

EuroLegalBot

CRIMINAL MATTERS CHATBOT

Enhancing Judicial Cooperation through a European Arrest Warrant (EAW) Chatbot

D2.2 DATASET



Co-funded by the
European Union

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

EuroLegalBot

Project Number: 101192405

JUST-2024-JCOO

D2.2 – DATASET

Version 1.0

1.	Change Control	- 4 -
1.1	Document Properties.....	- 4 -
1.2	Revision History	- 4 -
1.	Executive Summary	- 5 -
2.	Training of the Chatbot and the technical infrastructure	- 7 -
2.1	Imprint of the Avatar and training methodology	- 7 -
2.2	Data Transparency	- 8 -
2.3	Criteria for the judicial reasoning and organisation of the initial prompts-	14
	-	
2.4	Scraping Methodology and Compliance	- 18 -
2.5	General Overview of the 'scraping' process feeding the Avatars.....	- 20 -
2.5.1	File Distribution by Directory	- 20 -
2.5.2	Quantitative Considerations	- 22 -
2.5.3	Update and Data Acquisition Method	- 22 -
2.5.4	Purpose of the Dataset	- 22 -

1. Change Control

1.1 Document Properties

Deliverable No.	D2.2		
Work Package No.	WP2	Work Package Title	Dataset Creation
Author/s	European Public Law Organization - EPLO		
Contributor/s	Corte di Appello di Venezia (AppCourtVE) Fondazione Agenfor International (AGF) Institut Europeen De L'Expertise Et De L'Expert (EEEI) Expleo Italia S.P.A. (EXPLEO)		
Name	Dataset		
Date	22.12.2025		
Dissemination Level	PU		

1.2 Revision History

Version	Date	Comments
1.0		

1. Executive Summary

EuroLegalBot is an innovative project aiming to enhance judicial cooperation at the European level, particularly focusing on streamlining the use of the European Arrest Warrant (EAW).

The project focuses on the development and implementation of a Chatbot designed to support practitioners active in the justice sector in the efficient and effective utilisation of the EAW mechanism.

By leveraging cutting-edge technology, EuroLegalBot aims to facilitate smoother cross-border extradition processes, ultimately contributing to a more cohesive and efficient European legal framework.

More specifically, the project entails the development of an advanced Chatbot, EuroLegalBot, equipped with sophisticated natural language processing (NLP) capabilities. This intelligent Chatbot is designed to engage with practitioners active in the justice sector, including judges, prosecutors, lawyers, and law enforcement officials, providing them with tailored guidance, assistance, and up-to-date information regarding the EAW. Furthermore, EuroLegalBot integrates a comprehensive knowledge base containing relevant legal information, procedural guidelines, and case law associated with the EAW. This knowledge base is continuously updated to reflect legislative developments and best practices, while machine learning algorithms are employed to analyse user queries and retrieve contextually relevant information. The project foresees a pilot phase, during which the Chatbot will be tested extensively, over a period of four months, by 180 practitioners (60 per target country) active in the justice sector. Work Package 2 on Dataset Creation aims to collect and organise systematically in one dataset all material related to the European Arrest Warrant from official sources of the European Union and from each of the member states involved in the project. More concretely, it aims to create a methodology that will guide the collection and labelling of data, as well as the structure of the dataset, to identify for each target country the sources for the data that will compose the dataset on which the Chatbot will be trained, to collect & systematise the data and to lead to the creation of the dataset. The dataset comprises structured data organised into a machine-readable format suitable for machine learning algorithms. The dataset is cleaned, validated, and annotated to ensure its accuracy and relevance for training purposes.

Detailed documentation accompanying the dataset provides metadata, data dictionaries, and usage guidelines, enabling researchers and developers to effectively utilise the dataset for training and evaluation purposes.

2. Training of the Chatbot and the technical infrastructure

The chatbot was trained using dual criterion.

