Report on the Network Analysis

Grant Agreement No.:	101087342
Project Acronym:	POLINE
Project Title:	Principles of Law in National and European VAT
Website:	https://site.unibo.it/poline/en
Contributing WP:	WP3
Deliverable ID:	D3.3
Contractual delivery date:	30/09/2025
Actual delivery data:	30/09/2025
Dissemination level:	PU
Deliverable leader:	**UNITO**

This project is funded by the European Union's Justice Programme (2022)

1. Document History

Version	Date	Author	Partner	Description
1.0	08/08/2025	Rachele Mignone	UNITO	Definition of the document structure
2.0	22/08/2025	Davide Audrito, Ivan Spada	UNITO	First draft
3.0	28/08/2025	Davide Audrito, Rachele Mignone	UNITO	First round of revisions
4.0	15/09/2025	Rachele Mignone, Giovanni Siragusa, Ivan Spada	UNITO	Final draft
5.0	26/09/2025	Luigi Di Caro	UNITO	Final revision

2. Contributors

Partner	Name	Role	Contribution
UNITO	Luigi Di Caro	WP Coordinator	Coordination of the Deliverable activities
UNITO	Davide Audrito	WP Participant	Contribution to the Deliverable activities
UNITO	Rachele Mignone	WP Participant	Contribution to the Deliverable activities
UNITO	Giovanni Siragusa	WP Participant	Contribution to the Deliverable activities
UNITO	Ivan Spada	WP Participant	Contribution to the Deliverable activities

3. Table of Contents

4. List of Figures and Tables

5. List of Acronyms

CJEU	Court of Justice of the European Union
ECLI	European Case Law Identifier
JIF(s)	Judicial Interpretative Formula(s)
LLM(s)	Large Language Model(s)
VAT	Value Added Tax

6. Executive Summary

This report details a comprehensive, LLM-driven methodology to map and quantify the doctrinal relationships within European VAT case law. Building on the prior extraction and classification of Judicial Interpretative Formulas (JIFs) from a multilingual corpus, the project first uses Gemini 2.5 Pro to standardize and normalize all legal citations. This structured data is then used to construct a formal legal citation network of 6,034 nodes and 15,556 edges. The network analysis, using metrics like degree centrality, provides a powerful, quantifiable map of judicial influence and authority, moving beyond traditional qualitative methods to trace the complex evolution of legal doctrine in the VAT domain.

This deliverable is structured as follows. Section 7 is the Introduction, which provides the context for the POLINE project, introduces the concept of JIFs, and outlines the contributions of the report. Section 8, Methodology, details the technical steps involved in the process, covering the LLM-driven standardization of legal citations using Gemini 2.5 Pro and the formal construction of the directed graph using the networkx library. This is followed by Section 9, Network Analysis, which presents the core findings of the report, focusing on the application of network theory metrics, such as degree centrality, to quantify the influence, impact, and authoritativeness of specific legal texts and judgments. Finally, Section 10, Conclusions, summarizes the key achievements of the deliverable, reiterating the significance of the quantitative, relational knowledge graph for future analysis of European VAT law.

7. Introduction

The core challenge addressed by this work stems from the unique reasoning style of the Court of Justice of the European Union (CJEU), which frequently relies on the verbatim reproduction or subtle rephrasing of specific interpretative statements from previous judgments using a "copy-pasting" or "LEGO" technique. This means the precedential weight often lies not in an entire judgment, but in these recurring passages, which we term Judicial Interpretative Formulas (JIFs). Understanding the relationship between these JIFs is crucial for analyzing evolving trends in judicial reasoning across both the EU and national legal systems.

Building on prior project phases, this report details a comprehensive methodology to map, analyze, and quantify the doctrinal proximity of legal statements. The initial phase of the POLINE project involved automatically extracting Judicial Interpretative Formulas (JIFs) from a multilingual corpus of Value Added Tax (VAT) judgments spanning the CJEU and national courts in Italy, Sweden, and Bulgaria. Following extraction, the knowledge was organized: a multilingual VAT ontology was developed, and Large Language Models (LLMs) were then used to classify each JIF, transforming the raw collection into a structured and conceptually organized repository of legal knowledge.

