

D2 – Method of processing legal texts

Project funded by the European Union's Next Generation EU through the Slovak Republic's Recovery and Resilience Plan under project no. 09I05-03-V02-00049.

Introduction

The project “**Automated Legal Text Analysis based on Machine Learning**” (hereinafter “ALTAML”) is a key initiative aimed at integrating innovative approaches in the field of data processing and subsequent analysis, with a focus on the field of law, which also includes information and communication technology law. The aim of this project is to develop and validate effective methods for the automated analysis of legal texts using machine learning techniques. This involves, in particular, the development of tools that facilitate the processing and analysis of large volumes of data in the form of various legal documents, as well as the extraction of relevant information (attributes) from such documents, including the identification of key terms, references to legislation and other attributes.

The aim of the ALTAML project is therefore to contribute to more efficient access to legal information and to the acceleration of legal processes, thereby ensuring a higher degree of legal certainty and improving the availability of legal texts, the results of their analysis and relevant legal information for both professionals and the general public.

As part of work package KPB1, deliverable **D2 – Method of processing legal texts** - determining the methods of legal text processing for any entity interested in processing legal texts.

Contents

Introduction	1
1 The processing of court decisions	3
2 Bulk downloaded decisions	5
2.1 Decisions from the source justice.gov.sk	5
2.2 Decisions from the source nsud.sk	6
2.3 Decisions from the nssud.sk source	6
2.4 Decisions from the source ustavnysud.sk	7
3 Manually downloaded decisions	9
3.1 Collection of decisions and opinions of the Supreme Court of the Slovak Republic	9
3.2 Collection of decisions and opinions of the Supreme Administrative Court of the Slovak Republic	9
3.3 Collection of decisions of the Constitutional Court of the Slovak Republic	9
3.4 XSLT transformation	9
4 Tool for exporting metadata and creating data packages	11
5 Final evaluation	12

1 The processing of court decisions

The processing of court decisions within the ALTAML project consists of four layers:

- (a) Data acquisition (crawl/API/manual),
- (b) Transformation (PDF→XML, XSLT, relinking),
- (c) Quality and enrichment (validations, metadata mapping, relationship inference),
- (d) Publication and export (CMS, data packages, HTML outputs).

Each layer has defined inputs, outputs, checks, and SLA responses to errors. Key principles: uniform XML structure, reproducibility (scripted pipeline), traceability (logs and audit), and interoperability (ECLI, metadata standards).

As part of the ALTAML project, Wolters Kluwer (hereinafter "WK") obtained decisions using two methods: automated downloading, mostly using program robots, and manual downloading.

During the process of **bulk acquisition of court decisions**, in addition to the PDF documents themselves, detailed information about individual decisions was also obtained, which served as metadata and supplemented the information in the decisions.

The PDF documents were first converted en masse to XML format using a script developed by WK, which used poppler-utils and XSL transformations for technically precise data conversion.

In a subsequent step, the application developed by WK added links to relevant legal regulations, court decisions, and European Union legislation, thus ensuring the completeness and reference value of these documents for further use.

The documents in XML format were then imported into the Content Management System (CMS), where formal validation processes were carried out to ensure the accuracy and integrity of the data according to established standards and requirements.

The details of the individual decisions obtained were then used to systematically generate metadata related to these decisions. This process involved mapping and comparing the data obtained against existing categories and a terminology dictionary, thus ensuring the consistency and accuracy of the information.

In addition, further metadata was created in a subsequent process from the already created XML files of decisions, enabling a detailed overview of each decision through comprehensive and systematic information storage.

In the **manual process of obtaining decisions**, individual decision documents or complete files classified as decision notebooks were downloaded, using PDF format to preserve the integrity and protection of the content. This procedure requires careful checking and verification of each downloaded document to ensure its accuracy and completeness in the context of the decision-making documentation.

As part of this process, PDF documents were converted to XML format using an internal script created by WK. Subsequently, using a specific application also developed by WK, hypertext links to related legal regulations, court decisions, and European Union legislation were implemented in the text. For documents that were available in image PDF format, an OCR tool, specifically Abbyy FineReader, was used to digitize and convert the content into text form.

The transformed decisions in XML format were imported into a content management system (CMS), where the necessary validation processes were carried out to ensure their technical and content accuracy.

The metadata of manually downloaded court decisions were individually recorded in the content management system (CMS) and obtained from the PDF document or from the website of the original source.

2 Bulk downloaded decisions

The schedule of downloads is set to regular batches (cron). Each download checks for changed or new GUIDs/IDs and deduplicates them according to the hash file and file number.

Validations after import into the CMS include: (1) XML schema, (2) presence of mandatory metadata (type, date, file reference, ECLI if available), (3) consistency of links to legislation and case law, (4) anti-duplication rules.

2.1 Decisions from the source [justice.gov.sk](https://www.justice.gov.sk)

These are decisions of general courts published on the official website of the Ministry of Justice of the Slovak Republic.