2.1 Imprint of the Avatar and training methodology

First, an API database was built on the basis of an initial dataset consisting exclusively of legal information derived from data of absolute certainty and reliability. This first group of data and knowledge was used as the foundation for training a constellation of multiple LLMs, while giving preference to European technological solutions such as Mistral and Apertus.

Therefore, we can state that this phase of the work **CREATED A EUROPEAN CHARACTER** for the artificial intelligence system. Its neural methodologies and logical connections, as well as its foundational cultural framework, are built upon logical-cultural infrastructures and European patterns, which were prioritized from a technological standpoint.

The development of training solutions based on the integration of AI solutions represents the future of Europe, in our opinion.

In the practical application, we considered that these Tech technologies in the justice sector is considered a high-risk activity under the AI Act. Consequently, the project has highlighted four fundamental principles- which should be the Best Practices for future developments-, which served as the guiding framework for advancements in this field. We considered these essential methodologies in the technical set-up:

1. **It is essential to strengthen the technological infrastructure underlying vertical AI solutions.** Strengthening means decentralizing and scaling up the capabilities of European servers, where the data must be stored. It is of **VITAL IMPORTANCE** that these servers and their connected technological architectures remain within the national digital perimeter. In this regard, Agenfor has already initiated this process using an **ACN2-certified server**, which meets these requirements.
2. **Data must be made interoperable among various developers of AI-based solutions** so that European legal knowledge can be consolidated, and European knowledge hubs can assert their cognitive strength compared to the

internet-based data systems used by the U.S. and China. In our view, the **'open data' strategy** for European judicial AI platforms should become mandatory in EU-funded projects, with no copyright or privacy barriers for data created with EU support, or at the very least, such data should be freely accessible and without artificial restrictions by their creators. Data must be certified by a qualified agency, avoiding any 'scraping' access, because from the quality of the data depends the reliability of the outcomes.

3. **Every AI platform must be usable both via internet solutions and, more importantly, through offline systems**, where both data and LLMs operate in completely closed environments. This is crucial for sectors such as **justice**, **security**, and certainly **defense**, where the data managed by avatars must not exit secure circuits. No judicial or security authority will ever upload data to knowledge hubs that rely on internet access.
4. **It is of VITAL importance that the LLMs used are compatible with the principles of the AI Act**. The adoption of non-European solutions should therefore be discouraged. Furthermore, all algorithms must be **verifiable through agentic-based systems**, and the platforms must feature **user-friendly interfaces** that allow human operators to correct errors as they work, prioritizing human agent corrections over the logical systems of the AI.

These principles guided our initial strategic activity in the building of the dataset and in the preliminary training of the 'brain' of the avatar, its reasoning based on neural connections, LLMs and prompts.

2.2 Data Transparency

For transparency, we list here the criteria and sources of the initial training, which formed the 'character' of the avatar based on 92 judicial cases (links provided):

1. [Case C-303/05, *Advocaten voor de Wereld*, Judgment of 3 May 2007.](#)
2. [Case C-66/08, *Kozłowski*, Judgment of 17 July 2008.](#)
3. [Case C-296/08 PPU, *Santesteban Goicoechea*, Judgment of 12 August 2008.](#)