This structured dataset now serves as the foundation for the network analysis presented in this report. This contribution focuses on the construction and analysis of a legal citation network. To achieve this, all legal citations within the JIFs were first standardized and normalized using Gemini 2.5 Pro[1] for structured entity recognition and assignment of official unique identifiers (like CELEX and ECLI). This structured output was then used to construct a formal, directed graph with 6034 nodes and 15556 edges. The report concludes by presenting the results of the network analysis, demonstrating the use of metrics like degree centrality to quantify the influence, impact, and authoritativeness of specific legal texts and judgments within the European VAT framework.

8. Dataset Creation

Previous phases of the POLINE project successfully produced a comprehensive corpus of 3836 JIFs, creating a substantial and heterogeneous foundation for comparative legal analysis. This dataset offers a diverse multi-jurisdictional perspective by incorporating

[1] Google. (2025). Gemini 2.5 Pro [Large language model]. https://deepmind.google/models/gemini/pro

1402 JIFs from the CJEU, 1511 from Italian courts, 478 from Swedish courts, and 445 from Bulgarian courts.

This extensive dataset was generated by automatically extracting the JIFs from a multilingual collection of VAT-related court decisions that had been manually compiled by legal experts. To accomplish this, we employed LLMs, first conducting a rigorous selection process to identify the most effective tool for the task. By evaluating models like Claude 3.7 Sonnet[2][3], DeepSeek-R1[4][5], and Gemini 1.5 Pro[6][7] against a manually annotated standard (Table 1), we determined that Claude 3.7 Sonnet using a few-shot prompt delivered the highest performance. The extraction task was framed as a paragraph-level binary classification, where the model was prompted not only to identify a JIF but also to retrieve its associated legal citations, returning the findings in a machine-readable JSON format. The method's validity was confirmed by legal experts, achieving an average F1-score of 0.932 on the CJEU test set and a precision of approximately 0.987 on Italian Supreme Court judgments.

		Precision	Recall	F1-score
Zero-Shot	Gemini 1.5 pro	0.806	0.815	0.810
	Deepseek R1	0.864	0.961	**0.910**
	Claude 3.7 Sonnet	0.887	0.922	0.904
Few-Shot	Gemini 1.5 pro	0.844	0.877	0.86
	Deepseek R1	0.891	0.817	0.852
	Claude 3.7 Sonnet	0.901	0.972	**0.935**
Chain-of-Thoughts	Gemini 1.5 pro	0.829	0.76	0.793
	Deepseek R1	0.895	0.955	0.924
	Claude 3.7 Sonnet	0.874	0.972	**0.920**
Few-Shot Chain-of-Thoughts	Gemini 1.5 pro	0.865	0.707	0.778
	Deepseek R1	0.887	0.972	0.928
	Claude 3.7 Sonnet	0.884	0.983	**0.931**

Table 1. Model selection scores.

[2] Anthropic. (2025). *Claude 3.7 Sonnet* [Large language model]. https://www.anthropic.com/
[3] model snapshot: 2025-02-19. Temperature: 0.5, top-p: 0.7, top-k: 35
[4] DeepSeek-AI. (2025). *DeepSeek-R1* [Large language model]. https://chat.deepseek.com/
[5] model snapshot: 2025-05-28. Temperature: 0.35, top-p: 0.7, top-k: 35
[6] Google. (2024). *Gemini 1.5 Pro* [Large language model]. https://gemini.google.com/
[7] model snapshot: gemini-1.5-pro-002. Temperature: 0.20, top-p: 0.7, top-k: 35

