots, w_k)$ be a text consisting of a sequence of k tokens. An entity mention is defined as a tuple $x_{true} = (p_s, p_e, t_x)$, where $s, e \in [1, n]$ rep-

Entity	Count	Distinct Types with Percentages	Ambiguity type
entity	216	ORG(99.07%)/ORG-INS(0.46%)/ORG-EU(0.46%)	Subtype
competent authorities	198	ORG-INS(98.99%)/ORG(1.01%)	Subtype
body	173	ORG(91.91%)/ORG-INS(8.09%)	Subtype
bodies	92	ORG (90.22%) / ORG-INS (3.26%) / ORG-EU (6.52%)	Subtype
parties	46	ORG (30.43%) / ORG-COM (69.57%)	Subtype
applicant	26	ORG (26.92%) / PER-IND (73.08%)	Cross-Type
EEA	19	LOC-GPE (63.16%) / ORG-EU (36.84%)	Cross-Type
group	16	ORG (68.75%) / ORG-COM (6.25%) / PER-GRP (25%)	Sub/Cross-Type
stakeholders	16	ORG (62.5%) / PER-GRP (37.5%)	Cross-Type
participants	15	PER-GRP (93.33%) / ORG-COM (6.67%)	Cross-Type
third parties	13	ORG (76.92%) / ORG-COM (23.08%)	Subtype
candidates	9	ORG (11.11%) / LOC-GPE (88.89%)	Cross-Type
March	3	LOC-LND (66.67%) / TIM-DAT (33.33%)	Cross-Type

Table 9: Selection of ambiguous entities in the GOLD corpus. The columns show the entity, its count, and the distinct entity types it has been assigned accompanied with the distribution expressed in percentages.

represent the character boundaries of the mention in W , and $t_x \in \Omega$ denotes the type of the entity in a given entity type taxonomy Ω . $|x_{true}| = p_e - p_s$ denotes the *length* of an entity mention. For the purpose of our span-based evaluation, we enumerate all characters c within an entity mention x_{true} to obtain a non-empty sequence of character tuples: $x_{true} = \{(p_1, c_1), \dots, (p_n, c_n)\}$. Assume all sets of character tuples are strictly ordered by their position index p such that $p_1 < p_2 < \dots < p_n$, allowing them to be treated as sequences for string-matching algorithms.

E.0.2 Evaluation Approach

Given two entity mentions x_{true} and $y_{true} = (r_s, r_e, t_y)$ we say that x_{true} is *nested* in y_{true} if and only if $r_s \leq p_s$ and $p_e \leq r_e$ and $|x_{true}| < |y_{true}|$. We define performance criteria according to the approach introduced by Rojas et al. (2022). In a given corpus of entity mentions $X = \{x_1, \dots, x_m\}$ we say that an entity mention x_i is *flat* if there is no $x_j \in X \setminus \{x_i\}$ so that x_i is nested in x_j . An entity mention x_i is called *nested* if there exists an entity mention $x_j \in X \setminus \{x_i\}$ so that x_i is nested (contained) in x_j or x_j is nested in x_i . All nested mentions are subdivided into: (a) *outer* mentions, i.e., the ones that contain other nested entity mentions, but are not nested within any other entity mention, and (b) *inner* mentions, i.e., the ones that are nested (contained) within other entity mentions. An entity mention is *top-level* mention if it is a flat or outer mention.

For the *Type* evaluation, we defined the composite score which is a weighted average of *Type* scores of inner entity mentions and the *Type* score for the remaining characters within the outer entity

mention. Weights are ratios of the inner entities' character lengths to the total character length of the outer entity. The weight of the outer entity is the reminder after dropping all characters associated with inner entities from the levels below. Hence, the composite score can be defined recursively. To do so, we first extend the aforementioned definition of an entity mention.

At a given level of nesting depth, let there be b mutually disjoint inner entities. An i -th inner entity x_{true}^{in} be a sequence of characters, such that $x_{true}^{in} = \{(p_1^{in}, c_1^{in}), \dots, (p_{k_i}^{in}, c_{k_i}^{in})\}$. Then

$$x_{true}^{out} = x_{true} \setminus \bigcup_{i=1}^b x_{true}^{in}. \quad (1)$$

Also, let x_{pred}^{out} be a corresponding predicted outer entity mention, such that

$$x_{pred}^{out} = x_{pred} \setminus \bigcup_{i=1}^b x_{true}^{in} \quad (2)$$

Note that for the purpose of the evaluation, we mask out the true inner entities in the predicted outer entities as they are scored at lower nesting depths. This is to ensure that the outer entity similarity score (Eq. 4) excludes true inner entities and avoids counting inner entities at each nesting level twice.

Next, let the similarity function be such that

$$sim(x_{true}, x_{pred}) = \begin{cases} 1 & \text{if } \frac{|LCS(x_{true}, x_{pred})|}{\max(|x_{true}|, |x_{pred}|)} \geq \tau \\ 0 & \text{if } \frac{|LCS(x_{true}, x_{pred})|}{\max(|x_{true}|, |x_{pred}|)} < \tau \\ 0 & \text{if } \max(|x_{true}|, |x_{pred}|) = 0, \end{cases} \quad (3)$$

where $|LCS(\cdot, \cdot)|$ denotes the length of the longest common subsequence, and $\tau \in (0, 1]$ is the minimum acceptable normalized LCS threshold. Then the character-weighted average of similarity functions (or the composite metric) can be recursively defined as,

$$m_{comp}(x_{true}, x_{pred}) = \frac{|x_{true}^{out}|}{|x_{true}|} * sim(x_{true}^{out}, x_{pred}^{out}) + \sum_{i=1}^b \frac{|x_{true}^{in_i}|}{|x_{true}|} * m_{comp}(x_{true}^{in_i}, x_{pred}^{in_i}), \quad (4)$$

where b is the number of direct inner entities at a given nesting level. If a true inner entity has no corresponding prediction, then $x_{pred}^{in_i} = \emptyset$, which results in $|x_{pred}^{in_i}| = 0$, leading the $|LCS|$ in Equation 3 to equal 0.

Our composite metric allows for additional compensation/penalization of outer entities for correct/incorrect recognition of the nested entities contained within.

F Models details

F.1 Knowledge-based approach

The knowledge-based approach relies on lexical resources: (a) named entities derived from Authority Tables, and (b) trigger words obtained through word n-gram frequency analysis of the *EU acquis* (excluding the documents of the corpus presented in this paper). We summarize the lexical resources in Table 10.

F.2 Transformer

We fine-tuned the gliner-v2.1 on each fold’s training set with the parameters specified by its authors⁸. Before that, we pre-processed our corpus to match GLiNER’s input format⁹.

We executed the fine-tuning on one epoch. Within the training set (32 documents of each fold), we randomly split (80-20) for training and validation and observed both losses to avoid overfitting. We recorded the training and validation loss every 100 steps and selected the model with the lowest validation loss. The remaining 8 documents were kept exclusively for evaluation.

⁸<https://github.com/urchade/GLiNER/blob/main/examples/finetune.ipynb>

⁹<https://huggingface.co/datasets/urchade/synthetic-pii-ner-mistral-v1/resolve/main/data.json>

RESOURCE	NUMBER	MULTILINGUAL
organization triggers	966	✓
EU/int organizations	38542	✓
national courts	683	✓
other organizations	32234	✓
first names	26473	✓
surnames	88613	✓
person demonyms	551	
person name triggers	1438	
countries	16131	✓
country adjectives	292	
cities	535	
other geopolitical entities	303	✓
gpe triggers	240	
continents	97	✓
landform triggers	189	
landforms	215	
facility triggers	370	
legislation act names	150	
legal reference triggers	270	
subdivision triggers	89	
identifier triggers	425	
product triggers	1412	
products	68	
services/initiatives	8179	✓
event triggers	138	
temporal triggers	538	
currency triggers	143	✓
measurement triggers	171	
numerical triggers	595	

Table 10: Lexical resources exploited by the Knowledge-based approach. The column MULTILINGUAL indicates whether the resources are multilingual, i.e., cover all EU languages.

```

learning_rate=5e-6,
bf16=True,
others_lr=1e-5,
lr_scheduler_type="linear",
gradient_accumulation_steps=4,
warmup_ratio=0.03,
per_device_train_batch_size=8,
per_device_eval_batch_size=8,
focal_loss_alpha=0.75,
focal_loss_gamma=2,
num_train_epochs=1,
save_steps = 100,
save_strategy="steps",
eval_strategy="steps",
eval_steps=100,
load_best_model_at_end=False,
metric_for_best_model="loss",
greater_is_better=False,
save_total_limit=20,
use_cpu = False

```

Figure 5: GLiNER hyperparameters

When fine-tuning with the SILVER corpus, we used the constant scheduler and set *focal-loss-alpha* hyperparameter to .4 and also increased the warm up ratio to .1.