4. [Case C-388/08, **Leymann and Pustovarov**, Judgment of 1 December 2008.](#)
5. [Case C-123/08, **Wolzenburg**, Judgment of 6 October 2009.](#)
6. [Case C-306/09, **IB**, Judgment of 21 October 2010.](#)
7. [Case C-261/09, **Mantello**, Judgment of 16 November 2010.](#)
8. [Case C-192/12 PPU, **Melvin West**, Judgment of 28 June 2012.](#)
9. [Case C-42/11, **Lopes Da Silva Jorge**, Judgment of 5 September 2012.](#)
10. [Case C-396/11, **Radu**, Judgment of 29 January 2013.](#)
11. [Case C-399/11, **Melloni**, Judgment of 26 February 2013.](#)
12. [Case C-168/13 PPU, **Jeremy F**, Judgment of 30 May 2013.](#)
13. [Case C-237/15 PPU, **Lanigan**, Judgment of 16 July 2015.](#)
14. [Case C-463/15 PPU, **A**, Order of 25 September 2015.](#)
15. [Joined Cases C-404/15 and C-659/15 PPU, **Aranyosi and Căldăraru**, Judgment of 5 April 2016.](#)
16. [Case C-108/16 PPU, **Dworzecki**, Judgment of 24 May 2016.](#)
17. [Case C-241/15, **Bob-Dogi**, Judgment of 1 June 2016.](#)
18. [Case C-294/16 PPU, **JZ**, Judgment of 28 July 2016.](#)
19. [Case C-182/15, **Petruhhin**, Judgment of 6 September 2016.](#)
20. [Case C-452/16 PPU, **Poltorak**, Judgment of 10 November 2016.](#)
21. [Case C-453/16 PPU, **Özçelik**, Judgment of 10 November 2016.](#)
22. [Case C-477/16 PPU, **Kovalkovas**, Judgment of 10 November 2016.](#)
23. [Case C-640/15, **Vilkas**, Judgment of 25 January 2017.](#)
24. [Case C-579/15, **Popławski**, Judgment of 29 June 2017 \(**Popławski I**\).](#)
25. [Case C-270/17 PPU, **Tupikas**, Judgment of 10 August 2017.](#)
26. [Case C-271/17 PPU, **Zdziaszek**, Judgment of 10 August 2017.](#)
27. [Case C-571/17 PPU, **Ardic**, Judgment of 22 December 2017.](#)
28. [Case C-367/16, **Piotrowski**, Judgment of 23 January 2018.](#)
29. [Case C-191/16, **Pisciotti**, Judgment of 10 April 2018.](#)
30. [Case C-216/18 PPU, **Minister for Justice and Equality \(Deficiencies in the System of Justice\)**, Judgment of 25 July 2018.](#)
31. [Case C-220/18 PPU, **ML \(Conditions of detention in Hungary\)**, Judgment of 25 July 2018.](#)
32. [Case C-268/17, **AY \(Mandat d'arrêt – Témoin\)**, Judgment of 25 July 2018.](#)

33. [Case C-327/18 PPU, **RO**, Judgment of 19 September 2018.](#)
34. [Case C-247/17, **Raugevicius**, Judgment of 13 November 2018.](#)
35. [Case C-551/18 PPU, **IK \(Enforcement of an additional sentence\)**, Judgment of 6 December 2018.](#)
36. [Case C-514/17, **Sut**, Judgment of 13 December 2018.](#)
37. [Case C-492/18 PPU, **TC**, Judgment of 12 February 2019.](#)
38. [Case C-573/17, **Popławski**, Judgment of 24 June 2019 \(*Popławski II*\).](#)
39. [Joined Cases C-508/18 and C-82/19 PPU, **OG and PI \(Parquets de Lübeck and Zwickau\)**, Judgment of 27 May 2019.](#)
40. [Case C-509/18, **PF \(Prosecutor General of Lithuania\)**, Judgment of 27 May 2019.](#)
41. [Case C-489/19 PPU, **NJ \(Parquet de Vienne\)**, Judgment of 9 October 2019.](#)
42. [Case C-128/18, **Dorobantu**, Judgment of 15 October 2019.](#)
43. [Joined Cases C-566/19 PPU and C-626/19 PPU, **Parquet général du Grand-Duché de Luxembourg and de Tours**, Judgment of 12 December 2019.](#)
44. [Case C-625/19 PPU, **Openbaar Ministerie \(Swedish Public Prosecutor's Office\)**, Judgment of 12 December 2019.](#)
45. [Case C-627/19 PPU, **Openbaar Ministerie \(Public Prosecutor, Brussels\)**, Judgment of 12 December 2019.](#)
46. [Case C-813/19 PPU, **MN**, Order of 21 January 2020.](#)
47. [Case C-717/18, **X \(European arrest warrant – Double criminality\)**, Judgment of 3 March 2020.](#)
48. [Case C-314/18, **SF \(Mandat d'arrêt européen – Garantie de renvoi dans l'État d'exécution\)**, Judgment of 11 March 2020.](#)
49. [Case C-897/19, **Ruska Federacija**, Judgment of 2 April 2020.](#)
50. [Case C-195/20 PPO, **Generalbundesanwalt beim Bundesgerichtshof \(Speciality rule\)**, Judgment of 24 September 2020.](#)
51. [Case C-510/19, **Openbaar Ministerie \(Faux en écritures\)**, Judgment of 24 November 2020.](#)
52. [Case C-416/20 PPU, **Generalstaatsanwaltschaft Hamburg**, Judgment of 17 December 2020.](#)