With thousands of JIFs extracted, the next challenge was to organize this raw data into a conceptually meaningful structure for deeper analysis. We addressed this by developing a multilingual ontology for the VAT domain. Its framework was built upon the EUR-LEX[8] Directory of Case-Law taxonomy and then manually enriched with additional terms and hierarchical relationships by tax law experts. The final taxonomy contains 127 labels (nodes) and 130 edges. A crucial step was creating a multilingual dictionary to translate these labels into Italian, Swedish, and Bulgarian. Finally, we utilized LLMs once again to classify each extracted JIF against this ontology, transforming the flat list of formulas into a structured and contextualized repository of legal knowledge. During the classification process, the model was also tasked with the extraction of all legal references within each JIF, resulting in a list of Python dictionaries having shape [{'citation': raw citation text, 'src_doc': citing document identifier, 'src_par': citing paragraph number}, ...]. This deliverable details how, starting from his data format, it was possible to create and analyse a citation network of the VAT domain.

9. Citation Formatting with LLMs

The standardization of legal citations from the extracted data was performed using Gemini 2.5 Pro. This process was a crucial step in preparing the data for the Network Analysis, specifically for integration into the POLINE platform and the creation of a robust citation network.
The LLM was instructed to perform a structured, entity-recognition and normalization task on raw citation data. The prompt was designed to guide the model to not only classify the legal source but also to normalize key identifiers and create a standardized output format suitable for database entry and subsequent network analysis.

The input to the LLM was a list of Python dictionaries, each containing the original citation text and its context: [{'citation': raw citation text, 'src_doc': citing document identifier, 'src_par': citing paragraph number}, ...]

The prompt provided to Gemini 2.5 Pro was:

> """You will receive an input list of legal citations with shape[{'citation': the citation as written in the source document,

[8] eur-lex.europa.eu

'src_doc': the document citing the citation,
'src_par': the paragraph citing the citation}...]

For each citation in the list, output a JSON object with the following fields:

- citation: the citation as in the input,
- src_doc: as in the input,
- src_par: as in the input,
- citation_type: treaty/eu_legislation/case_law/national_legislation/other classifies the citation,
- case_number: number present in the citation (if present). If the citation type is case-law the case_number has shape C-number/year,
- name: name of the cited document,
- id: unique identifier for the cited docuement. It should be CELEX, ECLI or other official forms of iddentification. If in doubt, look online,
- id_type: string describing the type of id used above ('celex'/'ecli'/...),
- article: if present, the article cited,
- paragraph. if present, the paragraph cited

If a citation references several articles from the same document, create different entries.
for example "D.P.R. n. 633 del 1972, artt. 23, 24 e 25" will result in 3 different entries.

Input list:[{'citation': 'D.P.R. n. 633 del 1972, art. 9', 'src_doc': 'Sez 5 Ordinanza n 27389 del 01_12_2020 (Rv 659695 01)', 'src_par': '4'},...]
JSON list output: “””

This instruction explicitly requires the model to:

1. Classify the citation type (citation_type).
2. Extract contextual information (src_doc, src_par).
3. Normalize specific fields like the case number and, most critically, find the official unique identifier (id) and its type (id_type) by leveraging its knowledge base (simulated by the phrase "If in doubt, look online"). This step is essential for creating reliable nodes in the citation graph and reducing the risk of “hallucinations”.

4. Parse specific articles and paragraphs, and implement a splitting logic to create separate entries for citations that reference multiple articles within the same document.

A qualitative analysis was carried out by legal experts on a sample of 150 extracted citations. This evaluation confirmed that the LLM achieved an adequate level of performance in terms of accuracy and adherence to the required data structure. The high-quality output generated by Gemini 2.5 Pro ensures the reliability of the dataset for the intended purposes: integration into the POLINE platform and the subsequent creation of the citation network. The standardization of the unique identifiers (CELEX, ECLI, *etc.*) guarantees that the nodes in the network are consistently defined, which is fundamental for accurate network analysis.