To fine-tune **GLiNER**, we used a single GPU instance (NVIDIA A10G Tensor Core) with 24GB of GPU memory. We spent the total amount of approx. 60 GPU-hours for tuning, training, and inference on the GOLD and SILVER sub-corpora.

F.3 Generative AI

Figure 6 shows the template of the prompt passed to the LLM. The total cost for the processing of GOLD corpus was approx. USD250 (multiple runs), whereas the processing of the SILVER corpus cost approx. USD500 (single run).

G Additional results

Figure 11 provides evaluation results per fine-grained types (all mentions - STRICT only) for best models in each model class.

The evaluation results of **GLiNER** fine-tuned on the SILVER corpus are provided in Table 12.

A comparison of recall results for nominal and named entity mentions for fine- and coarse-grained entity types is provided in Table 13 and 14 respectively.

H SILVER corpus details

H.1 Document Selection

The SILVER corpus targets EU legislation (from selected legal sectors) that is currently in force while ensuring that responsible agents are represented across the extended corpus (as in the GOLD corpus). We analyzed the metadata of nearly 46k EU

System Prompt:

You are an Entity Recognition assistant for EU legislation. Given the input string from the user, use the `extract_entities` tool to extract named and nominal mentions of entities for each entity type in the tool according to its description. Consider there might be nested entities, where one entity contains another. There are also nominal entities, such as, operator, entity, body, natural person, etc. Generally, when the context does not indicate whether a nominal mention is a person or an organisation (operator, person, party, etc.), it should be considered as an organisation. The word 'entity' itself should be recognized as an organisation. The extraction lists should be one dimensional. If no instance of a given entity type is present in the text, return an empty array.

Here are some examples of correct extractions:
<examples>/DOCUMENT #40/<examples>

Configuration of the `extract_entities` tool:

```

{
  "tools": [
    {
      "toolSpec": {
        "name": "extract_entities",
        "inputSchema": {
          "json": {
            "type": "object",
            "properties": {
              "/ALL ENTITY TYPES/": {
                "type": "array",
                "description": "/CODEBOOK DESCRIPTION/",
                "items": {
                  "type": "string"
                }
              }
            }
          },
          "required": [
            "/ALL ENTITY TYPES/"
          ]
        }
      },
      "description": "Extracts a list of named and nominal mentions of entities from input string for each entity type, strictly following the annotation guidelines in the taxonomy. In case of uncertainty whether a given text fragment is a mention of an entity, it takes the conservative approach, i.e., the mentions are not included as an instance in the type. When assigning the type to an entity mention in general, it assumes that the local document context, and common knowledge is considered and exploited for resolving ambiguities. It does not extract pronominal mentions of entities (i.e., pronouns referring to entities, such as it, he, they, etc.). They shall not be annotated. In case of uncertainty on which subtype of a main entity type to choose (e.g. a mention of an organization which could potentially be an international or national one), the mention should be included only in the main type. If it is clear that no subtype of the main entity type matches an entity that is being considered, the tool appends the considered entity only to the main entity type."
    }
  ],
  "toolChoice": {
    "tool": {
      "name": "extract_entities"
    }
  }
}

```

User input:
/INPUT TEXT/

Figure 6: Prompt Template (Few Shot)

Type	KB			gliner(FT)			Claude(FS+DI)		
	P	R	F_1	P	R	F_1	P	R	F_1
PER	-	-	-	.000	-	-	.000	-	-
PER-IND	.258	.203	.238	.695	.335	.451	.761	.325	.456
PER-GRP	.491	.389	.434	.624	.636	.630	.400	.536	.458
ORG	.200	.397	.266	.522	.533	.528	.445	.190	.266
ORG-COM	.331	.263	.293	.565	.692	.622	.696	.323	.441
ORG-EU	.928	.701	.799	.844	.855	.850	.800	.873	.835
ORG-INS	.613	.544	.576	.634	.708	.669	.612	.543	.575
ORG-INT	.855	.437	.578	.787	.734	.760	.798	.675	.731
LOC	-	.000	-	.008	.688	.016	-	.000	-
LOC-FAC	.865	.481	.618	.359	.600	.449	.472	.634	.541
LOC-GPE	.800	.749	.774	.862	.691	.767	.896	.699	.785
LOC-LND	.743	.391	.513	.737	.668	.701	.805	.742	.772
LEG	-	.000	-	.000	.000	.000	-	.000	-
LEG-EU	.954	.758	.845	.662	.440	.528	.701	.763	.731
LEG-INT	-	.000	-	.420	.404	.412	.535	.442	.484
LEG-SUB	.981	.960	.970	.938	.946	.942	.864	.858	.861
LEG-NAT	-	.000	-	.228	.377	.284	.744	.674	.707
IDT	1.00	.194	.325	.027	.387	.050	.000	.000	.000
IDT-DOS	.870	.588	.702	.000	.000	.000	.556	.735	.633
IDT-EMA	1.00	1.00	1.00	.400	.667	.500	1.00	1.00	1.00
IDT-GEO	.783	.621	.692	.148	.931	.256	.377	.793	.511
IDT-PHO	.769	.566	.652	-	.000	-	.967	.547	.699
IDT-POS	.815	.210	.333	.000	.000	.000	.589	.819	.685
IDT-PRO	1.000	1.000	1.000	-	.000	-	.265	.563	.360
IDT-PUB	.548	.531	.540	.455	.028	.053	.847	.974	.906
IDT-URL	.900	.574	.701	1.00	.021	.041	.887	1.000	.940
PRO	.600	.045	.083	.005	.102	.010	.000	.000	.000
PRO-DOC	-	.000	-	.005	.030	.009	.015	.278	.029
PRO-INI	.674	.139	.230	.328	.403	.362	.587	.481	.529
PRO-SER	.907	.409	.564	.067	.090	.077	.489	.648	.556
PRO-STD	.318	.280	.298	.027	.120	.004	.387	.480	.429
EVT	-	-	-	.000	-	-	-	-	-
EVT-CON	.000	.000	.000	.000	.000	.000	.000	.000	.000
EVT-IND	.000	.000	.000	.077	.015	.025	.111	.750	.194
EVT-MT	.035	.222	.061	.100	.333	.154	.021	.111	.036
TIM-DAT	.957	.490	.648	.911	.887	.899	.907	.656	.762
TIM-DE	.475	.543	.507	.514	.400	.450	.604	.400	.481
TIM-FR	.543	.708	.614	.489	.611	.543	.640	.604	.621
TIM-DUR	.868	.853	.861	.773	.371	.501	.836	.592	.693
TIM-PER	.621	.295	.400	.163	.279	.206	.500	.574	.534
TIM-TIM	.000	.000	.000	1.00	.100	.189	1.00	.400	.571
NUM	-	-	-	.000	-	-	-	-	-
NUM-CUR	1.00	.837	.911	.114	.209	.148	.911	.953	.932
NUM-CT	-	.000	-	.114	.209	.147	.459	.355	.400
NUM-EX	.135	.759	.229	.543	.318	.401	.202	.482	.285
NUM-ME	.671	.424	.519	.222	.186	.202	.835	.491	.618

Table 11: Evaluation for per fine-grained types (all mentions - STRICT only) for selected models in terms of precision (P), recall (R) and F_1 with: FS - few shot, DI - dynamic input, FT - fine-tuned.