53. [Case C-398/19, *Generalstaatsanwaltschaft Berlin \(Extradition vers l'Ukraine\)*, Judgment of 17 December 2020.](#)
54. [Joined Cases C-354/20 PPU and C-412/20 PPU, *Openbaar Ministerie \(Indépendance de l'autorité judiciaire d'émission\)*, Judgment of 17 December 2020.](#)
55. [Case C-414/20 PPU, *MM*, Judgment of 13 January 2021.](#)
56. [Case C-649/19, *Spetsializirana prokuratura \(Déclaration des droits\)*, Judgment of 28 January 2021.](#)
57. [Case C-648/20 PPU, *Svishtov Regional Prosecutor's Office*, Judgment of 10 March 2021.](#)
58. [Case C-488/19, *Minister for Justice and Equality \(Mandat d'arrêt – Condamnation dans un État tiers, membre de l'EEE\)*, Judgment of 17 March 2021.](#)
59. [Case C-665/20 PPU, *X \(Mandat d'arrêt européen – Ne bis in idem\)*, Judgment of 29 April 2021.](#)
60. [Case C-206/20, *Prosecutor of the regional prosecutor's office in Ruse, Bulgaria*, Order of 22 June 2021.](#)
61. [Joined cases C-428/21 PPU and C-429/21 PPU, *Openbaar Ministerie \(Droit d'être entendu par l'autorité judiciaire d'exécution\)*, Judgment of 26 October 2021.](#)
62. [Case C-479/21 PPU, *Governor of Cloverhill Prison and Others*, Judgment of 16 November 2021.](#)
63. [Case C-203/20, *AB and Others \(Révocation d'une amnistie\)*, Judgment of 16 December 2021.](#)
64. [Joined Cases C-562/21 PPU and C-563/21 PPU, *Openbaar Ministerie \(Tribunal établi par la loi dans l'État membre d'émission\)*, Judgment of 22 February 2022.](#)
65. [Case C-804/21 PPU, *C and CD \(Obstacles juridiques à l'exécution d'une décision de remise\)*, Judgment of 28 April 2022.](#)
66. [Case C-105/21, *Spetsializirana prokuratura \(Informations sur la décision nationale d'arrestation\)*, Judgment of 30 June 2022.](#)
67. [Case C-480/21, *Minister for Justice and Equality \(Tribunal établi par la loi dans l'État membre d'émission – II\)*, Order of 12 July 2022.](#)