URL:

- <https://www.justice.gov.sk/sudy-a-rozhodnutia/sudy/rozhodnutia> (search form)
- <https://obcan.justice.sk/pilot/api/ress-isu-service/swagger-ui/index.html#/rozhodnutie> (API documentation)

Detailed information about the decision, including its metadata, is stored in JSON format and is available for download via the API endpoint available at <https://obcan.justice.sk/pilot/api/ressis-service/v1/rozhodnutie/{guid}>.

Then, the text files of court decisions in PDF format are downloaded via the URLs specified in the details["document"]["url"] field.

Example of a URL with decision details (JSON):

<https://obcan.justice.sk/pilot/api/ress-isu-service/v1/rozhodnutie/6e12997b-519c-4799-99fd-a3dbcf8db7cb:47305804-6213-45b8-ae44-0d009d23e52a>

Example URL with PDF text:

<https://obcan.justice.sk/content/public/item/47305804-6213-45b8-ae44-0d009d23e52a>

Example of decision details from the source:



2.2 Decisions from the source nsud.sk

This is a source of decisions of the Supreme Court of the Slovak Republic published on the official website of this court.

URL: <https://www.nsud.sk/rozhodnutia/>

Metadata is obtained from an HTML page containing detailed information about a specific legal decision, as well as a hyperlink to an attached PDF file containing the decision itself.

Metadata extracted directly from HTML includes the title of the decision, the date of its publication, and the case reference number. At the same time, certain metadata are extracted via a created XML document, which includes the specific categorization of the decision, the file identification number, the European case law identifier, and relevant links to existing legal acts and other case law texts.

2.3 Decisions from the nssud.sk source

On the website of the Supreme Administrative Court of the Slovak Republic, decisions are divided into five main categories:

- **decision-making activities in administrative matters:**

<https://www.nssud.sk/rozhodovacia-cinnost-vo-veciach-spravnych/>,

https://www.nssud.sk/wp-admin/admin-ajax.php?action=get_wdtable&table_id=70 (JSON data)

- **decisions in disciplinary matters**

<https://www.nssud.sk/rozhodovacia-cinnost-vo-veciach-disciplinarnych/>,

https://www.nssud.sk/wp-admin/admin-ajax.php?action=get_wdtable&table_id=67 (JSON data)

- **decision-making activities in electoral matters**

<https://www.nssud.sk/rozhodovacia-cinnost-vo-veciach-volebnych/>,

https://www.nssud.sk/wp-admin/admin-ajax.php?action=get_wdtable&table_id=61 (JSON data)

- **decisions of the competence senate**

<https://www.nssud.sk/rozhodnutia-kompetencneho-senatu/>,

https://www.nssud.sk/wp-admin/admin-ajax.php?action=get_wdtable&table_id=64 (JSON data)

- **decisions of the Grand Senate**

<https://www.nssud.sk/rozhodovacia-cinnost-velkeho-senatu/>

As part of the process of extracting metadata from electronic documents, the following information is extracted from the JSON format: document title, file number, register, and subject area. On the other hand, metadata including document type, file identification number, European Case Law Identifier (ECLI), and date of decision are extracted from the XML format.

2.4 Decisions from the source [ustavnysud.sk](https://www.ustavnysud.sk)

These are decisions of the Constitutional Court published on the official website of this court.

URL: https://www.ustavnysud.sk/rozhodovacia_cinnost/rozhodnutia

During the initial stage, metadata is extracted from the downloaded document objects. In the next step, metadata is extracted from the texts of decisions processed in XML format. Specifically, the metadata extracted from document objects includes the type of decision, the date of issue of the decision, the European Case Law Identifier (ECLI), and the file reference number. However, it is preferable that the file reference be verified using XML format. Metadata is obtained from XML-formatted texts, including the file reference, information on the relationship of the decision to the relevant legislation and existing case law.