Type	gliner(FTS)		
	P	R	F_1
PER	.000	-	-
PER-IND	.642	.696	.668(↑)
PER-GRP	.563	.672	.613(↓)
ORG	.394	.055	.097(↓)
ORG-COM	.477	.701	.568(↓)
ORG-EU	.869	.891	.880(↑)
ORG-INS	.541	.758	.631(↓)
ORG-INT	.764	.762	.763(↑)
LOC	.000	.000	.000(↓)
LOC-FAC	.288	.545	.377(↓)
LOC-GPE	.831	.838	.835(↑)
LOC-LND	.717	.755	.735(↑)
LEG	.000	.000	.000
LEG-EU	.730	.481	.580(↑)
LEG-INT	.397	.500	.443(↑)
LEG-SUB	.949	.961	.955 (↑)
LEG-NAT	.169	.721	.274(↓)
IDT	-	.000	-
IDT-DOS	.370	.588	.455(↑)
IDT-EMA	.750	1.00	.857(↑)
IDT-GEO	.472	.862	.610(↑)
IDT-PHO	.810	.642	.716(↑)
IDT-POS	.055	.048	.051(↑)
IDT-PRO	.005	.063	.009(↑)
IDT-PUB	1.00	.071	.133 (↑)
IDT-URL	1.00	.234	.379 (↑)
PRO	.000	.000	.000(↓)
PRO-DOC	.001	.056	.002(↓)
PRO-INI	.311	.413	.355(↓)
PRO-SER	.319	.384	.348(↑)
PRO-STD	.094	.320	.145 (↑)
EVT	.000	-	-
EVT-CON	.000	.000	.000
EVT-IND	.018	.500	.034 (↑)
EVT-MT	.036	.222	.063(↓)
TIM-DAT	.805	.921	.859(↓)
TIM-DE	.474	.147	.225(↓)
TIM-FR	.390	.639	.484 (↑)
TIM-DUR	.785	.853	.833(↑)
TIM-PER	.487	.311	.380(↑)
TIM-TIM	.000	.000	.000(↓)
NUM	.000	-	-
NUM-CUR	.262	.837	.400(↑)
NUM-CT	.032	.082	.046 (↓)
NUM-EX	.088	.809	.159(↓)
NUM-ME	.620	.480	.541 (↑)

Table 12: Evaluation for per fine-grained types (all mentions - STRICT only) for gliner fine-tuned on SILVER corpus (FTS) in terms of precision (P), recall (R) and F_1 . The arrows (↑↓) indicate increase/decrease compared to fine-tuning only on the GOLD corpus (FT).

Model	Recall (nominal)	Recall (named)
top-level mentions (w/o comp/penalty)		
KB	.419(.094)	.656(.084)
gliner(FT)	.660(.080)	.659(.073)
gliner(FTS)	.643(.041)	.725(.036)
Claude(FS+DI)	.430(.069)	.759(.041)
Claude+KB	.435(.071)	.771(.043)
all mentions (w/ compensation/penalty)		
KB	.420(.094)	.636(.073)
gliner(FT)	.662(.082)	.654(.073)
gliner(FTS)	.645(.043)	.730(.032)
Claude(FS+DI)	.435 (.073)	.718(.024)
Claude+KB	.439(.074)	.738(.027)
only nested (w/ compensation/penalty)		
KB	.371(.131)	.547(.071)
gliner(FT)	.421(.073)	.505(.063)
gliner(FTS)	.409(.124)	.556(.024)
Claude(FS+DI)	.534(.152)	.572 (.038)
Claude+KB	.498(.155)	.619(.029)

Table 13: Recall for fine-grained types in *Strict* evaluation: *nominal* (LEFT) and *named* (RIGHT). Standard deviations are given in brackets. ZS - zero shot, FS - few shot, FT - fine-tuned, FTS -fine-tuned on SILVER, DI - dynamic input.

Model	Recall (nominal)	Recall (named)
top-level mentions (w/o comp/penalty)		
KB	.500(.106)	.671(.087)
gliner(FT)	.795(.062)	.763(.017)
gliner(FTS)	.758(.046)	.757(.025)
Claude(FS+DI)	.748(.013)	.790(.022)
Claude+KB	.510(.056)	.794(.030)
all mentions (w/ compensation/penalty)		
KB	.496(.106)	.649(.075)
gliner(FT)	.796(.063)	.752 (.024)
gliner(FTS)	.759(.048)	.758(.024)
Claude(FS+DI)	.523 (.075)	.748(.013)
Claude+KB	.513(.059)	.758(.015)
only nested (w/ compensation/penalty)		
KB	.408(.112)	.551(.070)
gliner(FT)	.446(.062)	.548(.026)
gliner(FTS)	.410(.124)	.563(.020)
Claude(FS+DI)	.562(.128)	.584(.042)
Claude+KB	.526(.132)	.629(.030)

Table 14: Recall for coarse-grained types in *Strict* evaluation: *nominal* (LEFT) and *named* (RIGHT). Standard deviations are given in brackets. ZS - zero shot, FS - few shot, FT - fine-tuned, FTS -fine-tuned on SILVER, DI - dynamic input.

legal acts in force. We excluded corrigenda from the query and removed duplicate rows from the dataset (amendments, corrigenda, etc.), resulting in a dataset of over 22k legislative acts. The fields of interest were the year of publication, act type, thematic descriptors (EuroVoc terms¹⁰, subject matter, and directory codes), and the responsible agent.

To execute the stratified sampling procedure, we first pre-processed the metadata by: (1) imputing missing responsible agents, (2) merging main act types into 3 categories (decision, regulation, directive), (3) considering only the first author of the legal act, and (4) assigning the acts to 5-year time windows. We also organized the legal acts hierarchically according to their metadata. The hierarchy reflected our prioritization to represent legal acts through their responsible agents' along with legal act authors, legal act type, and publication year, in this order. If a leaf group had less than 6 legal acts, we combined all the groups at the level and assigned a shared label, i.e., 'all periods', 'multiple act types', or 'multiple authors'. This work resulted in a metadata tree with 540 groups at the leaf nodes. We drew a stratified sample of 540 legal acts from the target dataset, so that each stratum had at least 1 legal act in the SILVER corpus. The resulting corpus represents the target legal acts according to their characteristics expressed with metadata.

H.1.1 Imputation of Responsible Agent

Since only 40% of this defined population of legal acts had a currently used agent label assigned, we used thematic descriptors of labelled legal acts (EuroVoc terms, subject matter and directory codes) to impute the responsible agents for the remaining legal acts. We collapsed least frequent responsible agents into an 'other' class (< 1% of the total).

We used a stratified 5-fold validation on the remaining 9237 legal acts for the responsible agent classes. The average accuracy across the 5 folds was .963. The macro averages of Precision, Recall, and F1-scores were respectively .908, .910, and .905.

We imputed the missing values in the *responsible_agent* metadata field. As a final verification step, we manually inspected 170 of the inferred agents (selected at random) and observed .91 accuracy compared to human annotations. This confirmed that using controlled vocabularies (EuroVoc,

¹⁰<https://op.europa.eu/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/eurovoc>

directory codes, and subject matter topics) helped effectively fill in the missing values for the *responsible_agent* metadata field.

We also considered synthetic data augmentation by (1) prompting LLMs (Mistral-Large and Claude-Sonnet-3.5) to create hypothetical titles of legal acts based on the descriptions of each DGs mission and responsibilities (available on European Commission’s website) and (2) generating the thematic descriptors (EuroVoc terms, subject matter, and directory codes). We observed a decrease in the performance metrics and decided not to proceed with the LLM metadata augmentation.

H.2 Statistics

SILVER corpus statistics for the fine-grained entity types, are provided in Table 15.

I Annotation Guidelines

This section provides an excerpt from the annotations guidelines

I.1 General Guidelines

We provide here the general annotation rules:

- the scope of the annotation covers primarily named mentions of entities, including: persons, organisations, locations, products, identifiers, legal references, temporal expressions, events and numerical expressions,
- in addition to named mentions, in certain specific cases also nominal mentions of entities are to be annotated, more specifically for: (a) persons, (b) organisations, (c) locations, but only for geo-political entities, and (d) some type of temporal expressions (i.e., deadlines and frequencies),
- pronominal mentions of entities shall not be annotated (i.e., pronouns referring to entities),
- when assigning the type to a named-entity mention in general, it is assumed that the local document context and common knowledge is considered and exploited for resolving ambiguities,
- in case of uncertainty whether a given text fragment constitutes a mention of an entity a **conservative approach** should be applied, i.e., such a text fragment should not be annotated, or alternatively discuss the specific case with other annotators,

TYPE	NUMBER (%)
PER-IND	3812 (19.2%)
PER-GRP	16009 (80.8%)
ORG	1143 (2.6%)
ORG-COM	5455 (12.4%)
ORG-EU	27251 (61.8%)
ORG-INS	7525 (17.1%)
ORG-INT	2689 (6.1%)
LOC	0 (0%)
LOC-FAC	150 (.4%)
LOC-GPE	36628 (95.4%)
LOC-LND	1614 (4.2%)
LEG	0 (0%)
LEG-EU	12532 (26.6%)
LEG-INT	1631 (3.4%)
LEG-NAT	827 (1.8%)
LEG-SUB	32172 (68.2%)
IDT	2 (.01%)
IDT-DOS	608 (8.9%)
IDT-EMA	18 (.3%)
IDT-GEO	403 (5.9%)
IDT-PHO	185 (2.7%)
IDT-POS	369 (5.4%)
IDT-PRO	46 (6.7%)
IDT-PUB	4676 (68.6%)
IDT-URL	514 (7.5%)
PRO	60 (.6%)
PRO-DOC	3972 (40.0%)
PRO-INI	4236 (42.6%)
PRO-SER	1192 (12.0%)
PRO-STD	480 (4.8%)
EVT-CON	57 (7.4%)
EVT-IND	422 (54.9%)
EVT-MT	290 (37.7%)
TIM-DAT	15660 (59.2%)
TIM-DE	5504 (20.8%)
TIM-DUR	2214 (8.4%)
TIM-FR	1322 (5.0%)
TIM-PER	1739 (6.6%)
TIM-TM	14 (.01%)
NUM-CT	1015 (1.2%)
NUM-CU	1312 (1.6%)
NUM-EX	76689 (94.8%)
NUM-ME	2709 (3.3%)
ALL	275146

Table 15: Overall SILVER corpus statistics for the fine-grained entity types, also expressed in percentages (in brackets).