68. [Case C-168/21, *Procureur général près la cour d'appel d'Angers*, Judgment of 14 July 2022.](#)
69. [Case C-492/22 PPU, *Openbaar Ministerie \(Décision de remise différée en raison de poursuites pénales\)*, Judgment of 8 December 2022.](#)
70. [Case C-237/21, Generalstaatsanwaltschaft München \(Demande d'extradition vers la *Bosnie-Herzégovine*\), Judgment of 22 December 2022.](#)
71. [Case C- 158/21, *Puig Gordi and Others*, Judgment of 31 January 2023.](#)
72. [Joined Cases C-514/21 and C-515/21, *Minister for Justice and Equality \(Levée du sursis\)*, Judgment of 23 March 2023.](#)
73. [Case C-699/21, E.D.L. \(Motif de refus fondé sur la maladie\), Judgment of 18 April 2023.](#)
74. [Case C-700/21 O.G. \(Mandat d'arrêt européen à l'encontre d'un ressortissant d'un État tiers\), Judgment of 6 June 2023.](#)
75. [Case C-142/22, The Minister for Justice and Equality \(Demande de consentement – Effets du mandat d'arrêt européen initial\), Judgment of 6 July 2023.](#)
76. [Case C-71/21, Sofiyska gradska prokuratura and Others \(Mandats d'arrêt successifs\), Judgment of 14 September 2023.](#)
77. [Case C-164/22, Juan, Judgment of 21 September 2023.](#)
78. [Case C-636/22, PY \(R ressortissant d'un État tiers dans l'État membre d'exécution\), Order of 16 November 2023.](#)
79. [Case C-261/22, GN \(Motif de refus fondé sur l'intérêt supérieur de l'enfant\), Judgment of 21 December 2023.](#)
80. [Case C-396/22, Generalstaatsanwaltschaft Berlin \(Condamnation par défaut\), Judgment of 21 December 2023.](#)
81. [Case C-397/22, Generalstaatsanwaltschaft Berlin \(Condamnation par défaut\), Judgment of 21 December 2023.](#)
82. [Case C-398/22, Generalstaatsanwaltschaft Berlin \(Condamnation par défaut\), Judgment of 21 December 2023.](#)
83. [Case C-318/24 PPU, Breian, Judgment of 29 July 2024.](#)
84. [Case C-202/24, Alchaster, Judgment of 29 July 2024.](#)
85. [Case C-504/24 PPU, Anacco, Order of 20 September 2024.](#)

86. [Case C-763/22, Procureur de la République \(Concurrence of a European arrest warrant and extradition request\), Judgment of 20 March 2025.](#)
87. [Case C-443/24, Alchaster II, Judgment of 3 April 2025.](#)
88. [Case C-481/23, Sangas, Judgment of 10 April 2025.](#)
89. [Case C-219/25, Kamekris, Judgment of 19 June 2025.](#)
90. [Case C-305/22 C.J. \(Enforcement of a sentence further to an EAW\), Judgment of 4 September 2025.](#)
91. Case C-215/24, Fira, Judgment of 11 September 2025.
92. Case C-798/23, Abbottly, Judgment of 9 October 2025.

2.3 Criteria for the judicial reasoning and organisation of the initial prompts

In order to reinforce judicial reasoning, we established a set of techno-judicial criteria in the analysis, based on keywords deducted from the EAW framework decision:

Keyword	Articles
Additional information	Article 15(2) EAW FD
Admissibility	Article 267 TFEU
Agreement on Surrender between EU, Norway and	
Appeal with suspensive effect, <i>see also</i> time limits	
Arrest warrant (content and form of the EAW)	Article 8(1) EAW FD
Citizenship, <i>see</i> EU citizenship	
Competing requests	Article 16 EAW FD
Consent, <i>see</i> subsequent surrender and <i>see also</i> speciality rule	
Convention (EU) on Extradition (1996)	Articles 31 and 32 EAW FD
Detention	Articles 12 and 26 EAW FD
Double criminality	Articles 2(2), 2(4) and 4(1) EAW FD

Equality (principle of)	
EU citizenship and free movement; extradition of EU citizens	Articles 18 and 21 TFEU
Extraterritoriality	Article 4(7) EAW FD
<i>Force majeure</i> , see also time limits	Article 23 EAW FD
Fundamental rights scrutiny	—Right to be heard (Articles 47 and 48 Charter)
	—Member States' constitutions (Article 53 Charter)
	—Prohibition of inhuman or degrading treatment (Article 4 Charter)
	—Right to respect for private and family life and rights of the child (Articles 7 and 23 Charter)
	—Protection in the event of removal, expulsion or extradition (Article 19 Charter)
	—Right to an independent tribunal
	—Right to a tribunal previously established by law
	—Right to a fair trial (Article 47(2) Charter)