The output below illustrates the model's transformation of a raw Italian legislative citation into the structured JSON format:

Row citation: {'citation': 'D.P.R. n. 633 del 1972, art. 9', 'src_doc': 'Sez 5 Ordinanza n 27389 del 01_12_2020 (Rv 659695 01)', 'src_par': '4'}

Model output:


```
{
  "citation": "D.P.R. n. 633 del 1972, art. 9",
  "src_doc": "Sez 5 Ordinanza n 27389 del 01_12_2020 (Rv 659695 01)",
  "src_par": "4",
  "citation_type": "national_legislation",
  "case_number": null,
  "name": "Decreto del Presidente della Repubblica 26 ottobre 1972, n. 633",
  "id": "urn:nir:stato:decreto.presidente.repubblica:1972-10-26;633",
  "id_type": "normattiva",
  "article": "9",
  "paragraph": null
 }
```


This process resulted in a set of **5427** parsed citation records.

10. Network Creation

The creation of the legal citation network is a fundamental step for subsequent analysis. This process transforms the structured citation data, parsed as described in Section 9, into a formal graph structure. The graph is constructed using the networkx[9] Python library, which provides a robust framework for creating, manipulating, and studying complex networks.

The core of the network creation process involves establishing a dictionary of nodes and one of corresponding edges. This approach ensures that each legal entity (such as a judgment or a paragraph) is represented by a unique node, preventing duplication. Furthermore, it allows for the assignment of specific attributes to both nodes (e.g., document type) and edges (e.g., relation type), which is crucial for detailed analytical queries.

The procedure for populating the network graph from the parsed citation records is executed for each record as follows.

10.1 Creating Citing Nodes

For each parsed citation record, the process begins by identifying and creating nodes for the citing document and the specific paragraph containing the citation, known as the JIF.

To initiate the graph construction, the **citing document node (src_doc)**, which represents the document containing the citing JIF, is generated first. The attributes for these nodes are determined as follows:

- **ID:** This identifier is assigned based on the document's original jurisdiction:
  - For judgments from the CJEU, the CELEX number serves as the node ID.
  - For national judgments that exist within the record set as cited documents, the ID output by the LLM during the prior processing step is utilized. Should disambiguation be required, where a single document is associated with multiple candidate IDs, the ID exhibiting the highest frequency of occurrence is chosen.
  - In any other case, the document name doubles as its ID.
- **Doc_id:** in the case of src_doc nodes, this attribute equals the ID, but is nonetheless kept for structural consistency with other document nodes.

[9] NetworkX Developers. (2025). **NetworkX** (Version 3.5) [Computer software]. Available from https://networkx.org/

- **Type:** Uniformly set to "case_law", reflecting the source dataset's exclusive composition of legal judgments.
- **'Contains_jif':** Invariably set to True. This is a functional constraint, as all citations are derived directly from the JIFs embedded within these source documents.

The next step involves creating a node for the **citing paragraph node (src_par)**, which represents the specific JIF from which the citation is extracted. This process encompasses the following steps:

1. **ID generation:** A unique ID is generated for the paragraph, composed of the parent document's ID and the source paragraph index.
2. **jif_id generation:** A unique JIF identifier (jif_id) is created to link the node back to the initial dataset, structured as: origin_code#src_doc#src_par. This compound ID is necessary because the src_doc value does not consistently correspond to the src_id.
3. **Node initialization and dictionary update:** The node is added to the node dictionary with the attributes
    - 'id': as previously computed
    - 'doc_id': ID of the src_doc (parent node)
    - 'type': 'jif'
    - 'jif_id': jif_id

    If the node is already present (which can happen if the paragraph was previously included as a cited paragraph from a different JIF), the process ensures that its attributes are updated, specifically setting its 'type' to 'jif' and assigning the calculated jif_id. This guarantees that any existing node is correctly recognized as a JIF source.

At this point, a directed edge with attribute name= 'divided_into' can be added, linking the document node to the paragraph node.

