D13 – Final report on the achievements of the project

Project funded by the European Union – NextGenerationEU through the Recovery and Resilience Plan of the Slovak Republic under project No. 09I05-03-V02-00049

Financované Európskou úniou NextGenerationEU | PLÁN OBNOVY | ÚRAD PODPREDSEDU VLÁDY SLOVENSKEJ REPUBLIKY PRE PLÁN OBNOVY A ZNALOSTNÚ EKONOMIKU | VAIA VÝSKUMNÁ A INOVAČNÁ AKTIVITA | VÝSKUMNÁ AGENTÚRA

Contents

Introduction **3**

1 PhD candidates focusing on issues related to this project **4**

1.1 Definition of the indicator *4*

1.2 Dávid Varga *4*

1.3 Zoltán Szoplák *4*

1.4 Andrej Oriňák *5*

2 Joint projects supported by other schemes **5**

2.1 Definition of the indicator *6*

2.2 APVV-21-0336 — Analysis of court decisions using artificial intelligence methods *6*

2.3 VV-MVP-24-0038 — Analysis of liability for offences committed on the internet using machine learning methods *7*

3 Research publications disseminating the project's results **9**

3.1 Definition of the indicator *9*

3.2 List of publications *9*

4 National and international collaborations in the academic sphere and with the public and private sectors **12**

4.1 Definition of the indicator *12*

4.2 National collaborations *12*

4.3 International collaborations ..... ***Chyba! Záložka nie je definovaná.***

5 Activities aimed at disseminating results to potential partners and customers **14**

5.1 Definition of the indicator *14*

5.2 ALTAML final conference *14*

5.3 Conference: 'Law and Innovation — Is the Legal Profession Ready for Artificial Intelligence?' *14*

5.4 Workshop: 'Analysis of Court Decisions' *15*

5.5 Workshop on 'Automated Analysis of Legal Texts' at the conference 'Milestones of Law in the Area of Central Europe 2026' *16*

5.6 Workshop at the academic conference for PhD students and early-career researchers 'Law in an Era of Global Instability: Resilience, Security and New Challenges' *16*

6 Published datasets **17**

6.1 Definition of the indicator *17*

6.2 List of deliverables *17*

7 Published methodologies **21**

7.1 Definition of the indicator *21*

7.2 List of deliverables 21

8 Published methods 24

8.1 Definition of the indicator 24

8.2 List of deliverables 24

Introduction

This document details the achievement of the measurable indicators for the ALTAML project. For each indicator, it sets out its definition, the method used to determine the target value, the requirements for demonstrating achievement, and the specific results achieved within the project. It serves as a basis for monitoring reports as well as an overview for project researchers.

A summary of performance is set out in the table below. Details for each indicator are provided in the individual chapters.

Expected results and impact of project activities	Expected number	Actual achievement
PhD candidates focusing on issues related to this project	**3**	**3**
Joint projects supported by other schemes	**2**	**2**
Research publications disseminating the project's results	**4**	**5**
National collaborations within the academic sphere and with both the public and private sectors	**3**	3
International collaboration in the academic sphere and with the public and private sectors	**2**	**3**
Activities aimed at disseminating results to potential partners/customers	**4**	**5**
Published datasets	**3**	**3**
Published methodologies	**2**	**2**
Published methods	**14**	**14**

1 PhD candidates focusing on issues related to this project

Expected number	Actual achievement	Status
3	3	Achieved

1.1 Definition of the indicator

This indicator was estimated on the basis of current and planned PhD positions, the topics of which reflect the issues defined in the ALTAML project. The indicator tracks the development of research capacity in the field of automated analysis of legal texts.

1.2 Automatic attribute extraction and analysis of court decisions (Dávid Varga)

Data	Value
Thesis topic	Automatic attribute extraction and analysis of court decisions
Student	RNDr. Dávid Varga
Department	Institute of Computer Science, Faculty of Science, UPJŠ
Supervisor	Prof. RNDr. Stanislav Krajči, PhD.
Year the topic was announced	2021
Supporting documentation	Dávid_Varga_thesis_assignment_SK_EN

Link to the project: the topic directly addresses the core content of work package KPB2 – the extraction of attributes and relationships from court decisions. The PhD student is a co-author of the project's published deliverables and co-led the workshop 'Analysis of Court Decisions'.

1.3 The use of automatic story summarisation methods for the analysis of court decisions (Zoltán Szoplák)

Data	Value
Thesis topic	The use of automatic story summarisation methods for the analysis of court decisions
Student	RNDr. Zoltán Szoplák

Data	Value
Department	Institute of Computer Science, Faculty of Science, UPJŠ
Supervisor	Prof. RNDr. Stanislav Krajči, PhD.
Year the topic was announced	2020
Supporting documentation	Zoltán_Szoplák_thesis_assignment_SK_EN

Link to the project: automatic summarisation is a prerequisite for making the content of extensive court decisions accessible and ties in with the tasks of work packages KPB2 and KPB4. The PhD student is a co-author of a project publication presented at the JURIX 2025 conference.

1.4 Data analysis of legislation and case law in commercial law and information technology law (Andrej Oriňák)

Data	Value
Thesis topic	Data analysis of legislation and case law in commercial law and information technology law
Student	JUDr. Andrej Oriňák
Department	Faculty of Law, UPJŠ in Košice
Supervisor	Doc. JUDr. Regina Hučková, PhD.
Year the topic was announced	2023
Supporting documentation	Andrej_Oriňák_thesis_assignment_SK_EN

Link to the project: the topic combines legal analysis with data methods and corresponds to the legal strand of the project, namely the interpretation of attributes extracted from legal texts in practical application.

2 Joint projects supported by other schemes

Expected number	Actual achievement	Status
2	2	Achieved

2.1 Definition of the indicator

The estimate was based on ongoing and planned project proposals relating to the issues addressed in the ALTAML project. The indicator demonstrates that the project's theme is being systematically developed even outside the scope of its own funding.

2.2 APVV-21-0336 — Analysis of court decisions using artificial intelligence methods

Data	Value
Project number	APVV-21-0336
Scheme	Slovak Research and Development Agency (APVV) — General Call for Proposals VV 2021
Lead organisation	Pavol Jozef Šafárik University in Košice
Project duration	1 July 2022 – 30 June 2026
Amount of funding	€249,959.00

Link to the contract: Central Register of Projects

Objective: The main objective of the project is to lay the groundwork for the wider use of machine learning methods in the decision-making processes of courts. The sub-objectives are to identify and extract attributes and relationships from court decisions and to develop a pilot search tool for locating relevant decisions (including methods for extracting explicit and implicit attributes, and for representing and storing the extracted attributes and relationships), and the analysis of attributes extracted from court decisions.

Abstract: The use of artificial intelligence and innovative technologies in the public sector, particularly in the decision-making processes of courts, is a subject of current interest to the European Union, as stated in several official documents. The decision-making process of a court requires knowledge of legal principles and rules, as well as an understanding of legal texts. The research objectives of the project are therefore focused on extracting knowledge from the texts of court decisions issued by the courts of the Slovak Republic and their automated analysis — in relation to legal institutions and principles in practical application, as well as to the manner in which evidence is presented in court proceedings. The research team comprises researchers from the n the fields of legal theory, information and communication

technology law, procedural law, data analysis, computer science and natural language processing.

Connection to the ALTAML project: this project addresses the same subject matter, namely the automated analysis of court decisions. This project is the direct methodological predecessor of the ALTAML project. The overlap lies primarily in the extraction of attributes and relationships from court decisions (work package KPB2 of the ALTAML project) and in the retrieval of relevant decisions (KPB3 and KPB4). ALTAML builds on these results and extends them to include the processing of legislation, a semantic graph of the extracted data, and the production-ready dissemination of results. There is also an overlap in the research teams of both projects.