	—Principle that offences and penalties must be defined by law (Article 49(1) Charter)
Fundamental rights	—Right to liberty and security (Article 6 Charter) —Right to an effective remedy and to a fair trial (Article 47 Charter) —Right of defence (Article 48(2) Charter)
Guarantees	Article 5 EAW FD
<i>In absentia</i>	Article 4a(1) EAW FD
Judicial authority (issuing) <i>and</i> effective judicial protection	Article 6(1) EAW FD
Judicial authority (executing)	Article 6(2) EAW FD
Judicial decision, <i>see also</i> arrest warrant	Article 1(1) EAW FD
	Article 8(1)(c) EAW FD
Legality (principle of)	

Minors	Article 3(3) EAW FD
Mutual recognition (principle of)	Article 82(1) TFEU Article 67(3) TFEU
<i>Ne bis in idem</i>	Article 3(2) EAW FD
	Article 4(3) EAW FD
	Article 4(5) EAW FD
	Article 50 Charter
Non-discrimination (principle of) Non-discrimination on ground of nationality	Article 18 TFEU
Postponed surrender	Article 23(4) EAW FD
	Article 24(1) EAW FD
Prison conditions, <i>see</i> fundamental rights scrutiny	
Resident	Article 4(6) EAW FD
Speciality rule	Article 27 EAW FD
Statute barred	Article 4(4) EAW FD
Staying in	Article 4(6) EAW FD
Subsequent surrender	Article 28 EAW FD
Time limits	Article 17 EAW FD

Trade and Cooperation Agreement	Article 632 TCA
	Article 524(2) and Article 604(c) TCA
Transitional regime	—Articles 31 and 32 EAW FD
	—Article 62(1)(b) Withdrawal Agreement (WA)
	—Article 632 TCA
Withdrawal	—Article 50 TEU
	—Article 62(1)(b) WA

2.4 Scraping Methodology and Compliance

In the second phase of the technical setup, we selected highly qualified data sources based on these 4 guiding principles.

The project team followed a structured way of extracting continuous data from digital sources using automated tools or scripts, following defined steps and rules. The implemented methodology covered the choice of targets, how data is fetched and parsed and how the results are cleaned, stored, and used. Core idea All Digital Sources were accessed, read through their output (for example HTML), and specific pieces of information were pulled out into a structured format such as CSV, JSON, or a database.

The basic pillars of the adopted methodology were based on the following check list:

- ❖ **Non-Public Data:** If the data isn't publicly available on the internet, then careful consideration will be needed as to whether this data should be accessed. Legality will be determined through the acquisition of permission or review of the digital source terms
- ❖ **Explicit Agreement to Terms:** If the scrap processor agrees to the digital source's terms of services or other policies in any way, then they must abide by them. They will read the relevant policies in their entirety to determine legality

❖ **Copyrighted Data:** If any of the data being scraped is protected by copyright, then the project team will determine if the use constitutes copyright infringement or is an exception (such as fair use). They will descope the copyrighted data if no exception. The exception to the authors' monopoly stated under Article 3 of Directive (EU) 2019/790 applies only to "research organisations and cultural heritage institutions" acting for scientific research purposes. Users must have "lawful access" to the works, and rightsholders cannot opt out. It authorises, without compensation, the automated reproduction of works for scientific research purposes, but only by public or non-profit organisations. The exception under Article 4 is open to any user, including commercial entities. Rightsholders may "expressly reserve their rights in an appropriate manner," for example through machine-readable metadata, contract terms, or technical measures. The scope of Article 4 is currently being examined in *Like Company v. Google* Version 1.0 December 1, 2025 - 12 - Project 101192405 D2.2 | Dataset Ireland (C-250/25), the first generative AI copyright dispute referred to the Court of Justice of the European Union. The ruling is expected to clarify how EU copyright law applies to generative AI training practices.