- in case of uncertainty which subtype of a main entity type to choose (e.g. a mention of an organization which could potentially be an international or national one), the main entity type (generic type) should be chosen,
- analogously, if it is clear that no subtype of main entity type matches an entity that is being considered, simply the main entity type should be used, e.g., “*entity*” and “*body*” often refer to some organization, but it might be unclear what is their specific type,
- in case of type ambiguities, i.e., different interpretations of the possible type of an entity (e.g. person or organization), specific rules for resolving such ambiguities should be applied (they are provided in the subsequent section),
- in principle, annotations can be nested, for instance, a reference to a date can be part of a mention of a reference to legal documents, e.g., “*Council Regulation (EC) No 1/2003 of 16 December 2002*” contains the mention of “*16 December 2002*”. All nested entity mentions shall be annotated with the exception of the following cases:
 - nested nominal mentions of entities within nominal mentions of entities,
 - nested mentions of numerical expressions contained in numerical expressions,
 - nested mentions of products contained in product mentions,
 - nested mentions of identifiers contained in identifier mentions,
- if the annotated mention of an entity is nominal, the value of the **nominal** attribute in the interface of the annotation tool should be set to true,
- pairs and enumerations of entities are often referred to using nominal elliptical constructions. A nominal elliptical construction is a complex noun phrase in which a noun or nominal group is omitted that is clearly understood from context, making the phrase more concise and avoiding repetition. For instance, in “*European and Asian countries*” the word “*countries*” is omitted in the first constituent of the entire noun phrase. In such cases the entire noun phrase is to be annotated.

I.2 Entity Definitions and Specific Annotation Guidelines

This Section provides detailed definitions for all entity types using the following template:

- **Entity type name:** name of the entity type as it appears in the Taxonomy (and Annotation Tool)
- **Entity type code:** the code used to denote the entity type
- **Definition:** formal definition of the entity type
- **Examples:** positive examples
- **Counter examples:** negative examples
- **Annotation rules:** type-specific annotation rules

In particular, we include here definitions and annotation rules for 5 main entity categories: persons, organizations, locations, legal resources and products to exemplify the granularity of the descriptions in the full guidelines.

I.2.1 Person

This category covers references to individual people and groups of people. Both named and nominal mentions of persons or groups of persons fall under this category.

Individual

Entity type name: PERSON: INDIVIDUAL

Entity type code: PER-IND

Definition: Mentions of individual persons, disregarding whether living, dead, real or fictive. This category includes also nominal mentions of individuals like, for instance, “*president*”.

Examples: “*John Smith*”, “*George W. Bush*”, “*Ursula von der Leyen*”, “*EU officials*”, “*analyst*”, “*lawyer*”, “*representative*”, “*Ukrainian*” (citizen of Ukraine), “*president of the USA*”, “*member of the management board*”

Counter examples: “*He*”, “*They*”

Annotation rules:

- Functions, titles and roles which are mentioned in the context of person mentions should not be annotated, for instance, in the text “*President John Smith said ...*” only “*John Smith*” should be annotated.
- In case a nominal mention of a person appears ambiguous, i.e., it could also potentially refer to an organization or a geo-political entity in the specific context it appears in (e.g. “*operator*”, “*complainant*”, could both refer to an organization) **then the mention should be interpreted as a mention of an organisation or geo-political entity respectively.**

- In case of a nominal mention of a person only “minimal” noun phrase (basic noun phrase) should be annotated, i.e., descriptive parts that provide details should not constitute part of the annotation, e.g., in “*staff involved in carrying out supervisory tasks*” only the phrase “*staff*” should be annotated, analogously, in “*member of the committee for the development of the service*” only the phrase “*member of the committee*” should be annotated.
- In coordinated noun phrases being nominal mentions of person names, each constituent should be annotated separately, e.g., in “*agents and operators*” both “*agents*” and “*operators*” should be annotated separately. If some conjunct in a coordinated noun phrase is incomplete (e.g., in “*security and financial officers*” the “*security*” conjunct is incomplete – the phrase is an elliptical noun phrase), the entire noun phrase is to be annotated.
- Determiners and (possessive) pronouns do not constitute part of the text to be annotated, e.g. in “*their operators*” only “*operators*” should be annotated.

Group

Entity type name: PERSON: GROUP

Entity type code: PER-GRP

Definition: Named and nominal mentions of groups of people. In particular, toponym-related references to groups of people linked to geopolitical entities or geographical zones are covered by this type.

Examples: “*Ukrainians*”, “*Catholics*”, “*Africans*”, “*Polish Citizens*”, “*French football players*”, “*European Parliamentarians*”, “*Members of the European Parliament*”, “*representatives of the Republic of Albania*”, “*operators*” (if operator refers to a person), “*lawyers*”

Counter examples: “*European*”, “*lawyer*”

Annotation rules:

- In principle the same rules as in the case of individual persons apply.
- Mentions of committees, boards, task forces, working groups, networks, commissions, which have formal organizational structures are not considered person groups, however, references to the members of such entities are considered references to person groups. For instance, the text fragment “*Members of the Green Deal Task Force*” should be labelled as person group, while “*the Green Deal Task Force*” should not.

I.2.2 Organization

This category covers organizations, including EU, international, national and commercial organizations. In the context of any type of organization enumerated in this section, mentions of committees, boards, task forces, working groups, commissions, which do have some formal organizational structure, and which are either an internal part of a single organisation or are cross-organizational, are considered as organizations, and should be annotated accordingly.

EU Corporate body

Entity type name: ORGANISATION: EU-BODY

Entity type code: ORG-EU

Definition: Mentions of all EU institutions, including European Commission DGs, EU agencies, etc. It includes also nominal mentions of such entities like “EU agencies” or “EU institutions”.

Examples: “*DG MARE*”, “*JRC*”, “*European Commission*”, “*Frontex*”, “*EMSA*”, “*European Court of Justice*”, “*EU agencies*”, “*EU institutions*”.

Counter examples: “*European countries*”, “*WHO*”, “*National Assembly*”, “*German Government*”, “*European Patent Office*”

Annotation rules:

- Exclude preceding determiners and prepositions (“*the*”, “*in*”, “*for*”, ...) and descriptive phrases preceding or following (e.g., “*in 2004*”, “*of 2004*”), and any qualificative adjective from the mention of an EU body. Examples of text snippets in which an ORG-EU entity mention is highlighted (1) :
“*The new **European Commission** 2024*”
“*M. X works for **JRC Seville**.*”
- In case a nominal mention of an EU body appears ambiguous, i.e., it could also potentially refer to a person in the specific context it appears in, **then the mention should be interpreted as a mention of an EU organisation.**
- In case of a nominal mention of an EU body only “minimal” (simple) noun phrase should be annotated, i.e., descriptive parts that provide details should not constitute part of the annotation, e.g., in “*EU institutions responsible for carrying out these tasks*” only the phrase “*EU institutions*” should be annotated, analogously, in “*target EU institutions*” only the phrase “*EU institution*” should be annotated, however, if a specific term is introduced in text to describe the role of an EU body, then the entire term should be annotated, in “*rapporteur EU institution*” the entire phrase should be annotated.
- In coordinated noun phrases being nominal mentions of EU bodies, each constituent should be annotated separately, e.g., in “*EU agencies and the European Commission*” both “*EU agencies*” and “*European Commission*” should be annotated separately. If some conjunct in a coordinated noun phrase is incomplete (e.g., in “*executive and non-executive EU agencies*” the “*executive*” conjunct is incomplete – this is an elliptical nominal construction), the entire noun phrase is to be annotated.
- If an acronym is introduced in brackets after a mention of an organization name, then it should be annotated separately, e.g., in the text fragment “*European Commission (EC)*” one should annotate “*European Commission*” and “*EC*” separately.