2.3 VV-MVP-24-0038 — Analysis of liability for offences committed on the internet using machine learning methods

Data	Value
Project number	VV-MVP-24-0038
Scheme	APVV — public call for young researchers' projects 2024
Lead organisation	Pavol Jozef Šafárik University in Košice
Project duration	1 March 2025 – 31 December 2028
Amount of funding	€170,250

Link to the contract: Central Register of Projects

Objective: The main objective of the project is to analyse liability for private and public law offences committed in the digital environment and the methods of sanctioning them, based on a review of relevant legal sources and the design and application of machine learning methods. Sub-objectives include defining the legal framework for liability and classifying categories of offences, creating a coded dataset of extracted legal texts alongside the design of automated methods for extracting selected attributes, and finally analysing liability, comparing methods of sanctioning, and formulating proposals de lege ferenda.

Abstract: The incidence of unlawful conduct in the digital environment of the internet is constantly rising. The publication of illegal content or other unlawful activities on the internet has the potential to fulfil the elements of numerous torts under private or public law. Discussions centre not only on the liability of the perpetrators themselves, but also on the possibility of establishing liability for intermediaries whose services are used for unlawful activities. The project envisages the creation of a separate, coded or annotated dataset of legal texts, in particular court judgements and decisions by other state authorities, concerning offences committed on the internet.

Connection to the ALTAML project: the project utilises the same methodological core, namely the data analysis of legal texts with the implementation of machine learning methods. It applies this approach to a specific subject area. It builds directly on the annotation activities and the creation of coded datasets of legal texts (KPB1 and KPB2 of the ALTAML project) and extends the scope of the analysed legal texts to include decisions by state authorities other than courts.

3 Research publications disseminating the project's results

Expected number	Actual achievement	Status
4	5	Exceeded

3.1 Definition of the indicator

The estimate was based on the number of previous scientific articles published by key members of the research team. The indicator includes peer-reviewed publications that disseminate the project's results within the professional community. The publications are deliverable D24 — Peer-reviewed scientific articles.

3.2 List of publications

Sokol, P., Bachňáková Rózenfeldová, L., Varga, D., Rudohradská, S., & Hučková, R. (2025). **Anomaly detection in legal text — A case study on cybersecurity contracts**. In R. Markovich, L. Di Caro, A. Rapp, & C. Schifanella (Eds.), Legal Knowledge and Information Systems: JURIX 2025 (Frontiers in Artificial Intelligence and Applications, Vol. 416, pp. 350–355). IOS Press.

Link to the article: https://doi.org/10.3233/FAIA251609

Abstract**:** The growing availability of legal documentation creates opportunities for the application of empirical research methods that enable a comprehensive analysis of complex legal data. Systematic Content Analysis offers a structured approach to the examination of legal texts. This paper examines the possibility of detecting anomalies in legal texts through the use of machine learning methods. The conclusions presented are based on a case study focusing on contracts concluded between public authorities and private service providers in the field of cyber security. In this context, we propose an approach to detecting anomalous legal text, providing a categorisation of the attributes that influence the identification of anomalies and describing their significance within the chosen domain.

Horvát, Š., Varga, D., Gurský, P., Fedurcová, N., Baran, S., & Szoplák, Z. (2025). **A self-supervised method for legal reference resolution**. In R. Markovich, L. Di Caro, A. Rapp, & C. Schifanella (Eds.), Legal Knowledge and Information Systems: JURIX 2025 (Frontiers in Artificial Intelligence and Applications, Vol. 416, pp. 300–305). IOS Press.

Link to the article: https://doi.org/10.3233/FAIA251601

Abstract: Distinguishing references to legal provisions in court decisions is challenging when abbreviations correspond to multiple candidate legal provisions. At the same time, however, it is crucial, as reliable reference disambiguation supports the retrieval of legal information. We present a self-supervised graph-based method that constructs a weighted co-occurrence graph from court decisions, in which nodes represent provisions of legal legislation and edge

weights reflect the frequency of their co-occurrence. Ambiguous references are resolved by measuring distances from contextual provisions using a multi-source Dijkstra's algorithm, whereby the nearest candidate is selected. Experiments show that the proposed approach improves accuracy compared with simple reference solutions, provides a reliability score, and represents a practically applicable method of differentiation. The study also includes an annotated dataset intended for further research.

Sokol, P., & Bachňáková Rózenfeldová, L. (2025). **Prospects for the use of the systematic contextual analysis method in legal research**. Forum Iuris Europaeum, 13(2). ISSN 2644-4364.

Link to the article: http://fie.iuridica.truni.sk/wp-content/uploads/2025/12/FIE-2-2025_na-web.pdf

Abstract: The aim of this paper is to present the method of systematic contextual analysis and its individual stages, and to explore the possibilities for its application in legal research. In this paper, we outline the advantages and disadvantages associated with applying this approach, whilst taking into account the specific characteristics of the Slovak legal environment, particularly the free availability of various types of legal texts. Based on the authors' research to date, we present specific examples of the use of this method, as well as the specific features associated with it.

Rudohradská, S. (2025). **Practical application of the Digital Services Act — the case of the Slovak Republic.** EU and Comparative Law Issues and Challenges Series (ECLIC). Paper presented at the international academic conference 'Strong and Secure Europe: Legal and Economic Aspects'.

Link to the article: https://doi.org/10.25234/eclic/38129

Abstract: The Digital Services Act is currently the subject of considerable attention not only from experts but also from the general public. As stated in Article 93 of the DSA, the Regulation will apply in full from 17 February 2024. Member States were required to transpose its provisions into their national legal systems and commence its practical implementation. This paper addresses issues relating to the transposition of the Digital Services Act into the legal system of the Slovak Republic, which transposed the DSA through an amendment to Act No. 264/2022 Coll. on Media Services. In the Slovak Republic, we are currently witnessing the first decisions by the Media Services Council imposing an obligation to remove unlawful content. The aim of this paper is to examine the process of implementing and applying the Digital Services Act and to analyse the relevant decisions of the Media Services Council, with a focus on comparing the concept of unlawful content under Slovak national law and European Union law.

Szoplák, Z., Varga, D., Gurský, P., Horvát, Š., & Krajči, S. (2025). **Enhancing document retrieval in large corpora: A keyphrase and reference-based approach.** In K. Satoh & K. Nitta (Eds.), *Local Proceedings of the 19th International Workshop on Juris-Informatics (JURISIN 2025) in association with the 17th JSAI International Symposia on AI (JSAI-isAI 2025)*, 26–27 May 2025.

Link to the article:
https://jurisinformaticscenter.github.io/jurisin2025/jurisin2025localproceedings.pdf

Abstract: Document retrieval, particularly in large corpora, remains a challenging task despite recent advances in natural language processing (NLP) and large language models (LLMs). The context windows of these models are often not large enough to accommodate a large number of documents. To address this problem, many NLP tasks — such as document retrieval for RAG systems — employ a two-step procedure: first, the most relevant documents are filtered out, and these are then used as input for a language model that ranks them. However, the larger the corpus, the greater the need for a robust retrieval method. The authors propose a solution that extends the standard retrieval method, based on the similarity of vector representations, to include a score derived from the match of extracted key phrases, as well as an extension utilising references in legal documents. The first approach achieved a mean average precision (mAP) of 80.25 %; the second improved this result to 92.85 %.

4 National and international collaborations in the academic sphere and with both the public and private sectors

Type of collaboration	Expected number	Actual achievement	Status
National collaborations	3	3	Fulfilled
International collaborations	2	3	Exceeded

4.1 Definition of the indicator

The estimate was based on existing collaborations and the potential for future collaborations with entities that have a natural interest in the project's results.

A Confirmation of Interest in Collaboration has been signed with each of the organisations listed.