❖ **Personal Data:** If any of the data being scraped includes personal data, the project team will ensure compliance with GDPR. In case that this is not possible, then they will descope or anonymize the personal data.

❖ **Improperly Sourced IP Addresses:** The project team will ensure that the IPs being sourced are obtained legally and ethically. If residential IPs are being used, then they will ensure the compliance with applicable data protection laws.

❖ **External Use of Data:** Most scraped data should only be used for internal business analysis. In case the data will be used externally, then the team will ensure the compliance Quality of Data The project team will implement a comprehensive quality assurance (QA) process taking into consideration the pivotal importance of data integrity. The data quality dimensions that will be thoroughly checked are: • accuracy, • validity, • timeliness, • completeness, • uniqueness, and • consistency Cleaning, validation, and storage The project team will be normalising formats (dates, numbers, encodings), removing duplicates, and handling missing or inconsistent values. Then they will loading the data into the dataset with logging for traceability and quality checks. Monitoring, maintenance, and ethics The project team will continuously

monitoring for structural changes in the sources. Moreover, it will periodically reassess legal, ethical, and compliance aspects (e.g. personal data, competition issues, contractual limits) and updating the methodology.

2.5 General Overview of the ‘scraping’ process feeding the Avatars

In this part of the document we describe the structure and composition of a dataset consisting of an initial knowledge hub of 6,339 files, organized into thematic directories mainly related to courts, tribunals, legal portals, and national and European judicial institutions.

The directory-based organization reflects the institutional origin of the documents and allows for a preliminary classification by legal system, level of jurisdiction, and territorial scope.

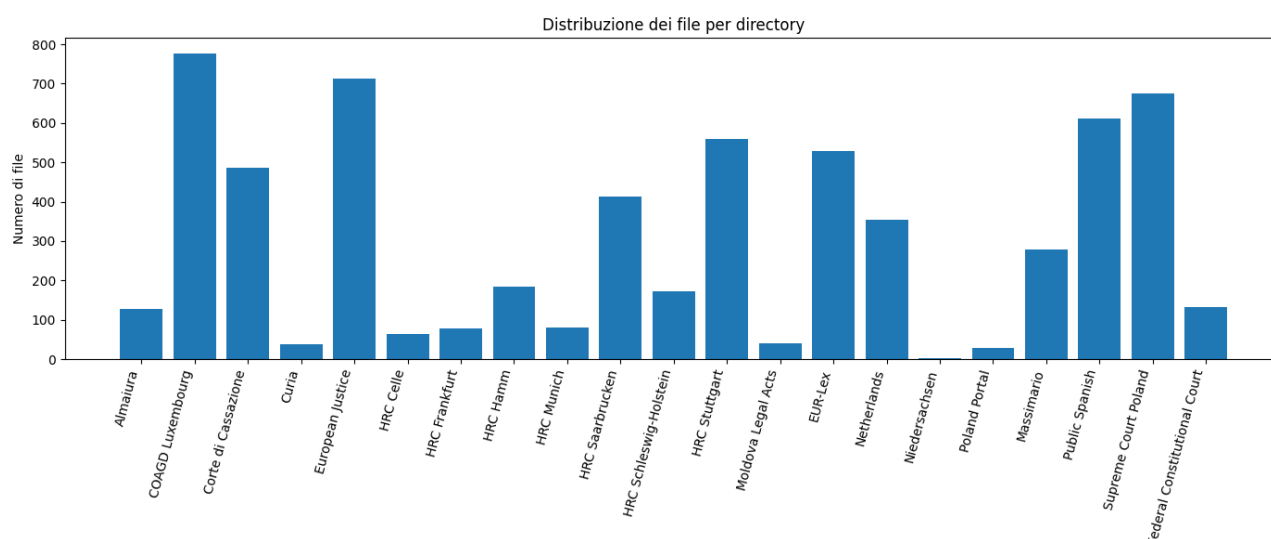
2.5.1 File Distribution by Directory

Below is the list of directories included in the dataset, along with the corresponding number of files.