The final step involves searching for corresponding JIFs based on textual similarity (JIFs with distinct IDs but identical content). For CJEU JIFs, this check is rigorously performed across all available linguistic versions. This process yields a list of IDs for every JIF that shares the exact text of the citing paragraph but possesses a different identifier. Each of these newly identified JIFs is added to the node dictionary, and two dedicated "corresponds_to" edges are established (one per direction), linking each corresponding JIF pair. This final measure ensures that all textual duplicates are accurately represented and linked within the network.

10.2 Creating Cited Nodes

The process of creating **Cited Document Nodes** involves an initial check to determine the presence of JIFs, followed by the creation of nodes for the document, article, and paragraph levels of granularity.

A preliminary check is executed to determine if the cited document is present in the **initial project dataset**. This verification establishes whether the document contains JIFs, which is used to set the node's 'contains_jif' attribute. The cited document node is then initialized with the following core attributes:

- **ID:** The unique document identifier (cit_id) extracted during the citation parsing phase.
- **Doc_ID:** Set to be identical to the **ID**.
- **Type:** The **citation type** (e.g., 'case_law', 'legislation') as classified during the initial extraction.
- **'Contains_jif':** A **Boolean** value indicating the presence of JIFs in the document.

If the citation specifies a finer level of granularity, corresponding article and paragraph nodes are created sequentially.

If the cited document is divided into articles and the citation specifies one, a separate **article node** is created. These nodes possess the following attributes:

- **ID:** Composed of the document ID and the article number.
- **Doc_Id:** The ID of the parent document.
- **Type:** Set to 'article'.

A directed edge labeled 'divided_in' is added, connecting the document node to the newly created article node.

If the cited document or article is further divided into paragraphs and the citation specifies one, a paragraph node is created. If the paragraph node is identified as a JIF, its 'is_jif' attribute is set to True.

A directed 'divided_in' edge is added. This edge originates from the article node if the paragraph is part of an article, or from the document node otherwise.

Finally, the citing paragraph is connected to the cited text of the finest granularity (document, article, or paragraph) using two inverse directed edges: 'cites' (Source → Target) and 'is_cited'

11. Network Structure

The network resulting from the creation presents a schema as depicted in Figure 1. For legibility purposes, all document types have been represented by the 'Document' node.

Node types:

- **treaty:** document type assigned by the LLM during the citation parsing phase
- **eu_legislation:** document type assigned by the LLM during the citation parsing phase
- **national_legislation:** document type assigned by the LLM during the citation parsing phase
- **case_law:** document type assigned by the LLM during the citation parsing phase
- **other:** document type assigned by the LLM during the citation parsing phase
- **article:** manually assigned according to the LLM's extraction
- **paragraph:** manually assigned according to the LLM's extraction
- **JIF:** manually assigned

Node attributes:

As described in Section 10, each node has a set of attributes according to its type. Document nodes (i.e., treaties, legislation, case law…) have the following attributes:

- ID: ID identifying the node
- Doc_id: ID identifying the document
- type: one value among 'treaty', 'eu_legislation', 'national_legislation', 'case_law', and 'other' as determined during the parsing process
- contains_jif: defaulted to False, can be True only if the document has 'type'= 'case_law' as only judgments can contain JIFs.

Article and paragraph nodes, on the other hand, have the following:

- ID: ID identifying the specific article/ paragraph
- Doc_id: ID identifying the document
- type: 'article'/ 'paragraph'.

Finally, nodes representing JIFs are described by:

- ID: ID identifying the specific paragraph
- Doc_id: ID identifying the document
- type: 'jif'

- Jif_if: allowing the mapping to the original JIF dataset.

Edges:

- **divided_in:** links a node representing a document or an article to one of its subdivisions.
- **corresponds_to:** links a JIF with a corresponding one.
- **cites:** links a JIF to a document or a section if cites.
- **is_cited:** links a document or a section to the JIF that cites it.