4.2 National collaborations

Organisation	Sector
Ministry of Justice of the Slovak Republic	Public sector
Institute of State and Law, Slovak Academy of Sciences, v. v. i.	Academic sector
Slovak Bar Association	Professional body

The Ministry of Justice of the Slovak Republic is the central state administration body responsible for the judiciary and the prison service, and the administrator of the justice sector's information systems, including sources of court decisions. As the body responsible for the digitisation of the justice system, it is a natural user of the project's results in making case law accessible and analysing it.

The Institute of State and Law of the Slovak Academy of Sciences (**SAV**) is a research unit of the Slovak Academy of Sciences focused on fundamental research in the fields of legal theory, constitutional law and legal history. Its research focus complements the legal strand of the project, particularly in the interpretation of attributes extracted from legal texts.

The Slovak Bar Association is a self-governing professional body established by the Act on the Legal Profession, which brings together all lawyers in the Slovak Republic, maintains a register of lawyers and ensures their training and disciplinary accountability. Lawyers represent the

largest target group of end-users of legal information systems, as confirmed by the survey carried out as part of deliverable D01, and the Bar Association is therefore a natural partner in verifying the practical applicability of the project's results and in disseminating them into legal practice.

4.3 International collaborations

Organisation	Country
Palacký University in Olomouc, Faculty of Law	Czech Republic
Institute of State and Law, Czech Academy of Sciences	Czech Republic
Tallinn University of Technology (TalTech)	Estonia

Palacký University in Olomouc, Faculty of Law, is one of the leading Czech law faculties with a long-standing focus on information technology law and the use of modern technologies in legal practice. A representative from the faculty spoke at the 'Law and Innovation' conference organised as part of the project.

The Institute of State and Law of the Czech Academy of Sciences is a research unit of the Czech Academy of Sciences focused on theoretical and interdisciplinary legal research, including law and technology. This collaboration facilitates an international comparison of approaches to the automated analysis of case law.

Tallinn University is an Estonian public university operating within one of the most digitised public administrations in Europe and dedicated to digital transformation and interdisciplinary research. This collaboration brings to the project experience in advanced e-government and digitised justice.

5 Activities aimed at disseminating results to potential partners and customers

Expected number	Actual achievement	Status
4	5	Exceeded

5.1 Definition of the indicator

The estimate was based on the planned activities for disseminating results within the project, namely conferences and workshops. These activities fulfil the deliverables of work package KPB5, namely D30 – Conference and D31 – Workshop.

5.2 ALTAML Final Conference

Data	Value
Date	24 June 2026
Venue	Amphitheatre, 2nd floor of the old building, Faculty of Law, Comenius University in Bratislava
Type	Conference

Link: https://altaml.upjs.sk/news/konferencia-altaml-otvorila-diskusiu-o-zodpovednom-vyuzivani-umelej-inteligencie-v-sudnictve/

Abstract: The conference was dedicated to the results of the project and an expert discussion on the possibilities, risks and conditions for the responsible use of artificial intelligence in the judiciary. The event brought together perspectives from public administration, judicial practice and the academic community. A key part of the event was the round table discussion 'AI in the courts: decision-making support, training and the limits of human control', which focused on the use of artificial intelligence in legal research, working with case law, summarising case files, and the training of judges and court officials. The discussion also addressed the risks of generative artificial intelligence — lack of transparency, inaccuracies, hallucinations and algorithmic bias — and the need to maintain meaningful human oversight.

5.3 Conference: 'Law and Innovation – Is the Legal Profession Ready for Artificial Intelligence?'

Data	Value
Date	14 November 2025
Venue	Faculty of Law, UPJŠ, Kováčska 26, Košice
Type	Conference

Link: https://altaml.upjs.sk/news/je-pravna-prax-pripravena-na-ai-konferencia-pravo-a-inovacie-otvorila-klucove-temy/

Abstract: The conference was organised by the Department of Commercial Law and Business Law in collaboration with the Institute of Informatics at the Faculty of Natural Sciences. The event brought together experts from academia, the legal profession, the judiciary and the technology sector. Representatives from the Ministry of Justice of the Slovak Republic, the Faculty of Law at Comenius University, the Slovak Bar Association and Palacký University in Olomouc spoke at the expert forum. The panel discussions focused on the application of artificial intelligence in day-to-day legal work, the automation of legal processes, the future of the legal profession and education. The programme included a presentation of the interim results of the ALTAML project and an introduction to the tools provided by partner Wolters Kluwer.

5.4 Workshop: 'Analysis of Court Decisions'

Data	Value
Date	14 May 2025
Venue	Faculty of Natural Sciences, UPJŠ, Jesenná 5, Košice
Type	Workshop

Link: https://altaml.upjs.sk/news/workshop-analyza-sudnych-rozhodnuti-pomocou-dat-a-strojoveho-ucenia/

Abstract: The workshop was led by members of the research team – RNDr. Dávid Varga and JUDr. Simona Rudohradská, PhD. Participants learnt about the use of data analytics and machine learning methods in the analysis of court decisions, with an emphasis on graph structures that enable individual decisions to be represented as nodes and their interrelationships to be visualised. They gained hands-on experience in representing decisions as graphs, visualising relationships between decisions, and identifying anomalous decisions using the outlier detection method. The session concluded with a demonstration of the COIN (Contextual Outlier Interpretation) method. The event was aimed at students, researchers and experts with an interest in law, data analytics and natural language processing.

5.5 Workshop on 'Automated Analysis of Legal Texts' at the conference 'Milestones of Law in Central Europe 2026'

Data	Value
Date	27 March 2026
Venue	Special-purpose facility of the **Office** of the National Council of the Slovak Republic, Častá – Papiernička
Type	Workshop

Link: https://altaml.upjs.sk/news/workshop-automated-analysis-of-legal-texts/

Abstract: The workshop was led by JUDr. Zoltán Gyurász, PhD. It focused on the specific characteristics of legal texts, which, whilst formally precise, represent complex structures that are highly context-dependent from the perspective of automated processing. Attention was focused on the shift from traditional keyword-based approaches to advanced methods of computational understanding — natural language processing, information extraction and machine learning methods. The event included demonstrations of solutions inspired by the research activities of the ALTAML project and a discussion on the need for a critical approach to artificial intelligence, particularly in relation to bias and transparency.

5.6 Workshop at the academic conference for PhD students and early-career researchers 'Law in an Era of Global Instability: Resilience, Security and New Challenges'

Data	Value
Date	30 April 2025
Venue	Faculty of Law, UPJŠ, Kováčska 26, Košice
Type	Workshop

Link: https://altaml.upjs.sk/news/altaml-workshop-na-konferencii-pre-doktorandov/

Abstract: The conference included a specialist workshop on the use of machine learning in the analysis of court decisions, led by Simona Rudohradská from the Department of Commercial and Economic Law. The workshop demonstrated that the convergence of law, data analytics and artificial intelligence is no longer a vision of the future, but a reality that is gradually changing the way we work with legal texts. Both the conference and the workshop are outcomes of the ALTAML project.

6 Published datasets

Expected number	Actual achievement	Status
3	3	Achieved

6.1 Definition of the indicator

The estimate was based on the number of datasets that need to be prepared to achieve the project's objectives. The indicator covers 'dataset' deliverables intended for publication.

6.2 List of deliverables

Number	Deliverable	Work package	Access
D04	Sample data	KPB1.4	Public
D08	Database of Slovak and English legal phrases	KPB2.4	Public
D14	Semantic graph of extracted data	KPB3.1	Public

The expected value for the three published datasets has been fully achieved. All three datasets are intended for public access in open repositories in accordance with the principles of open science and the FAIR principles.