International organisation

Entity type name: ORGANISATION: INTERNATIONAL

Entity type code: ORG-INT

Definition: Mentions of international organizations and non-profit institutions

Examples: “WIPO”, “World Health Organization”, “UNICEF”, “European Patent Office” (not part of EU institutions).

Counter-examples: “European Space Agency”, “National Assembly”, “Coca-Cola”.

Annotation rules:

- Exclude preceding determiners and prepositions (“the”, “in”, “for”, ...) and descriptive phrases preceding or following (e.g., “in 2004”, “of 2004”), and any qualitative adjective from the mention of an international organisation. Examples of text snippets in which an ORG-INT entity mention is highlighted:

“The strong **World Health Organization** in 2024”

“M. X works for **WHO** in Geneva.”

“M. X works for **WHO Geneva**.”

- In case a nominal mention of an international organisation appears ambiguous, i.e., it could also potentially refer to a person in the specific context it appears in, **then the mention should be interpreted as a mention of an international organisation.**
- In case of a nominal mention of an international organisation only “minimal” (simple) noun phrase should be annotated, i.e., descriptive parts that provide details should not constitute part of the annotation, e.g., in “*International organisations responsible for carrying out these tasks*” only the phrase “*International organisations*” should be annotated, analogously, in “*target international organisations*” only the phrase “*international organisations*” should be annotated, however, if a specific term is introduced in text to describe the role of an international organisation, then the entire term should be annotated, in “*rapporteur international organisation*” the entire phrase should be annotated.
- In coordinated noun phrases being nominal mentions of international organisations, each constituent should be annotated separately, e.g., in “*WHO and IOM*” both “*WHO*” and “*IOM*” should be annotated separately. If some conjunct in a coordinated noun phrase is incomplete (elliptical noun construction), the entire noun phrase is to be annotated.

National

Entity type name: ORGANISATION: NATIONAL

Entity type code: ORG-INS

Definition: Mentions of government, national non-profit, public and political institutions. It includes also nominal mentions of such entities like “National Government” or “Car Maker Union”.

Examples: “Polish government”, “French presidency”, “United Auto Workers”, “Member States Governments”, “Car Maker Unions”.

Counter-examples: “European Parliament”, “WHO”, “Volkswagen”.

Annotation rules:

- The annotation rules for EU bodies and International organisations apply here by analogy.

Commercial

Entity type name: ORGANISATION: COMMERCIAL

Entity type code: ORG-COM

Definition: Mentions of profit organizations, companies, corporations, banks, etc.

Examples: “Volkswagen”, “Coca-Cola”, “University of Leuven”, “Bank of Scotland”, “operators” (if operator refers to organisation or if not defined).

Counter-examples: “European Space Agency”, “National Assembly”, “WHO”.

Annotation rules:

- The annotation rules for EU bodies and commercial organisations apply here by analogy.

Additional examples of what to annotate in case of coordination constructions/complex noun phrases.

- “legal person, group or entity”
- “local and regional authorities”
- “branches and subsidiaries”
- “EU institutions, agencies and bodies”
- “EU and international organizations”

In the above list, the phrases in (b), (d) and (e) are elliptical constructions, hence the entire phrase should be annotated. In (a) each of the elements appears to be of different type, hence, each is to be annotated separately. The phrase in (c) does not constitute an ellipsis, hence both of the constituents are to be annotated separately.

I.2.3 Location

Geo-political Entity (GPE)

Entity type name: LOCATION: GPE

Entity type code: LOC-GPE

Definition: Geo-Political Entities (GPE) are complex entities referring to a geographical area which is associated with some sort of political structure, i.e., consists of a population, a government or some administrative body, and a physical location. For example, countries, provinces, states, counties and cities, and suchlike entities are considered GPEs. Furthermore, groups of countries and supranational unions of countries are also considered GPEs, whereas international organisations (e.g., ones which govern such supranational unions) are not considered GPEs.

Examples: “Luxemburg”, “Province of Brescia”, “European Union”, “rapporteur Member State”, “Euro area”, “Euro zone”.

Counter-examples: “European Commission”, “FIFA”, “South Portugal”.

Annotation rules:

- Mentions of landforms that contain mentions of GPEs are not to be annotated as GPEs. For example, the text fragment “South Portugal” refers to a geographical region, hence it should be annotated as a landform. However, in certain cases, the adjectives that might indicate certain geographical region are actually part of an official name of a country, e.g., “West Germany”. In the latter case the text fragment should be annotated as GPE.

- In case a nominal mention of an entity (e.g., “*applicant*”) could refer in some context to both a geopolitical entity and a person **then the mention should be interpreted as a mention of a geo-political entity**.
- In case of a nominal mention of a geopolitical entity only “minimal” noun phrase should be annotated, i.e., descriptive parts that provide details should not constitute part of the annotation, e.g., in “*Member States tasked to carry out this task*” only the phrase “*Member States*” should be annotated, similarly, in “*target Member States*” only the phrase “*Member States*” is to be annotated, however, if a specific term is introduced in text to describe the role of a geopolitical entity, then the entire term should be annotated, in “*rapporteur Member States*” the entire phrase should be annotated.
- In coordinated noun phrases being nominal mentions of geopolitical entities, each constituent should be annotated separately, e.g., in “*members states and third countries*” both “*member states*” and “*third countries*” should be annotated separately. If some conjunct in a coordinated noun phrase is incomplete (e.g., in “*European and African countries*” the “*European*” conjunct is incomplete), the entire noun phrase should be annotated.
- Determiners and (possessive) pronouns do not constitute part of the text to be annotated, e.g. in “*the Member State*” only “*Member States*” should be annotated.
- In a text fragment such as, for example, “*Gyza, Egypt*”, each mention of a geopolitical entity should be annotated separately, i.e., “*Gyza*” and “*Egypt*”.
- As part of the general annotation rules, nominal elliptical constructions should be annotated as one annotation, e.g., both “*overseas countries and territories*” and “*European and Asian countries*” constitute one single annotation resp.

Facility

Entity type name: LOCATION: FACILITY

Entity type code: LOC-FAC

Definition: Facilities are functional and human-made structures created for a particular purpose that have a geographical location and are recognizable from a distance. For instance, tourist attractions (e.g., museums, amusement parks, stadiums, monuments), urban and extra-urban infrastructure (e.g., streets, highways, tunnels, parkings, bridges, hospitals), transportation hubs (e.g., airports, sea ports, etc.), buildings (e.g., churches, municipalities), are considered facilities. Natural landmarks are not considered facilities.

Examples: “*Schiphol Airport*”, “*Kyiv City Duma*”.

Counter-examples: “*Danube River*”

Annotation rules:

- Nominal mentions of facilities are not to be annotated.

Landmark

Entity type name: LOCATION: LANDFORM

Entity type code: LOC-LND

Definition: Landforms are natural land features (geological formations) on the solid surface of the Earth, which together

make up a given terrain. Landforms include hills, mountains, canyons, valleys, plains, plateaus, peaks, mountain ranges, dunes, water bodies (e.g. oceans, seas, lakes, rivers, bays, waterfalls), islands, peninsulas, continents, etc. Human-made land features are not considered landforms (e.g., bridges).

Examples: “*Baltic Sea*”, “*Niagara Falls*”, “*South and East France*”

Counter-examples: “*Central Park*”, “*France*”.

Annotation rules:

- In text fragments with a named mention of landform which is preceded or followed by a nominal mention of a landform (that does not constitute part of the named mention) only the named mention part should be annotated. For example, in “*Baltic Sea coastline*” only “*Baltic Sea*” should be annotated. Analogously, in the phrase “*the mountains in the vicinity of Lago Maggiore*” only the mention of “*Lago Maggiore*” is to be annotated.
- Mentions of geographical (not administrative) parts of GPEs are to be annotated as landforms. For example, the text fragment “*South Portugal*” refers to a geographical region rather than political entity, hence it should be annotated as a landform,
- Nominal mentions of landforms are not to be annotated.

1.2.4 Identifier

Identifier is a unique name/label (usually a word, number, letter, symbol, or any combination of those) that allows identifying a unique object, where such an object may be a product, document, standard, location, building, person, etc. This category covers various types of identifiers.

Publication

Entity type name: IDENTIFIER: PUBLICATION

Entity type code: IDT-PUB

Definition: Mentions of publication identifiers.

Examples: “*ISSN 1977-091X*”, “*ISBN 978-951-45-9693-3*”, “*OJ L 53, 27.2.0008, p. 52.*”, “*ECLI:EU:C:2020:1007*”

Counter-examples: “*Article 2(a)*”, “*Official Journal 2025*”, “*CN codes 0401*”

Postal address

Entity type name: IDENTIFIER: POSTAL-ADDRESS

Entity type code: IDT-POS

Definition: Mentions of postal addresses or elements of postal addresses

Examples: “*Via Piccinni 20, 24061 Albano Sant’alessandro, Italy*”, “*Via Piccinni 20*”, “*NY 10202*”.