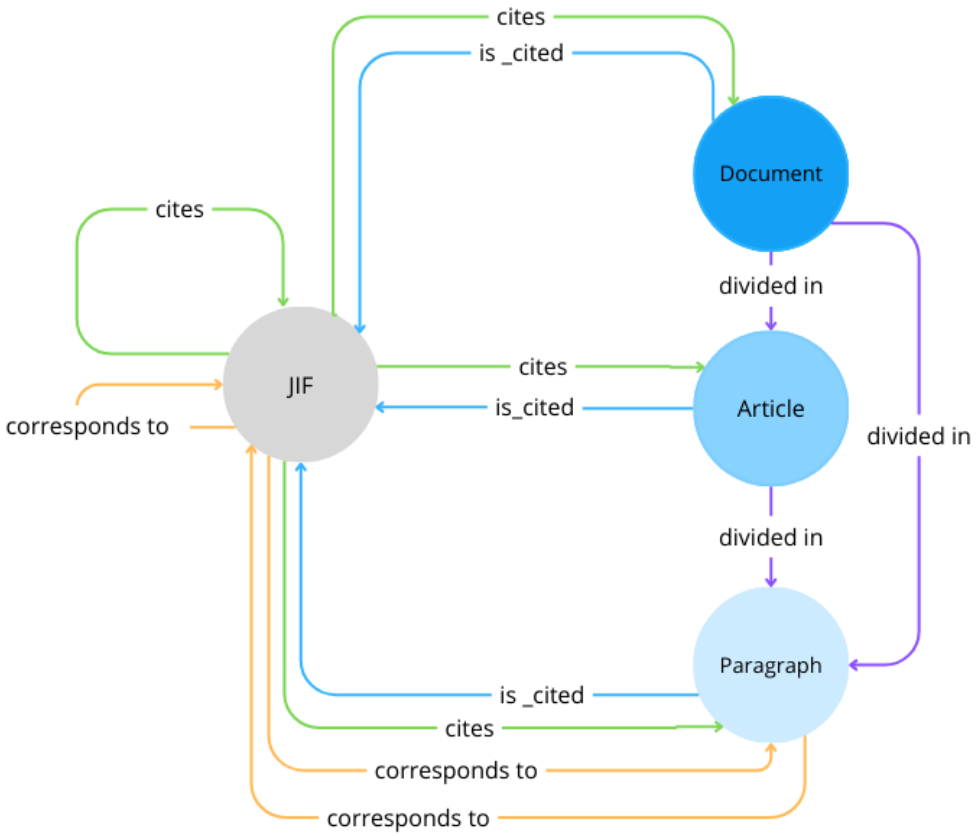

Figure 1. Citation network schema.

12. Network Analysis

Network analysis of legal citations provides a powerful method to map and understand the relations between legal documents. This data-driven approach moves beyond traditional analysis, allowing for the objective identification of influential judgments, the tracking of evolving legal precedents, and the discovery of non-obvious relationships within a large body of law. The primary objective is to transform legal text into a structured network to measure the influence and connectivity of legal arguments, unlocking new avenues for legal research and practice. This methodology offers several advantages for legal professionals and researchers. By analyzing network centrality metrics, researchers can quickly identify the most influential and authoritative documents within a specific legal field, streamlining legal research. The structure of citation networks can also be compared across different jurisdictions to understand variations in legal reasoning and judicial dialogue. Furthermore, patterns in these networks could potentially be used to forecast the evolution of legal doctrines or identify areas of law that are becoming more contested or significant.

Based on this network, we have developed specific applications to demonstrate its utility. By calculating the in-degree of each JIF, we can rank legal judgments based on their citation frequency, thereby identifying the most pivotal and authoritative cases within the dataset and providing a quantitative measure of a case's importance. This same approach can be applied to legislation. Analyzing citation patterns allows us to determine the most influential pieces of legislation at both the European and national levels, highlighting foundational legal texts in the VAT domain.

The citation network constructed for this analysis is composed of **6,034 nodes** and **15,556 edges**.