6.3 D04 – Sample data

Data	Value
Deliverable number	D04
Work package / task	KPB1, task KPB1.4
Responsible partner	Wolters Kluwer SR s. r. o.
Type	dataset
Access and dissemination	public — published in open repositories
Document with a detailed description	ALTAML-D04-KPB1.4-Sample-data (EN)
Availability	published on the project website https://altaml.upjs.sk/

Summary: The deliverable consists of a demonstration dataset of court decisions made available simultaneously in three formats — the original PDF, transformed XML and the resulting HTML — together with a complete set of metadata (source, URL, internal ID, case case reference number, date, type and title of the decision, summary, issuing court and the names of individual files). The dataset covers seven sources of judicial decisions: the website of the Ministry of Justice of the Slovak Republic (justice.gov.sk), the websites of the Supreme Court of the Slovak Republic, the Supreme Administrative Court of the Slovak Republic and the Constitutional Court of the Slovak Republic, as well as three collections of decisions and opinions from these courts. Given the scope of the processed sources, only a representative selection of the processed documents has been published for demonstration purposes. An integral part of the deliverable consists of two sets of accompanying documentation for the data packages: Part I describes 17 data packages for the domains of copyright and trade marks/unfair competition, whilst Part II describes 14 data packages for the domain of digital evidence in criminal proceedings. Each of the 31 packages is accompanied by a results set in .xlsx format (datovy-balik-1.xlsx to datovy-balik-31.xlsx) with an 'altaml' worksheet containing all documents and, in most cases, also a 'po_redukcii' sheet with a narrowed-down selection based on the analysis of the case reference number (registers 'Ca' and 'Co' for copyright, 'CbPV' and 'CoPV' for trade marks, 'T', 'To', 'Tdo', 'Tk', 'Nt' and 'ÚS' for criminal matters).

Link to project objectives: The deliverable fulfils Specific Objective 1 of the project (identification of a suitable method for processing legal text in selected case studies) and also forms the input data set for all subsequent work packages KPB2 to KPB5.

6.4 D08 – Database of Slovak and English legal phrases

Data	Value
Deliverable number	D08
Work package / task	KPB2, task KPB2.4
Responsible partner	Pavol Jozef Šafárik University in Košice
Type	dataset
Access and dissemination	public — published in open repositories
Document containing a detailed description	ALTAML-D08-KPB2.4-Database-of-Slovak-and-English-Legal-Phrases (EN)
Availability	Dataset and accompanying documentation in the UPJŠ GitLab repository

Summary: The deliverable is a bilingual database of legal terms and their definitions, published simultaneously in .csv, .xlsx and .json formats. The terms were collected from

publicly available online sources (Slov-Lex, Najpravo.sk) and supplemented with selected section headings from the Collection of Laws of the Slovak Republic; as not every section heading is a legal term, the headings were filtered using a combination of statistical and semantic metrics and by prompting large language models. The files `pravne_pojmy_sk_en.csv` and `pravne_pojmy_sk_en.xlsx` contain bilingual lists of terms (columns SK and EN), whilst the file `legal_terms_sk_en.json` contains an extended structure with fields `_id` (Slovak term), definitions (a list of definitions with their source; for sources marked 'zz', this also includes the identification of the Act and the heading within its structure) and english (the English equivalent and its definition). Where the text of a section was identified as a definition, the definition was extracted directly from its wording. The English equivalents and definitions were created by translating and summarising the Slovak definitions using the large language model LLaMA 3.3; their primary purpose is not to provide a definition that is accurate for human readers, but to generate vector representations of concepts, where context and thematic consistency are paramount.

Link to project objectives: The deliverable fulfils specific objective 2 of the project (extraction of explicit and implicit attributes from legal text). The database forms a closed set of candidate concepts for deliverable D09 — a Method for assigning key-phrases, thereby ensuring terminological accuracy and consistency of extraction across the entire corpus.

6.5 D14 — Semantic graph of extracted data

Data item	Value
Deliverable number	D14
Work package / task	KPB3, task KPB3.1
Responsible partner	Pavol Jozef Šafárik University in Košice
Type	dataset
Access and dissemination	public — published in open repositories
Document with a detailed description	ALTAML-D14-KPB3.1-Semantic-graph (EN)
Availability	the resulting RDF graph in the file model.rdf; ontology prefixes https://court.ics.upjs.sk/ontology/ and https://court.ics.upjs.sk/resource/

Summary: The deliverable is a unified RDF graph containing metadata on decisions of the Court of Justice of the European Union and Slovak courts in a common semantic format, including RDFS inference, published in the file model.rdf. The proposal is based on an analysis of the ontological layer of the Publications Office of the European Union, which comprises the Common Data Model (CDM) for document structure and relationships, SKOS/EuroVoc for

subject classification, and Dublin Core Terms for bibliographic attributes. As the Ministry of Justice of the Slovak Republic publishes not only case law but also ordinary first- and second-instance decisions, which are not recognised by the CDM and which have their own attributes (district and region of the court, nature and form of the decision, area and sub-area of legislation), the research team introduced a new superclass, `upjs_ontology:court_decision`, within the hierarchy `cdm:case-law_work ⊑ upjs_ontology:court_decision ⊑ cdm:Work ⊑ owl:Thing`. The extension was implemented exclusively 'upwards' via `subClassOf` and `subPropertyOf`, thereby preserving the integrity of the original CDM model, and every European court decision automatically becomes a valid instance of the new class . Four new classes (decision_nature, decision_form, legal_area, legal_sub_area) and thirteen new properties were proposed, covering the ECLI, case case reference number (spisová značka) and file identification number (ICS), links to the judge, court and country of origin, the district → region hierarchy, and the type of court.

Link to project objectives: This deliverable fulfils Specific objective 3 of the project (representation of data extracted from legal texts). It demonstrates that the data model used by EU institutions is sufficiently versatile for Slovak case law as well, without the need for further conversion, thereby laying the groundwork for the creation of a unified, interoperable database of decisions.

7 Published methodologies

Expected number	Actual achievement	Status
2	2	Achieved

7.1 Definition of the indicator

The estimate was based on the number of methodologies that need to be developed to achieve the project's objectives. Unlike individual methods and algorithms, these are comprehensive methodological frameworks that determine how a given task is approached.

7.2 List of deliverables

Number	Deliverable	Work package	Type	Access
D01	End-users´ survey	KPB1	Survey	Public
D15	Storage design	KPB3	Proposal	Public

The expected value of the two published methodologies was fully achieved. The methodological framework for systematic contextual analysis of law, which was validated and evaluated in Deliverable D23 — Evaluation Report, was also published in the form of a peer-reviewed scientific article (Chapter 3).

7.3 D01 – End-users´ survey

Data	Value
Deliverable number	D01
Work package / task	KPB1, task KPB1.1
Responsible partner	Wolters Kluwer SR s. r. o.
Type	survey / methodological framework
Access and dissemination	public — published on the project website
Document with a detailed description	ALTAML-D01-KPB1.1-End-users-survey (EN), ALTAML-D01-KPB1.1-End-users-survey (EN)
Availability	https://altaml.upjs.sk/wp-content/uploads/sites/29/2025/07/Uzivatelsky-

Data	Value
	prieskum_altaml_WK_2025_04_04.pdf and https://altaml.upjs.sk/wp-content/uploads/sites/29/2025/07/ALTAML_Users_survey_en.pdf

Summary: The aim of the survey was to identify the needs, practices and expectations of those who regularly work with legal texts, and to ascertain how end-users perceive current legal information tools, what they find lacking in them, and what features would enhance the efficiency of their work. The survey involved 102 legal professionals from various areas of law and regions of the Slovak Republic — attorneys and lawyers in the private sector, in-house lawyers at companies, and employees of the public administration and the judiciary. Data collection combined a questionnaire with structured and open-ended questions with direct in-depth interviews with selected respondents; the questionnaires were anonymous and the data were processed in accordance with Regulation (EU) 2016/679. Once data collection was complete, the relevance of the responses was assessed, and incomplete responses, blank forms and responses lacking meaningful information were excluded. The survey yielded three key findings: the need for a fast, accurate and reliable search function, the results of which are immediately usable without further verification; a demand for concise summaries of court decisions and for the extraction of keywords from those decisions; and a recurring call for a tool to predict the outcome of legal disputes. The deliverable includes two presentations — our own survey of 102 respondents and the international 2024 Wolters Kluwer Future Ready Lawyer Survey, conducted amongst 712 lawyers in the US and nine European countries, which provides a broader frame of reference.