Counter-examples: “*Two streets on the left*”, “*Next village*”, “*51° 28’ 38” N*”, “*abc@derf.com*”

Annotation rules:

- If all elements of the address are mentioned, they should be matched including city names and country names. These mentions of cities and countries should also be

annotated as nested entity mentions. Like in the example “Via Piccinni 20, 24061 Albano Sant’alessandro, Italy”, “Albano Sant’alessandro” and “Italy” should also be annotated respectively as GPE.

- The addressee does not constitute the part of the text to be annotated. For instance, in the text fragment “John Smith, Via Piccinni 20, 24061 Albano Sant’alessandro, Italy” only the part “Via Piccinni 20, 24061 Albano Sant’alessandro, Italy” should be annotated as postal address, while “John Smith” should be annotated as a mention of a person.

Geo-coordinates

Entity type name: IDENTIFIER: GEO-COORDINATES

Entity type code: IDT-GEO

Definition: Mentions of geographical coordinates.

Examples: “51° 28’ 38” N”.

Counter-examples: “NY 10202”, “255.000.000.000”, “+49 60 1234576”.

Phone number

Entity type name: IDENTIFIER: PHONE

Entity type code: IDT-PHO

Definition: Mentions of phone numbers

Examples: “+49 60 1234576”, “60 1234576”.

Counter-examples: “255.000.000.000”, “51° 28’ 38” N”.

Annotation rules:

- The trigger words preceding or following a phone number are not to be annotated, e.g., in “Tel. (43-1) 313 45-0” only “(43-1) 313 45-0” should be annotated.

Email address

Entity type name: IDENTIFIER: EMAIL

Entity type code: IDT-EMA

Definition: Mentions of email addresses.

Examples: “abc@derf.com”.

Counter-examples: “+49 60 1234576”, “Via Mazzini 27”.

Url

Entity type name: IDENTIFIER: URL

Entity type code: IDT-URL

Definition: Mentions of URLs.

Examples: “http://www.google.com”.

Counter-examples: “255.000.000.000”.

IP

Entity type name: IDENTIFIER: IP

Entity type code: IDT-IP

Definition: Mentions of IP addresses.

Examples: “255.000.000.000”.

Counter-examples: “http://www.google.com”.

VAT number

Entity type name: IDENTIFIER: VAT

Entity type code: IDT-VAT

Definition: Mentions of VAT (Value Added Taxation) numbers.

Examples: “ATU99999999”.

Counter-examples: “CN codes 0401”.

Bank Identity

Entity type name: IDENTIFIER: BANK

Entity type code: IDT-BAN

Definition: Mentions of bank account numbers, IBANs.

Examples: “DE44 5001 0517 5407 3249 31”.

Counter-examples: “2345 3523 2453 3453”.

Credit card

Entity type name: IDENTIFIER: CREDIT-CARD

Entity type code: IDT-CCN

Definition: Mentions of credit card numbers.

Examples: “2345 3523 2453 3453”.

Counter-examples: “DE44 5001 0517 5407 3249 31”.

Product identifier

Entity type name: IDENTIFIER: PRODUCT

Entity type code: IDT-PRO

Definition: Mentions of product codes such as TARIC codes or CN (Combined Nomenclature) codes.

Examples: “CN codes 0401”, TARIC examples such as “3809910000 10”, “1904209900”.

Counter-examples: “Iphone 16”.

Dossier identifier

Entity type name: IDENTIFIER: DOSSIER

Entity type code: IDT-DOS

Definition: Mention of a specific EU dossier, such as inter-institutional procedures, as cases from the Court of Justice, Merger cases, State aid cases, Digital Market Act cases, Antitrust and Cartel cases, Foreign Subsidies cases.

Examples: “2024/0224(COD)”, “22/71” , “F-59/09”, “T-264/11 P”, “F-59/09 RENV”, “M.0009”, “SA.00058”, “DMA.000028”, “AT.00042”, “FS.000011”.

For instance with a bit of context: “Case 22/71 ‘Béguelin’ [1971], paragraph 11, page 949” , “F-59/09, T-264/11 P and F-59/09 RENV.” , “M.0009 EssilorLuxottica/GrandVision”, “cases SA.00058, SA.00071, and SA.00059”, “Decision of 12 February 2024 in Case DMA.000028 – Microsoft – Web Browsers”, “Case AT.00042, Pierre Cardin” , “with reference FS.000011”.

Counter-examples: *“2015/C 016/71”*.

Annotation rules:

- The words like *“Case”* or *“cases”* may help detecting an IDT-DOS but should not be part of the annotation.
- Analogously, additional information on the case, e.g., the parties involved, should not be part of the annotation. For instance, in the text fragment: *“Case IV/M.0003 - British Petroleum/Amoco Corporation”* only the fragment *“IV/M.0003”* should be annotated as mention of a case dossier.

Online media identifier

Entity type name: IDENTIFIER: ONLINE-MEDIA

Entity type code: IDT-ONL

Definition: Mentions of online media-like identifiers

Examples: *“@johnS”*.

Counter-examples: *“john.smith@gmail.com”*.

I.2.5 Product

Product entities cover named mentions of entities such as services, initiatives, standards, documents and goods. Names of substances, materials, and especially type of products are not part of this category.

Service

Entity type name: PRODUCT: SERVICE

Entity type code: PRO-SER

Definition: Named mentions of services, where a service in this context refers to an instrument, tool, functionality put in place in order to help or facilitate accomplishing certain tasks and processes or to fulfill specific requirements in business of any kind. For instance, IT systems, online platforms, portals, human-operated assistance services, banking instruments, etc., are examples of services. Services are used to implement initiatives covered by a different entity category.

Examples: *“Official Journal of the European Union”*, *“Google Search Engine”*, *“VISA Information System”*, *“F-gas Portal”*, *“Customs Risk Management System”*, *“European Union Public Licence”*.

Counter-examples: *“Google”*, *“United Nations 2030 Agenda for Sustainable Development and its Sustainable Development Goals”*, *“various nice services”*.

Annotation rules:

- Nominal mentions of services are not to be annotated.

Initiative

Entity type name: PRODUCT: INITIATIVE

Entity type code: PRO-INI

Definition: Named mentions of initiatives, i.e., actions and endeavours such as plans, agendas, projects, programmes, strategies, visions, etc., designed to achieve specific goals or to bring about change in the context of any domain.

Examples: *“Horizon 2020”*, *“Circular Economy Action Plan”*, *“United Nations 2030 Agenda for Sustainable Development and its Sustainable Development Goals”*.

Counter-examples: *“Visa Information System”*, *“External Action Service”*, *“project launched by Microsoft”*.

Annotation rules:

- Nominal mentions of initiatives are not to be annotated.

Standard

Entity type name: PRODUCT: STANDARD

Entity type code: PRO-STD

Definition: Mentions of concrete standards, where standard is considered a set of agreed-upon and documented guidelines, specifications, accepted practices, technical requirements, or terminologies for diverse fields, which has been introduced by an organization (international or national) dealing with standards, e.g., ISO.

Examples: *“API Standard 2543”*, *“ASTM D1086-1964”*, *“International Electrotechnical Commission (IEC) 60335-2-89 and IEC 60335-2-40”*

Counter-examples: *“ISO”*, *“CEN”*, *“PDF”*, *“XML”*.

Annotation rules:

- Mentions of standard type/family, e.g., *“ISO standards”* are not to be annotated, only explicit standard with an ID should be annotated.
- Mentions of versions and amended versions of standards are to be annotated, e.g., the full text fragment *“EN 60598-1:1997/A13:1999”* should be annotated, not only *“EN 60598-1:1997”*.
- Mentions of data format standards are not to be annotated as PRODUCT: STANDARD, however, they can be annotated using the generic type PRODUCT.
- Nominal mentions of standards are not to be annotated.

Document

Entity type name: PRODUCT: DOCUMENT

Entity type code: PRO-DOC

Definition: Named mention of any kind of documents which do not fall under the LEGAL RESOURCE entity type. This is the case also for some non-legal publications produced by a European institution like JRC Scientific Report or reports produced by European Agencies. General Publication also falls under this category.

Examples: *“Annual Activity Report on Climate Action”*, *“United Nations World Water Development Report”*.

Counter-examples: *“Council Regulation (EEC, Euratom, ECSC) No 259/68 of 29 February 1968”*, *“2024 Charter Report - FRA summary of civil society consultation on funding for fundamental rights”*.