The network is a directed graph designed to capture the hierarchical nature of legal documents and their citation relationships.
The components are defined by several node types representing various granularity levels of legal documents. The Document is the highest-level entity, such as a complete court judgment or legislative act. Within a document, an Article represents a specific section, and the Paragraph is the most granular level. JIFs represent a particular kind of paragraph carrying interpretative value.

To measure the influence and connectivity of each node, we primarily use degree centrality. In-degree represents the number of incoming citations a node receives and is a direct measure of a node's influence, impact, and authoritativeness within the dataset.

Conversely, out-degree represents the number of outgoing citations a node makes, indicating the breadth of legal precedent and sources the document references.

The methodology allows for analysis at the document, article, and paragraph levels, providing a more nuanced understanding of what part of the document holds the most meaning.

Example Analysis: Document 62005CJ0434 - Stichting Regionaal Opleidingen Centrum Noord-Kennemerland/West-Friesland (Horizon College) contro Staatssecretaris van Financiën.) - Case C-434/05

The citation analysis shows that this judgment holds a strong position within the dataset, evidenced by 41 inbound citations and 43 references to prior works. This demonstrates significant activity in its overall citation profile.

By disaggregating, we can see the citation activity for specific parts of the document. For instance, it is possible to see that the document itself was cited 8 times, while its most cited paragraphs, namely 16 and 20, were cited 4 times each. Besides knowing the number of citations, the graph structure allows researchers to easily follow the nodes' edges to discover which documents cite the paragraphs under scrutiny.

For JIFs, the number of citing documents can be considered a metric of relevance, underlining the paragraph's interpretative nature as it serves as precedent to numerous future documents.

This citation-based approach is also instrumental for identifying the most salient legislation within a domain, such as the Value Added Tax (VAT) framework. By quantifying the outbound citations directed from the dataset's documents to legislative texts, we can empirically determine the core statutes.

The analysis of outbound citations confirms the concentration of legal influence within a few key legislative documents in the VAT framework. The most frequently cited documents, detailed below, are consistently referenced, with citation counts representing their established precedential authority:

- 907 citations - **Council Directive 2006/112/EC** of 28 November 2006 on the common system of value added tax. The Directive's influence is granular, with numerous specific articles drawing significant reference:
    - Article 90(1): 76 citations

    - Article 132(1)(b): 52 citations
    - Article 132(1)(m): 45 citations
    - Article 73: 43 citations
    - …
- 712 citations - **Sixth Council Directive 77/388/EEC** of 17 May 1977 on the harmonization of the laws of the Member States relating to turnover taxes - Common system of value added tax: uniform basis of assessment. This document also demonstrates high-granularity citation, with specific paragraphs serving as frequent references:
    - Article 13: 67 citations
    - Article 13A(1)(b), In-degree: 45
    - …
- 635 citations - **Decreto del Presidente della Repubblica 26 ottobre 1972, n. 633 -** the foundational Italian legislation that established and governs the VAT system

Our analysis of the VAT subset confirms this pattern: the most frequently cited legislative documents are Council Directive 2006/112/EC and the Decreto del Presidente della Repubblica 26 ottobre 1972, n. 633 (the foundational Italian VAT legislation). Consistent with our hypothesis, the documents receiving the highest number of citations are the pivotal European and national legislative instruments governing the Value Added Tax domain, confirming the centrality of these statutes to the legal field.

13. Conclusions

In conclusion, this deliverable successfully employs an LLM-driven methodology to transform a fragmented, multilingual corpus of VAT case law into a robust, quantifiable, relational knowledge graph. By extracting and classifying thousands of JIFs and meticulously standardizing their underlying citations, we have constructed a comprehensive legal citation network comprising 6,034 nodes and 15,556 edges. The subsequent network analysis, leveraging metrics such as degree centrality, provides an unprecedented empirical foundation for understanding the true flow of legal authority and identifying the most influential precedents within the European VAT system.

Ultimately, this report moves beyond traditional qualitative analysis, delivering a powerful, quantifiable map of doctrinal proximity and judicial influence that is essential for tracing the complex evolution and convergence of national and European VAT law.