Link to project objectives: The deliverable fulfils Specific Objective 1 of the project and provides methodological guidance for all subsequent work packages. The three identified requirements were directly incorporated into the project deliverables: the requirement for search functionality in D16 and D17, the requirement for summaries and keywords in D06, D09 and D10, and the requirement for dispute outcome prediction in D21.

7.4 D15 – Repository design

Data	Value
Deliverable number	D15
Work package / task	KPB3, task KPB3.2
Responsible partner	Pavol Jozef Šafárik University in Košice
Type	Architecture proposal / methodological framework
Access and dissemination	public — publication in open repositories

Data	Value
Document with a detailed description	ALTAML-D15-KPB3.2-Repository design (EN)
Availability	architecture documentation; components available in the UPJŠ GitLab repositories

Summary: This deliverable defines a modular data architecture for a system designed to automate the processing, enrichment and publication of court decisions. At its core is the ALTAML decision retriever component, which handles the retrieval of decisions from external sources, their conversion, versioning and the sequential application of independent extraction methods. The architecture combines three types of storage according to the processing phase: a MySQL relational database performs a control and reference function with a central 'raw_decision' table, court codes and stored JSON schemas for external APIs, thereby reducing the load on the Ministry of Justice's servers; the MongoDB document database stores the final, enriched decisions, including the version history of extraction methods, which allows users to revert to an older version in the event of a faulty method, as well as to run a new method on decisions that have already been processed without repeating the entire process; and Elasticsearch forms the final search layer with its own analyser based on the Slovak Hunspell dictionary. The proposal further describes four separate services and their repositories — the Tvaroslovník API for lemmatisation and grammatical categories, Referex for link extraction with five supporting MongoDB collections, Keyphrase Extractor for identifying key terms, and Annotate as a search and annotation interface. The document also transparently records a methodological change from the original plan: the proposal for a custom query language and a separate middleware layer has been replaced by the use of Elasticsearch's native query mechanism, which has proven to be sufficient and, moreover, allows for the combination of attribute-based, full-text and semantic search.

Link to project objectives: The deliverable fulfils specific objective 3 of the project (representation of extracted data and design of a suitable search method). The architecture is a direct prerequisite for deliverables D16 and D17 and, at the same time, ensures scalability, reproducibility of processing and the sustainability of results after the project's completion.

8 Published methods

Expected number	Actual achievement	Status
14	14	Achieved

8.1 Definition of the indicator

The estimate was based on the number of methods and algorithms that need to be developed to achieve the project's objectives. The indicator covers deliverables of the types 'method', 'algorithm' and 'model' across work packages KPB1 to KPB4. All are intended for publication.

8.2 List of deliverables

Number	Deliverable	WP	Type
D02	Method of processing legal texts	KPB1	Method
D03	Methods for continuous downloading	KPB1	Method
D05	Method for the extraction of references	KPB2	Method
D06	Method for verdict extraction	KPB2	Method
D07	Method for extraction of process-related phrases	KPB2	Method
D09	Method for assigning key-phrases	KPB2	Method
D10	Semi-automatic methods for segmenting legal texts	KPB2	Method
D16	A query model for filtering legal text	KPB3	model
D17	Middleware search method	KPB3	Method
D18	Relevance assessment method	KPB4	Method
D19	Algorithm for legal text linking	KPB4	algorithm
D20	Method for creating and describing clusters	KPB4	Method
D21	Model for predicting the verdicts	KPB4	model
D22	Algorithm for identifying outliers	KPB4	algorithm

The expected value for the fourteen published methods was achieved in full. The indicator has been met. The methods from deliverables D05 and D22 were also published in the form of

peer-reviewed scientific articles — a method for extracting and distinguishing references to legislation and a method for detecting anomalies in legal texts (Chapter 3). The source codes for all methods are managed in publicly accessible repositories in accordance with the open access principle declared in the project description.

8.3 D02 – Method of processing legal texts

Data	Value
Deliverable number	D02
Work package / task	KPB1, task KPB1.2
Responsible partner	Wolters Kluwer SR s. r. o.
Type	Method
Access and dissemination	public — publication in open repositories
Document with a detailed description	ALTAML-D02-KPB1.2-Method-for-processing-legal-texts (EN)
Availability	methodological document published on the project website

Summary: The method presents a comprehensive, reproducible procedure for processing court decisions, consisting of four layers with defined inputs, deliverables and checks: data acquisition (crawling, API, manual downloading), transformation (PDF → XML → HTML, XSLT, relinking), quality and enrichment (validation, metadata mapping, relationship inference) and publication and export (CMS, data packages, HTML deliverables). During bulk acquisition, PDF documents are converted to XML using a script that utilises poppler-utils and XSL transformations; the application then adds links to legal provisions, court rulings and EU legislation, and the documents are imported into the content management system, where XML schema validation takes place, along with checks on mandatory metadata (type, date, case reference number, ECLI), consistency of hyperlinks and anti-duplication rules; the schedule of runs is managed by cron, and deduplication is carried out based on the file hash and case reference number. In the manual process, OCR is applied to image-based PDFs using ABBYY FineReader, and decision compendia are segmented according to headings, numbering and templates. The document describes in detail four automated sources (justice.gov.sk, including API endpoints; nsud.sk; nssud.sk, with five categories of judicial activity; and ustavnysud.sk) and three manual sources of decision collections. The method also includes a tool for exporting metadata and creating data packages, which, based on user-specified SQL queries, generates result sets with a uniform metadata structure (Id, aspi_id, author, spzn, type, date, source, pub_number, full_source, area, relation, register, year, portion, edition).

Link to project objectives: This deliverable fulfils Specific Objective 1 of the project. The method is intended for any organisation interested in processing legal texts and forms the technical basis for the creation of dataset D04.

8.4 D03 – Methods for continuous download

Data item	Value
Deliverable number	D03
Work package / task	KPB1, task KPB1.3
Responsible partner	Pavol Jozef Šafárik University in Košice
Type	Method
Approach and dissemination	public — publication in open repositories
Document with a detailed description	ALTAML-D03-KPB1.3-Methods-for-continuous-downloading (EN)
Availability	https://gitlab.science.upjs.sk/altaml/decisions-downloader

Summary: The deliverable is a highly resilient system for the continuous downloading of court decisions, implemented in Python using a microservices architecture and communication via the RabbitMQ message broker. The system automatically retrieves the most recently published judgements from the Slovak Ministry of Justice portal, processes and transforms the downloaded metadata and PDF files into a uniform JSON format, and stores them in a MongoDB database; selected attributes are simultaneously stored in a MySQL relational database for fast searching. The microservices architecture enables the replication of individual services across multiple servers, thereby distributing the load, with each server identified by its own value of the SERVER_NUMBER variable. The system is designed to be resilient to failures of both external and internal components: it includes a monitoring service that sends email alerts regarding data flow failures, and administrative services for exporting failed decisions. The entire solution is fully containerised using Docker and Docker Compose, with separate Compose files for the databases, RabbitMQ, data flow, email monitoring service, stream and administrative operations, so that deployment does not require manual installation of dependencies.

Link to project objectives: The deliverable fulfils specific objective 1 of the project and ensures the continuous updating of the project's database. The system is also a key component of the architecture described in deliverable D15.