- Almaiura/ → 127 files
Collection related to academic or legal content, presumably doctrinal in nature or supporting case law.
- COAGD_of_Luxembourg_Investigating_Chamber/ → 777 files
Documentation originating from the Investigating Chamber (Chambre d’instruction) of Luxembourg, with a significant volume of judicial acts.
- Corte_di_Cassazione/ → 487 files
Judgments and decisions of the Italian Court of Cassation.
- Curia/ → 37 files
Limited number of documents, likely related to institutional decisions or communications.
- European_Justice/ → 713 files
Documents of a European scope, potentially including case law and acts of the institutions of the European Union.
- HRC_of_Appeals_Celle_/ → 63 files
- HRC_of_Appeals_Frankfurt/ → 78 files
- HRC_of_Appeals_Hamm/ → 184 files
- HRC_of_Appeals_Munich_/ → 81 files
- HRC_of_Appeals_Saarbrücken_/ → 412 files
- HRC_of_Appeals_Schleswig_Holstein_/ → 173 files
- HRC_of_Appeals_Stuttgart_/ → 559 files

Set of directories referring to the German High Regional Courts (Oberlandesgerichte), with a heterogeneous distribution of the number of decisions by court.

- Legal_acts_of_Republic_of_Moldova_National_Court's_Web_Portal_of_Ministry_of_Justice_of_the_Republic_of_Moldova/ → 40 files
Legal acts originating from the official Moldovan portal of the Ministry of Justice.
- Lex_(Acceso_a_la_legislación_de_la_Unión_Europea)/ → 528 files
Legislative and legal documents extracted from the EUR-Lex portal.
- Netherland_/ → 353 files
Collection of decisions or legal acts originating from the Dutch legal system.
- Niedersachsen_/ → 2 files
Marginal directory with an extremely limited number of documents.
- Portal_Orzeczen_Poland/ → 28 files
Judgments extracted from the Polish judicial decisions portal.
- Portale_del_Massimario/ → 279 files
Legal maxims and summaries of case law, presumably from the Italian legal context.
- Public_Spanish_/ → 611 files
Documentation of Spanish origin, with a substantial volume of public or judicial acts.
- Supreme_Court_of_Poland/ → 674 files
Judgments of the Polish Supreme Court.
- The_federal_Constitutional_Court/ → 133 files
Decisions of the German Federal Constitutional Court.



2.5.2 Quantitative Considerations

- The directories with the highest number of files are:
 - COAGD of Luxembourg – Investigating Chamber (777 files)
 - European Justice (713 files)
 - Supreme Court of Poland (674 files)
 - Public Spanish (611 files)
 - HRC of Appeals Stuttgart (559 files)
- The dataset shows a strong concentration of judicial documents, with particular emphasis on the German, Polish, Italian, and European legal systems.

2.5.3 Update and Data Acquisition Method

The files included in the dataset are collected through an automated scraping script designed to query the respective institutional sources.

The script is executed on a weekly basis, allowing for:

- progressive updating of the document corpus;
- acquisition of new decisions or acts published by the original sources;
- maintenance of a coherent and temporally aligned database.

This acquisition method implies that the number of files per directory is dynamic and subject to change over time.

2.5.4 Purpose of the Dataset

The structure of the dataset makes it suitable for:

- comparative analysis of national and European case law;
- legal and doctrinal research activities;
- text mining and NLP applications on legal texts;
- construction of structured archives of judicial decisions.

This document provides a structural and quantitative description of the collection, serving as an informational basis for subsequent technical or legal analyses.



EuroLegalBot

CRIMINAL MATTERS CHATBOT

EuroLegalBot

Project

Enhancing Judicial Cooperation through a European Arrest Warrant (EAW) Chatbot

DATASET

Deliverable No. 2.2