Annotation rules:

- A named mention of document on the topic of legislation, even EU legislation, may be annotated, like a reference to an academic book talking about legislation.
- Nominal mentions of documents are not to be annotated.

I.2.6 Legal Resource

Legal resource entity type covers mentions of EU, international and national legislation documents of any kind.

International legislation

Entity type name: LEGAL RESOURCE: INTERNATIONAL

Entity type code: LEG-INT

Definition: Mentions of international legislation. In this specific context, it includes only regulations, agreements, treaties, and other legal resources where at least two different countries are involved AND where NONE of the European Union institutions is a party thereof. International legislation includes any bilateral/multilateral agreement between two/multiple countries, even if these countries are EU Member States always with the condition that NONE of the European Union institutions is a party thereof.

Examples: “*North American Free Trade Agreement*”, “*Constitutive Act of the African Union*”, “*ASEAN*”, “*Franco-Polish strategic partnership agreement*”.

Counter-examples: “*Paris Agreement*”, “*Mercosur*”, “*Schengen Agreement*”.

Annotation rules:

- References to annexes, sections, points preceding or following the mention of EU legislation do not constitute part of the annotation.
- A free-text description that follows a mention of an international legislation should not be annotated.
- If there is in a text an acronym or name of any institution responsible/involved in releasing the legislation then it should be part of the text to be annotated, e.g. “*UN General Assembly Resolution ES-10/20*”.
- Nominal mentions of international legislation are not to be annotated.

EU legislation

Entity type name: LEGAL RESOURCE: EU

Entity type code: LEG-EU

Definition: All mentions of legislation, including agreement, treaties, regulations where any of the European Union institutions is one of the involved parties. All founding treaties (EURATOM, EU, EEC, etc) are considered as EU Legal Resources. Otherwise, having one or more EU member states as a party of an agreement without any involvement of EU institutions does not fall under EU legal resource.

A more detailed list of examples of EU legislation types is provided at:

https://eur-lex.europa.eu/content/tools/TableOfSectors/types_of_documents_in_eurlex.html

Examples: “*Paris Agreement*”, “*Mercosur*”, “*Schengen Agreement*”, “*Council Regulation (EEC, Euratom, ECSC) No 259/68 of 29 February 1968*”, “*Council Directive 75/268/EEC of 28 April 1975*”, “*Decision No 477/2010/EU of the European Parliament and of the Council*”.

Counter-examples: “*OJ L, 2024/573, 20.2.0004*”, should be annotated as publication identifier, “*ASEAN*”, “*Franco-Polish*

strategic partnership agreement”, “*Union law*” (does not refer to any concrete legislation).

Annotation rules:

- A free-text description that follows a mention of an EU legislation should not be annotated, unless there is no other identifier of the legislation in the text and the free-text description **appears to be the title of the legal document**. For instance, the entire text fragment “*Memorandum of Understanding on a Code of Conduct for the members of the Governing Council*” should be annotated as a mention of EU legislation.
- A date following the identifier of the EU legislation constitutes part of the mention and should be annotated, unless it is mentioned in a relative sub-clause. For instance, in “*Riga Ministerial Declaration of 11 June 2006*” (refers to a meeting of EU country ministers, hence it is treated as EU legal resource) the entire text fragment should be annotated, whereas, in “*Riga Ministerial Declaration which took place last year on 11 June*” only “*Riga Ministerial Declaration*” should be annotated.
- References to annexes, sections, points preceding or following the mention of EU legislation do not constitute part of the annotation.
- Sometimes, a short version of how the EU legislation will be referred to, e.g., “*Decision*”, is introduced in the vicinity of the full named mention. Such shorter versions should be annotated, but only when they are introduced, any subsequent mentions thereof (which appear nominal) should not be annotated.
- Nominal mentions of EU legislation are not to be annotated.

National legislation

Entity type name: LEGAL RESOURCE: NATIONAL

Entity type code: LEG-NAT

Definition: Mentions of any kind of national legislation.

Examples: “*Constitution of the Fifth Republic of 4 October 1958*”, “*Constitutional Law of 25 June 1992*”, “*Toubon Act*”.

Counter-examples: “*Paris Agreement*”, “*Decision No 477/2010/EU of the European Parliament and of the Council*”.

Annotation rules:

- References to annexes, sections, points preceding or following the mention of national legislation do not constitute part of the annotation.
- A free-text description that follows a mention of national legislation should not be annotated.
- Nominal mentions of national legislation are not to be annotated.

Subdivision element

Entity type name: LEGAL RESOURCE: SUBDIVISION

Entity type code: LEG-SUB

Definition: Any mention of document/resource sub-section or part.

Examples: “*Article 2*”, “*Annexe III*”, “*Paragraph 1.a*”, “*Article 2, 3 and 5*”.

Counter-examples: “*pages*”.

Annotation rules:

- In a consecutive sequence of references to various resource parts, each specific mention has to be annotated separately, e.g., in “*Article 2, page 2*” both “*Article 2*” and “*page 2*” should be annotated separately.
- A mention of “*Annex*” or “*Preamble and General Provisions*” can be annotated, where in the case of the former it should be clear from the text there is only one annex in the document.

I.2.7 Event

Named mentions of events such as meetings, incidents, conflicts and occasions, and named mentions of events which contain information which allows to tailor them to a specific point in time and/or contain named mention of the participants/locations, etc.

Meeting

Entity type name: EVENT: MEETING

Entity type code: EVT-MT

Definition: Mentions of any type of meetings and gatherings aims at discussions, presenting views and sharing information, including, for instance: parliamentary sessions, conferences, workshops, summits, fora, expositions, debates, seminars, etc.

Examples: “*14th Session of ICUMSA 1966*”, “*Dublin European Council meeting on 13 and 14 December 1996*”, “*Member States meeting within the Council of 16 May 1989*”, “*France-Ukraine meeting*”

Counter-examples: “*sessions in the EU parliament*”, “*future meetings*”

Annotation rules:

- Potential time references and/or named mentions of participants and locations should constitute part of the annotated text fragment.
- Analogously, the aforementioned “internal” mentions of locations, participants and temporal references should be annotated separately as well, e.g., in “*Member States meeting within the Council of 16 May 1989*” “*Member States*”, “*16 May 1989*” constitute nested annotations.
- In certain texts a reference to a date might appear as a hint to label a text fragment as an event, for instance, in the text fragment “*As underlined in the Commission communication of 19 May 2010 entitled ‘A Digital Agenda for Europe’, public authorities should play their part in promoting markets for online content.*” the phrase: “*Commission communication of 19 May 2010*” refers to a document released by EU (with a title of that document following), which falls under EU legislation category, and hence, should not be annotated as a mention of a meeting.

Conflict

Entity type name: EVENT: CONFLICT

Entity type code: EVT-CON

Definition: Mentions of any type of conflicts and disputes, including, for instance: wars, armed invasions, revolutions, disagreements between countries, coup d’etat, etc.

Examples: “*World War II*”, “*Russia-Ukraine war*”, “*ADF attack in Goma*”, “*Brexit*”

Counter-examples: “*future war*”, “*potential conflict*”

Annotation rules:

- Potential time references and/or named mentions of participants and locations should constitute part of the annotated text fragment.
- Analogously, the aforementioned “internal” mentions of locations, participants and temporal references should be annotated separately as well, e.g., in “*ADF attack in Goma*”, the mention of “*Goma*” constitutes a nested annotation.
- Long free-text descriptions of conflicts, e.g., “*Russia’s unprovoked and unjustified war of aggression against Ukraine on 24 February 2022*” are not considered named mentions of events and are not to be annotated.

Incident or Disaster

Entity type name: EVENT: INCIDENT-DISASTER

Entity type code: EVT-IND

Definition: Mentions of any type of natural and man-made disasters, including, for instance: pandemics, earthquakes, floods, humanitarian crisis, cyber attacks, assassinations, car and plane crashes, etc.

Examples: “*Mississippi flood*”, “*Chernobyl nuclear accident*”, “*COVID pandemic*”, “*Mont Louis accident in 1984*”

Counter-examples: “*natural disaster*”

Annotation rules:

- Potential time references and/or named mentions of participants and locations should constitute part of the annotated text fragment.
- Analogously, the aforementioned “internal” mentions of locations, participants and temporal references should be annotated separately as well, e.g., in “*Mont Louis accident in 1984*”, the mention of “*1984*” constitutes a nested annotation.
- Long free-text descriptions of incidents, e.g., “*the floods which destroyed large parts of Haiti in 2015*”, are not considered named mentions of events, and should not be annotated.