8.5 D05 – Method for the extraction of references

Data	Value
Deliverable number	D05
Work package / task	KPB2, task KPB2.1

Data	Value
Responsible partner	Pavol Jozef Šafárik University in Košice
Type	Method
Access and dissemination	public — publication in open repositories
Document with a detailed description	ALTAML-D05-KPB2.1-Method-for-extracting-references (EN)
Availability	https://gitlab.science.upjs.sk/altaml/referex

Summary: The method, implemented as the ALTAML Referex project, enables the automatic extraction of references to Slovak legislation, previous decisions of Slovak courts, decisions of the European Court of Human Rights and the Court of Justice of the European Union, and decisions of the Slovak Data Protection Authority. It combines rule-based approaches using regular expressions, Levenshtein distance and fuzzy search with named entity recognition (NER) based on the SlovakBERT model. The solution can be deployed as a REST API service with two endpoints: /v1/referex/laws for extracting references to legislation and /v1/referex/dockets for extracting references to court decisions. The underlying data layer consists of a MongoDB database with six collections (law_validities, law_aliases, law_aliases_lemmatized, codebook_courts, zakony_aliases, dict_lemmatized_laws), which enable the identification of a law even when only an abbreviation is used in the text, the assignment of the correct temporal version of the legislation, and the reliable identification of the court mentioned in the decision. The service uses the Tvaroslovník API for the lemmatisation of the input text. As a result of this method, it enables the creation of hypertext and semantic links between legal documents.

Link to project objectives: The deliverable fulfils Specific Objective 2 of the project and supports the long-term objective of building an interconnected and searchable ecosystem of Slovak and European legal texts. The method for distinguishing between references was also published in the form of a peer-reviewed scientific article at the JURIX 2025 conference (Chapter 3).

8.6 D06 – Method for verdict extraction

Data	Value
Deliverable number	D06
Work package / task	KPB2, task KPB2.2
Responsible partner	Pavol Jozef Šafárik University in Košice
Type	Method

Data	Value
Access and dissemination	public — publication in open repositories
Document with a detailed description	ALTAML-D06-KPB2.2-Method-for-extracting-judgements (EN)
Availability	https://gitlab.science.upjs.sk/altaml/verdict-extraction

Summary: The deliverable is a hybrid data extractor for criminal law judgements, which combines rule-based text processing, heuristic natural language processing methods and the large language model LLaMA 3.3, deployed locally via the Ollama platform. The system extracts the statement of facts, information on acquittal, sentences imposed, mitigating circumstances under Section 36 of the Criminal Code, aggravating circumstances under Section 37 of the Criminal Code, and evidence, which it categorises into witness statements, documentary evidence, expert reports, physical evidence, records and other evidence. The statement of facts is extracted as faithfully as possible directly from the text of the decision, without summarisation or rephrasing. Prior to the extraction itself, the system uses a language model to verify the number of defendants and automatically excludes decisions involving multiple offenders, thereby reducing the risk of incorrect attribution of data to individual persons. The identification of criminal offences takes place in several stages: heuristic selection of candidates based on lexical overlap between the statement of facts and descriptions of criminal offences (with tokenisation, stop-word removal and specialised heuristics for common factual scenarios) and subsequent re-ranking by a language model, which is strictly limited by a prompt to a closed database of Slovak criminal offences. If the model fails, a rule-based fallback mechanism is activated. The ijson streaming parser is used to process large corpora, with results saved on an ongoing basis. The recommended configuration is at least 32 GB of RAM with GPU support (CUDA).

Link to project objectives: The deliverable fulfils specific objective 2 of the project and also directly addresses the end-users' requirement for concise summaries of court decisions, as identified in deliverable D01. The extracted attributes form the input for deliverable D21.

8.7 D07 – Method for extraction of process-related phrases

Data	Value
Deliverable number	D07
Work package / task	KPB2, task KPB2.3
Responsible partner	Pavol Jozef Šafárik University in Košice
Type	Method

Data	Value
Access and dissemination	public — publication in open repositories
Document with a detailed description	ALTAML-D07-KPB2.3-Method-for-extracting-process-related-phrases (EN)
Availability	https://gitlab.science.upjs.sk/altaml/process-related-phrases

Summary: The method presents a hybrid extractor of procedural legal events aimed at the automatic identification of procedural acts and outcomes of proceedings from Slovak court decisions, specifically the dismissal of an appeal, a successful appeal, the discontinuation of proceedings, the annulment of a decision and the referral of a case back for further proceedings. Extraction is limited to a predefined ontology of procedural categories, in which each category contains a set of representative formulations serving as reference prototypes. Processing takes place in multiple layers: following text normalisation and division into sentences, a heuristic filtering of meaningful sentences is carried out based on the presence of procedural keywords, sentence length and the occurrence of legal terminology typical of the operative part; in the first phase of extraction, regular expressions are applied to precisely identify standardised formulations; if the rule-based approach fails to yield a result, the system utilises the locally deployed LLaMA 3.3 language model via Ollama, guided by a strict Slovak prompt. The extracted phrases are then converted into a vector representation using the SlovakBERT model, fine-tuned for the task of semantic similarity between sentences, and classified according to the highest cosine similarity with reference prototypes, applying a minimum threshold. Classification is followed by deduplication based on a combination of category and the normalised form of the phrase, and sorting by score. The deliverable is a JSON structure which, for each event, lists the category, the extracted phrase, the similarity score, the reference prototype, the original decision sentence and the extraction method.

Link to project objectives: This deliverable fulfils Specific Objective 2 of the project. The combination of a rule-based and a semantic approach makes it possible to process even the linguistically variable formulations typical of Slovak court decisions and to create structured databases of procedural events.

8.8 D09 – Method for assigning key-phrases

Data	Value
Deliverable number	D09
Work package / task	KPB2, task KPB2.5
Responsible partner	Pavol Jozef Šafárik University in Košice
Type	Method

Data	Value
Access and dissemination	public — publication in open repositories
Document with a detailed description	ALTAML-D09-KPB2.5-Method-for-assigning-key-terms (EN)
Availability	https://gitlab.science.upjs.sk/altaml/keyphrase-extractor

Summary: The method presents a hybrid extractor of legal key terms from Slovak court decisions, combining statistical text processing, semantic vector representations and a large language model. Extraction is restricted to a closed set of legal terms derived from the D08 deliverable, thereby ensuring terminological accuracy and minimising the risk of generating non-standard expressions. The scoring of candidate terms comprises four components: TF-IDF weighting calculated exclusively over the set of candidate terms (the rest of the document's lexical content is ignored and Slovak stop-words are explicitly removed), cosine similarity between the term's embedding and the embedding of the entire document, generated by the SlovakBERT model fine-tuned for semantic similarity of sentences; a modification of the score based on the term's occurrence in reference texts (cited decisions and referenced legislation), normalised by the number of reference texts; and finally, a linear combination of statistical and semantic components with an adjustable weighting parameter. The embedding of a candidate concept combines the concept's name itself with its definitions from the D08 database. Optionally, the principle of maximum marginal relevance (MMR) may be applied, which ensures semantic diversity among the selected concepts. The final selection is carried out by a locally deployed LLaMA 3.3 model via Ollama, which removes synonymous and duplicate terms and gives priority to legally established terms, thereby simulating a lawyer's expert decision-making.

Link to project objectives: This deliverable fulfils Specific Objective 2 of the project and responds to end-users' requirement for the extraction of keywords from decisions, as identified in deliverable D01. The extracted key terms also serve as input for deliverables D16, D18 and D19. The approach combining key phrases and references in document retrieval was published in a research paper at the JURISIN 2025 workshop, where it achieved a mean average precision (mAP) of 92.85 %.