Example of a document object from the source:

```
{
  "documents": [
    {
      "docType": "USSR_DECISION",
      "content": null,
      "title": "Rozhodnutie - Uznesenie o spojení vecí",
      "index": 0,
      "documentId": "132d0d4b-9c0e-4dec-9319-876a02128d23",
      "mkDocumentType": "Rozhodnutie - Uznesenie o spojení vecí",
      "mkDateOfDecision": "04/09/2025 00:00:00",
      "mkRSAPNumberOfFile": "Pls. ÚS 25/2025",
      "mkTypeOfDecision": [
        "Uznesenie o spojení"
      ],
      "mkYearOfLawReports": null,
      "contentType": null,
      "mkDecisionInTermsOf": [
        "§ 62 zákona o ústavnom súde v spojení s § 166 ods. 1 CSP"
      ],
      "mkLawReportsNumber": null,
      "mkParentIdDecision": null,
      "extension": null,
      "size": null,
      "mkVolumeOfLawReports": null,
      "mkAffectedLegalRegulation": null,
      "mkClauseTitle": null,
      "mkClauseText": null,
      "mkTimePeriodZnaU": null,
      "mkWebTitle": null,
      "mkECLI": "ECLI:SK:USSR:2025:Pls.ÚS.25.2025.1",
      "mkDateOfLegalForce": "04/09/2025 00:00:00",
      "mkPublicationDate": null,
      "mkTypeOfProposer": null,
      "mkTypeOfProceeding": "procesné konanie ÚS SR",
      "mkFormOfDecision": "Uznesenie",
      "mkTypeOfNegotiation": [
        "Prerokovanie vo veci"
      ],
      "mkResultOfNegotiation": [
        "spája na spoločné konanie",
        "nadviaha právoplatnosť a vykonateľnosť"
      ],
      "mkCause": null,
      "mkWordRegister": [
        "spojenie vecí"
      ],
      "mkMaterialRegister": [
        "konanie pred ústavným súdom-použitie ustanovení Civilného sporového poriadku"
      ],
      "mkComplainedLegalRegulation": null,
      "mkJudgeReporter": "uboš Szigeti",
      "mkDifferentView": null,
      "mkUnderage": null,
      "mkRVFNumberOfFile": "928/2025",
      "mkFileReference": null,
      "mkReferences": null,
      "mkIncludeToZnaU": false,
      "mkClarificationOfLegalRegulation": null,
      "mkEntryDate": null,
      "mkFormOfEntry": null,
      "mkTypeOfEntry": null,
      "mkDecisionInTermsOfForSort": "§ 62 zákona o ústavnom súde v spojení s § 166 ods. 1 CSP"
    }
  ]
}
```

3 Manually downloaded decisions

The process of acquiring this type of content is a more complex task, characterized by the predominant need for manual intervention in performing most of the related activities, which are not fully automated.

Manual workflow: (1) identification of the notebook/decision, (2) PDF download, (3) OCR in ABBYY FineReader for image PDFs, (4) PDF→XML transformation, (5) metadata addition and quality control, (6) import into CMS. In the case of notebooks, individual decisions are segmented based on headings, numbering, and patterns.

As part of our initiatives, we present a comprehensive overview of URL sources focused on these specific types of content:

3.1 Collection of decisions and opinions of the Supreme Court of the Slovak Republic

This document brings together selected decisions issued by the Supreme Court of the Slovak Republic, which have been edited and organized. The publication is designed in the form of a notebook, which means that it contains several decisions in one file.

URL: <https://www.nsud.sk/zbierka-stanovisk-a-rozhodnuti/>

3.2 Collection of decisions and opinions of the Supreme Administrative Court of the Slovak Republic

This is a collection of selected decisions of the Supreme Administrative Court of the Slovak Republic, which have been edited and organized. The publication is designed in the form of a notebook, which means that it contains several decisions in one file.

URL: <https://www.nssud.sk/zbierka-stanovisk-a-rozhodnuti-najvyššieho-správneho-sudu-slovenskej-republiky/>

3.3 Collection of decisions of the Constitutional Court of the Slovak Republic

These are selected decisions of the Constitutional Court of the Slovak Republic, which have been edited and organized. They are published either individually or once a year in a single PDF document.

URL: https://www.ustavnysud.sk/sk/rozhodovacia_cinnost/zbierka_rozhodnuti

3.4 XSLT transformation

For the purpose of implementing the ALTAML project, Wolters Kluwer has developed a script that implements the XSLT transformation process for converting XML documents to HTML format.

The HTML format is optimized for viewing in web browsers and ensures greater comprehensibility for other users.

XSLT templates define the header and structure of decisions. Styles are designed to be readable in a browser and consistent with the internal design manual.

4 Tool for exporting metadata and creating data packages

As part of the project implementation, Wolters Kluwer developed a software application that allows metadata to be exported from the results of SQL queries submitted by users, with the application also serving to generate data packages.

Users can formulate a set of questions using SQL queries that relate to individual topics or domains of the ALTAML project. The result sets generated on the basis of these queries adhere to the following metadata structure: Id; aspi_id; author; spzn; type; date; source; pub_number; full_source; area; relation; rejsrik; year; porc; edition.

This metadata structure allows the creation of reports that are fully compliant with the requirements of the project clients and that lead to specific overviews of data packages, including the data packages themselves containing html documents.

The output metadata (Id; aspi_id; author; spzn; type; date; source; pub_number; full_source; area; relation; rejsrik; year; porc; vydani) enable the composition of overviews and integration into downstream analytics. HTML packages with linked documents are created.

5 Final evaluation

The processing of court decisions is a complex process that combines automated technologies, structured data models, and expert domain knowledge. As part of the **ALTAML** project, we have created a unified, methodologically sound, and technically consistent procedure that ensures high quality, reliability, and repeatability of the entire workflow—from the initial acquisition of documents to the creation of data packages ready for further analytical and application use.

The key elements of success are:

- consistent application of validation mechanisms,
- standardization of formats and metadata,
- transparent linking to legislative and case law sources,
- risk management, and continuous process improvement.

The procedures put in place give the ALTAML project the potential to provide high-quality, reliable, and technically compatible judicial data. This document also serves as a comprehensive methodological framework for further developing, optimizing, and adapting the process to new requirements.