Occasion

Entity type name: EVENT: OCCASION

Entity type code: EVT-OCC

Definition: Mentions of an event happening in particular time marked by certain circumstances or occurrences, including, for instance: awards, sport events like world cups, festivals,

religious periods or holidays, anniversaries, commemorations, parades, etc.

Examples: “*Fatherland Day*”, “*Ramadan*”, “*Europe Day on 9 May*”, “*MTV Video Music Awards*”, “*Christmas*”

Counter-examples: “*religious celebration*”

Annotation rules:

- Potential time references and/or named mentions of participants and locations should constitute part of the annotated text fragment.
- Analogously, the aforementioned “internal” mentions of locations, participants and temporal references should be annotated separately as well, e.g., in “*Europe Day on 9 May*”, the mention of “*9 May*” constitutes a nested annotation.
- Analogously, as for the other event types, long free-text descriptions of occasions are not considered named mentions of occasions.

I.2.8 Timex (Temporal Expression)

Temporal expression are sequences of tokens that denote time, that is, express a point in time, a duration, period, frequency or deadline.

Time

Entity type name: TIMEX: TIME

Entity type code: TIM-TM

Definition: Explicit mentions of time using a reference to concrete hour/minute.

Examples: “*10 o’clock*”, “*20:30 CET*”, “*2 p.m.*”

Counter-examples: “*2 May 2024*”

Annotation rules:

- If time and date references are mentioned together in a consecutive sequence of tokens, e.g., “*20:30 CET, 2 May 2024*”, each of the two has to be annotated separately, i.e., “*20:30 CET*” as time and as “*2 May 2024*” date.
- Mentions of relative time references (e.g., “*next hour*”) are not subject to annotation.

Date

Entity type name: TIMEX: DATE

Entity type code: TIM-DAT

Definition: Mentions of point in time referring to a concrete date, month or year.

Examples: “*1 January 2000*”, “*March 2023*”, “*1 June*”, “*February*”, “*8 and 9 March 2020*”

Counter-examples: “*yesterday*”, “*last year*”, “*in future years*”

Annotation rules:

- Mentions of relative date references (e.g., “*last year*”) are not subject to annotation,
- Internal mentions of date elements, e.g., month references, are not subject to annotation.

Duration

Entity type name: TIMEX: DURATION

Entity type code: TIM-DUR

Definition: Mention of a concrete amount (quantity) of time expressed in, for instance, years, months, weeks, days, hours, minutes, etc. The mention can also refer to a specific time span, indicating minimum and maximum amount expressed using some time unit.

Examples: “*20 weeks*”, “*1 year*”, “*two months*”, “*2 to 4 months*”, “*10 to 20 years*”,

Counter-examples: “*ages*”, “*many years*”, “*few weeks*”

Annotation rules:

- Complex duration expressions, for instance including coordination should be annotated as one mention. For example “*2 years and 6 months*” should be one mention.
- References to approximate durations (e.g., “*several years*”) are not subject to annotation,
- Mentions of ranges of time should be annotated, and the whenever applicable the internal mentions to time quantities should be annotated as well.

Period

Entity type name: TIMEX: PERIOD

Entity type code: TIM-PER

Definition: Mention of a time span with a start and end date.

Examples: “*from 1 January 2024 to 31 December 2024*”, “*between March and April 2023*”, “*20 weeks*”, “*1970s*” (refers to the decade, i.e., 1970-1980), “*024 and 2025*”

Counter-examples: “*from one day to another*”, “*2 years*” (is a duration)

Annotation rules:

- The dates mentioned within the text span representing the period should be annotated separately as well.

Frequency

Entity type name: TIMEX: FREQUENCY

Entity type code: TIM-FR

Definition: Mentions of frequencies expressed in some time units, e.g., years, months, weeks, or expressions indicating periodicity, even if it is not specified exactly.

Examples: “*every year*”, “*each month*”, “*once in 2 years*”, “*at least twice a year*”, “*at regular intervals*”, “*semi-annually*”, “*quarterly*”, “*periodically*”, “*seven days a week*”

Counter-examples: “*as soon as possible*”, “*whenever possible*”

Annotation rules:

- The amount of time if present within the annotated span representing frequency should be annotated as well as duration accordingly.
- Approximative and nominal references to frequencies (e.g., “*at regular intervals*”, “*regularly*”) are to be annotated and flagged with the nominal attribute accordingly.

Deadline

Entity type name: TIMEX: DEADLINE

Entity type code: TIM-DE

Definition: Mentions of deadlines which include references to concrete dates or point in time, or expressions and phrases indicating a deadline that provides the sense of urgency, even if it is somewhat vague.

Examples: “*by 1 May 2020*”, “*until March 2023*”, “*at the latest within six months*”, “*without delay*”, “*as soon as possible*”, “*immediately*”, “*after a period of four years*”, “*from one year after*”, “*no more than 6 weeks before*”, “*within a period of 14 days*”, “*20 days following*”

Counter-examples: “*at some point in time*”, “*later*”

Annotation rules:

- In case of deadline mentions that are related to some specific event or fulfillment of certain conditions, only the core temporal expression is to be annotated, e.g., in the text fragment “*at the latest within six months after the accomplishment of the project*”, only “*at the latest within six months after*” should be annotated. In principle, the text that follows the preposition (e.g., “*after*”, “*before*”, “*from*”, etc.) should not be included as part of the annotation.
- The dates mentioned within the text span representing the deadline should be annotated separately as well.
- Relative, urgent time references such as , e.g., “*as soon as possible*”, “*immediately*” are to be annotated and flagged with the nominal mention attribute accordingly.

I.2.9 Numex (Numerical Expression)

Numex category covers all types of numerical expressions, like for instance percentages, currency expressions, measurements, counts of objects, etc.

Numerical

Entity type name: NUMEX: NUMERICAL

Entity type code: NUM-EX

Definition: Mentions of any numbers, including natural and rational numbers, percentages, written both using digits and words. Combinations of numbers and operators are not considered to fall under this category, although some elements they comprise of do so.

Examples: “*12 000*”, “*two hundred*”, “*30%*”, “*between one and ten*”, “*1,23-3,45*”

Counter-examples: “*3+5*”, “*first*”

Annotation rules:

- Numerical expressions containing some arithmetic symbols are not subject to annotation,
- Ordinal numbers should not be annotated.
- Mentions of ranges of numerical values should be annotated.

Currency expression

Entity type name: NUMEX: CURRENCY

Entity type code: NUM-CU

Definition: Mentions of numerical expressions that refer to an amount of money in an explicitly mentioned currency. The numerical values can be expressed both using numbers and words. Mentions of the currency itself alone does not fall in this category.

Examples: “*2.50\$*”, “*two dollars*”, “*23K Euro*”, “*three million pounds*”, “*two hundred €*”, “*30EUR*”, “*10-30 EURO*”, “*between 10 and 20 EUR*”

Counter-examples: “*\$*”, “*euro*”

Annotation rules:

- Stand-alone mentions of currencies are not subject to annotation.
- Mentions of ranges of values of currencies should be annotated.

Age

Entity type name: NUMEX: AGE

Entity type code: NUM-AG

Definition: Mentions of an age of a person, object or an artifact

Examples: “*20 years old*”, “*fifteen years old*”, “*10-years old*”

Counter-examples: “*very old*”, “*new born*”

Annotation rules:

- The mention of the person, object or an artefact whose age is being mentioned does not constitute part of the text span to be annotated.

Measurement expression

Entity type name: NUMEX: MEASUREMENT

Entity type code: NUM-ME

Definition: Mentions of any type of measurements, including, but not limited to: physical extent, space, volume, weight, temperature, speed, intensity, pressure, lighting, intensity, electrical power, etc. References to durations fall under a different category, namely, TIMEX: DURATION.

Examples: “*10 m3*”, “*thirty watts*”, “*60 km/h*”, “*98 decibel*”, “*28 °C*”, “*from 1,0 to 7,0 cm*”

Counter-examples: “*2 months*”, “*five years*”

Annotation rules:

- References to durations should not be annotated with this label.
- Mentions of ranges of values related to measurements should be annotated.

Count of objects

Entity type name: NUMEX: COUNTX

Entity type code: NUM-CT

Definition: Mentions of quantities of persons, living organisms, locations, organizations, events and objects, like for instance, products, etc.

Examples: “*20 persons*”, “*hundred computers*”, “*5 meetings*”, “*250 incidents*”, “*10 updates*”, “*one country*” “*two to three airplanes*”

Counter-examples: “*person*”, “*system*”

Annotation rules:

- Mentions of a person, object, event, etc. without an explicit mention of the quantity should not be annotated as instances of this entity type,
- Mentions of ranges of quantities should be annotated.