8.9 D10 – Semi-automated methods for segmenting legal texts

Data	Value
Deliverable number	D10
Work package / task	KPB2, task KPB2.6
Responsible partner	Pavol Jozef Šafárik University in Košice
Type	Method

Data	Value
Access and dissemination	public — publication in open repositories
Document with a detailed description	ALTAML-D10-KPB2.6-Semi-automated-method-for-the-segmentation-of-legal-texts (EN)
Availability	https://gitlab.science.upjs.sk/altaml/legal-text-segmentation

Summary: The deliverable is a hybrid segmentation and classification model that automatically divides Slovak court decisions into logical content segments and assigns a legally relevant segment type to each segment. The system operates on multiple levels. First, it performs coarse segmentation at the level of thematic blocks by analysing semantic distances between neighbouring sentences, whose vector representations are generated by the SlovakBERT model, fine-tuned for semantic similarity between sentences; candidate boundaries are determined based on local maxima of the thematic change score and supplemented by macro-boundaries identified using structural markers typical of Slovak judgements. The segments are subsequently classified into ten legal categories — Subject_of_the_Dispute, Ruling_Part, Court's_Assessment, Factual_Circumstances, Legal_Framework, Evidence, Previous_Proceedings_in_the_Case, Plaintiff_Prosecutor, Defendant_Accused and Others. The classification combines character-based TF-IDF n-grams of the char_wb type, which robustly capture morphological variants of the Slovak legal language, semantic scoring against prototypical class descriptions, and an extensive set of rule-based heuristics taking into account legal formulations, procedural terms, references to regulations, the identification of parties to the proceedings, the wording of judgements, judicial reasoning and the structural position of the segment. This is followed by sequential post-processing to remove inconsistencies between neighbouring segments and a second level of classification — sentence relabelling utilising embedding similarity, the local context of neighbouring sentences and sequential rules. The deliverable consists of a structured JSON representation of the document and analytical CSV files containing relabelling statistics and transition matrices between classes.

Link to project objectives: The deliverable fulfils specific objective 2 of the project. Segmentation makes it possible to narrow down the scope of relevant data extraction and to filter decisions according to the location of a reference or term within a specific part of the decision, which is directly utilised by deliverables D16, D17, D18 and D19.

8.10 D16 – Query model for filtering legal text

Data	Value
Deliverable number	D16
Work package / task	KPB3, task KPB3.3

Data	Value
Responsible partner	Pavol Jozef Šafárik University in Košice
Type	Model
Access and dissemination	public — publication in open repositories
Document with a detailed description	ALTAML-D16-KPB3.3-Query-model-for-searching-legal-texts (EN)
Availability	documentation of the query model; Elasticsearch indices based on the D15 architecture

Summary: The deliverable Defines a query model based on the Elasticsearch Query DSL, which enables the combination of full-text searches within the texts of judgements and legislation, searches based on extracted references to legislation and other judgements, searches based on automatically identified key terms, filtering by categorised attributes, and semantic search using vector representations, including hybrid queries. The model uses separate indexes for court decisions and for legislation. For full-text fields, a custom analyser called 'slovak_lemmatizer' is defined, based on the Slovak Hunspell dictionary, which enables matches to be found even between different grammatical forms of the same word. The court decision index contains seventeen attributes with assigned data types — text for full-text, 'keyword' for precise filtering and aggregations (ECLI, case number, type of decision, area and sub-area of law, key terms, name and type of court), 'date' for range queries, and 'nested' for nested reference structures. As decisions can be lengthy, the vector representation is not created for the entire document, but for smaller text segments (chunks), each of which carries information about the part of the decision from which it originates (header, introduction, operative part, reasoning, instructions); this allows semantic search to be restricted, for example, to the reasoning alone. The embeddings are 384-dimensional, generated by the paraphrase-multilingual-MiniLM-L12-v2 model, and are compared using cosine similarity via k-nearest-neighbour (kNN) queries with approximate nearest-neighbour search. The document contains eight annotated query examples, including a hybrid search that combines BM25 full-text scores with cosine similarity of embeddings using a weighted combination (0.7 to 0.3). In the case of legal regulations, the text is not divided into chunks; instead, a single vector representation of the entire regulation is used, together with filtering by period of validity.

Link to project objectives: This deliverable fulfils specific objective 3 of the project and directly addresses the most pressing requirement of end-users — fast and accurate searching — as identified in deliverable D01. It builds on the architecture of D15 and forms the basis for the user layer of D17.

8.11 D17 – Middleware search method

Data	Value
Deliverable number	D17
Work package / task	KPB3, task KPB3.4
Responsible partner	Pavol Jozef Šafárik University in Košice
Type	Method
Access and dissemination	public — publication in open repositories
Document with a detailed description	ALTAML-D17-KPB3.4-Search-Method-Middleware (EN)
Availability	https://gitlab.science.upjs.sk/altaml/middleware-search-method

Summary: The method addresses situations where a user knows only the general topic they wish to search for, but is unable to immediately formulate all the related terms and legal provisions leading to relevant results. The input is therefore a broadly formulated topic, such as 'cybercrime'. The method first converts this into a 384-dimensional embedding using the paraphrase-multilingual-MiniLM-L12-v2 model and, using KNN queries, searches for semantically related chunks of court decisions and legal provisions in two separate Elasticsearch indices. The decisions are stored in the index in chunks with a maximum size of 400 characters, generated from the main sections of the document (introduction, ruling, reasoning, instructions). Legal provisions are stored hierarchically at the level of section, paragraph and sub-section, with more specific documents always retaining their superordinate context to ensure they remain semantically complete; during a search, hierarchical deduplication is then performed, which, in the event of both a section and its corresponding paragraph appearing, retains the more specific provision. Two text contexts are created from the search results, which are fed into two separate calls to the language model, controlled by system and user prompts and constrained by a JSON schema. The first call suggests 5 to 10 Slovak search terms derived exclusively from of the retrieved decisions (for example, 'IP address', 'computer network', 'website'), whilst the second selects identifiers of relevant legal provisions (for example, Section 247a or Section 360b of the Criminal Code) exclusively from the list of candidates provided. The result is a single JSON object containing the fields `search_phrases` and `legal_provisions`, from which the user selects values for further searches. The method has been implemented as a working prototype, including the demo script `demo_cybercrime.py`.

Link to project objectives: The deliverable fulfils specific objective 3 of the project. It lowers the barrier to entry when searching for legal information for users who do not have detailed knowledge of the terminology and legal provisions in the given field, which was one of the recurring findings of survey D01.

8.12 D18 – Method for relevance evaluation

Data	Value
Deliverable number	D18
Work package / task	KPB4, task KPB4.1
Responsible partner	Pavol Jozef Šafárik University in Košice
Type	Method
Access and dissemination	public — publication in open repositories
Document with a detailed description	ALTAML-D18-KPB4.1-Method-for-assessing-relevance (EN)
Availability	https://gitlab.science.upjs.sk/altaml/document-similarity

Summary: The method determines the degree of similarity between two Slovak court decisions by combining five components: global semantic similarity of the documents, similarity of individual text segments, matches in references to related decisions, matches in cited legislation, and similarity of extracted key terms. Each judgement is first divided into chunks at the paragraph level, aggregated into units of approximately equal size, which allows the similarity to be analysed not only at the level of the entire document, but also between specific parts of the legal argumentation. A key step is the filtering out of procedural and 'template' text — court designations, procedural instructions, standardised formulations and recurring legal phrases — which reduces the influence of general procedural language on the resulting scores and places greater emphasis on the substantive content of the decisions. Vector representations of both documents and chunks are generated by the SlovakBERT model, which is fine-tuned for the semantic similarity of sentences. The similarity of chunks is evaluated symmetrically in both directions, and only the most relevant matches above a defined threshold are included in the result. The similarity of references and cited regulations is calculated using Jaccard similarity of sets of identifiers, whilst the similarity of key terms is calculated using weighted overlap, taking into account their relative significance. The final score is a linear combination of the aforementioned components, with the greatest weight given to the ' ' similarity of chunks, which best represents substantive and argumentative similarity. The deliverable is not merely a numerical score, but also interpretable information — partial scores for individual metrics, a list of shared legal references and regulations, shared key terms, and the most similar text segments.

Link to project objectives: The deliverable fulfils specific objective 4 of the project (identification of relationships between legal texts). The interpretability of the results is a prerequisite for the method's acceptability in legal practice, as confirmed by the feedback from legal research teams recorded in deliverable D23.

8.13 D19 – Algorithm for legal text linking

Data	Value
Deliverable number	D19
Work package / task	KPB4, task KPB4.2
Responsible partner	Pavol Jozef Šafárik University in Košice
Type	Algorithm
Access and dissemination	public — publication in open repositories
Document with a detailed description	ALTAML-D19-KPB4.2-Algorithm-for-linking-legal-texts (EN)
Availability	https://gitlab.science.upjs.sk/altaml/document-relevance-extractor

Summary: The algorithm builds on the similarity metric developed in deliverable D18 and applies it to the entire corpus: upon submission of a source document, it compares the reference decision with all other decisions in the dataset, calculates a similarity score for each pair, and returns a ranked list of identifiers for the most similar decisions. As with D18, a hybrid approach is employed, combining semantic similarity of the text, similarity of the document's content sections, matching of legal references, matching of cited legislation, and similarity of extracted key terms; prior to the calculation, documents are segmented into chunks and procedural and form-based text is filtered out. Vector representations are generated by the SlovakBERT model, fine-tuned for the semantic similarity of sentences; comparisons are made using cosine similarity, whilst the match between references and regulations is assessed using Jaccard similarity. The result is a network of interconnected legal documents reflecting the factual and semantic connections between decisions, which makes it possible to identify, for example, a 'family of decisions' on the same matter or decisions with similar legal reasoning or factual basis. The system is implemented in Python using the Sentence Transformers, scikit-learn and NumPy libraries and supports GPU acceleration (CUDA) for the efficient processing of large corpora.

Link to project objectives: The deliverable fulfils specific objective 4 of the project. The ranked list of relevant documents serves as a recommendation tool that helps lawyers identify decisions worthy of more in-depth manual examination.

8.14 D20 – Method for creating and describing clusters

Data	Value
Deliverable number	D20

Data	Value
Work package / task	KPB4, task KPB4.3
Responsible partner	Pavol Jozef Šafárik University in Košice
Type	Method
Access and dissemination	public — publication in open repositories
Document with a detailed description	ALTAML-D20-KPB4.3-Method-for-creating-and-describing-clusters (EN)
Availability	https://github.com/simonHorvat/GCDC

Summary: The method presents a graph-oriented categoriser of court decisions implemented in Python using the NetworkX library. Decisions, legal references and key terms form the nodes of a multipartite graph, whose edges capture the multidimensional nature of legal texts. Data preparation involves the extraction of legal references using regular expressions and similarity metrics, including the Levenshtein distance, the preparation of a set of candidate key terms, and the creation of vector representations of the texts, their references and the candidate terms (TF-IDF and the bge-m3 model, respectively). The score for a candidate term in a given document is derived from a combination of TF-IDF weighting and cosine similarity, whereby the score from the text itself is combined with the score from its references, and a threshold is then applied. The weight of the edge between a decision and a legal reference is calculated as the ratio of the local number of citations of that reference in a specific decision to the global number of citations of the same reference across the entire dataset, with all weights normalised at the decision level so that their sum equals one. Significant but rare citations thus receive a higher weight than those that occur frequently. Latent legal domains are identified through community detection in two phases: by moving nodes between communities to maximise the increase in modularity, and subsequently clustering communities into supernodes through recursive repetition of the process. The method was experimentally validated on a thousand criminal law court decisions, where it led to the identification of fourteen distinct communities with a modularity value of 0.53; the documentation also provides specific examples of clusters with their key legal references and keywords.

Link to project objectives: The deliverable fulfils specific objective 4 of the project (identification of groups of legal texts). The description of clusters through representative legal references and keywords enables faster filtering and supports the analysis of patterns in judicial practice.

8.15 D21 – Model for predicting the verdicts

Data	Value
Deliverable number	D21
Work package / task	KPB4, task KPB4.4
Responsible partner	Pavol Jozef Šafárik University in Košice
Type	Model
Access and dissemination	public — publication in open repositories
Document with a detailed description	ALTAML-D21-KPB4.4-Model-for-predicting-verdicts (EN)
Availability	https://gitlab.science.upjs.sk/altaml/verdict-prediction

Summary: The model presents a hybrid analytical pipeline for the automated classification of criminal proceedings' outcomes and the subsequent extraction of imposed sentences. The architecture is three-tiered. Tier 0 identifies cases that have been discontinued or suspended on the basis of explicit section references and excludes them from further processing. Tier 1 applies deterministic legal rules based on the occurrence of statutory provisions indicating acquittal, conviction and dominant provisions leading to conviction; if the rules lead to an unambiguous outcome, the decision is classified without further processing. The Tier 2 layer is activated only in the event of an ambiguous result and utilises a Retrieval-Augmented Generation approach, in which an embedding is calculated for the analysed case using the BGE-M3 model, incorporating the factual description, identified offences, mitigating and aggravating circumstances, and evidence; cosine similarity is used to find the most similar historical cases, which are then inserted as few-shot examples into the prompt of a locally deployed LLaMA 3.3 model, which makes the final decision. In the case of convictions, the system subsequently extracts the type of sentence, its severity, conditions, and category of enforcement, along with supplementary details. The consistency of deliverables is ensured by strictly defined Slovak prompts, limiting the deliverable to JSON structures, a closed set of defined sentence types, and zero-temperature inference. Embeddings are cached in a local JSON database and results are saved on an ongoing basis, making the pipeline resilient to computation interruptions.

Link to project objectives: The deliverable fulfils specific objective 4 of the project and also directly meets end-users' requirement for a tool to predict the outcome of legal disputes, which was one of the three key findings of the D01 survey. The model builds on the attributes extracted in the D06 deliverable.

8.16 D22 – Algorithm for finding outliers

Data	Value
Deliverable number	D22
Work package / task	KPB4, task KPB4.5
Responsible partner	Pavol Jozef Šafárik University in Košice
Type	Algorithm
Access and dissemination	public — publication in open repositories
Document with a detailed description	ALTAML-D22-KPB4.5-Algorithm-for-finding-outliers (EN)
Availability	https://gitlab.science.upjs.sk/altaml/method-for-outlier-detection

Summary: The algorithm identifies atypical legal decisions using the Isolation Forest unsupervised learning method and interprets the results using the SHAP method. The procedure consists of three steps. Data preparation transforms legal attributes into a numerical representation — it fills in missing values and creates indicators for them, normalises text attributes, converts binary categories into numerical form, assigns the value 'unknown' to missing categories, applies one-hot encoding to categorical attributes, and removes attributes with a constant value. The Isolation Forest model, which operates on the principle of random division of the data space, subsequently calculates an anomaly score for each decision, an atypicality indicator based on an internal threshold determined by the contamination parameter set at 10 %, and a ranking of decisions according to the degree of deviation. An interpretative layer based on SHAP makes it possible to explain why a particular decision was flagged as anomalous: waterfall plots show the contribution of individual attributes to the resulting score at the level of a single decision; the plot of the average absolute SHAP value and the beeswarm visualisation provide a global view of the importance and direction of the attributes' influence across the entire dataset; and the SHAP interaction heatmap captures the strength of the interaction between pairs of attributes. The method was experimentally validated on five datasets of court decisions manually annotated by lawyers in areas such as contractual penalties and unfair competition, thereby demonstrating its robustness across different legal domains. For each dataset, a separate deliverable folder is generated containing ten types of deliverable (CSV files with scores and rankings, SHAP values, global attribute importance, interaction pairs, summary files and visualisations), along with a common summary file across all datasets.

Link to project objectives: The deliverable fulfils specific objective 4 of the project (identification of legal texts that differ from others). The method for detecting anomalies in legal texts was also published as a peer-reviewed scientific paper at the JURIX 2025 conference (Chapter 3) and, according to feedback from legal research teams recorded in Deliverable D23, ranks among the project's most useful analytical tools alongside descriptive statistical analysis.