









Automated Extraction and Representation of Citation Network: A CJEU Case-Study

Galileo Sartor¹ , Piera Santin² , Davide Audrito³  , Emilio Sulis¹ ,
and Luigi Di Caro¹ 

¹ Computer Science Department, University of Torino, C. Svizzera 185, Turin, Italy
{galileo.sartor,emilio.sulis,luigi.dicaro}@unito.it

² CIRSFID - AI, University of Bologna, Via Galliera 3, Bologna, Italy
piera.santin2@unibo.it

³ Legal Studies Department, University of Bologna, Via Zamboni 27, Bologna, Italy
davide.audrito2@unibo.it

Abstract. Although with some discrepancy, both in common law and in civil law systems, previous judgments play an important role with respect to future decisions. Traditional legal methodologies usually involve the use of manual rather than automatic keyword search mechanisms to retrace the steps of the judicial decision-making. However, these methods are generally highly time-consuming and can be subject to different types of biases. In this work, we present an automated extraction pipeline to map and structure citations in rulings regarding fiscal state aids in the case-law of the Court of Justice of the European Union. In particular, by exploiting the available XML data in the EUR-Lex platform, we built an end-to-end parser based on a set of regular expressions and heuristics, which is able to iteratively extract all citations, finally creating a hierarchical structure of citations with their contextual information at the paragraph level. Such data structure can be projected into a graphical representation, enabling useful visualization and exploration features and insights, such as the diachronic study of the development of specific citations and legal principles over time. Our work suggests how the exploitation and analysis of citation networks through automated means can provide significant tools to support traditional legal methodologies.

Keywords: Legal knowledge extraction · Visual law · Citation network · Digital justice

1 Introduction

According to the general theory of legal sources, in common law systems previous judgments are binding on future decisions, whereas in civil law systems they only have a persuasive force. Over time this general distinction has been attracting criticism. Some scholars argue, for instance, that there is no considerable discrepancy in the use of precedents in civil and common law systems

or that in both jurisdictions courts are bound to comply with precedents [6]. Moreover, the CJEU style of citation is quite peculiar, since it is inspired by a civil law model but it uses to invoke its own case law as common law courts do [13].

The importance of precedent citations in CJEU case law increased over the past years, and this appears also from the growing number of explicit citations in each judgment (e.g. the older one in our dataset does not have any citation, whereas the most recent one has nine direct citations with several recursive ones). For such reason, mapping precedents in CJEU case-law became crucial in order to better understand the development of the judicial framework.

Traditional legal methodology provides analytical tools to retrace the steps of judicial decision-making through citation networks. However, these methods are highly time-consuming and can be subject to bias, for instance due to non-empiric legal doctrine based exclusively on *opinio iuris*. In 2014, for example, Derlen and Lindholm [8] addressed the citation network of the Court of Justice of the European Union (CJEU) case-law. They found out that *Dassonville* and *Bosman* should be considered the most relevant landmark judgments, taking into consideration the frequency by which subsequent CJEU decisions cite these cases. On the contrary, handbooks use to attribute the greatest pioneering value to *Cassis de Dijon* and *Van Gend en Loos*. This example shows that computational analysis of law can have a significant role in supporting traditional legal research, in that citation networks can provide new insights in the understanding of judicial decision-making.

This work focuses on citation to previous rulings made by the CJEU in appeal decisions in the area of fiscal state aid. Such field is an interesting and significant case study, since it has a reduced legal basis and the discipline has been entirely developed through the CJEU judgments [22]. In particular, we choose to focus on appeals against decisions of the General Court (GC), because the relatively short time span (first judgment is of 1997) allows us to have a more homogeneous style of citations [21].

Furthermore, the Publications Office of the European Union developed an XML schema, which has been consistently applied to CJEU judgments. That allows the automatic extraction of references on the basis of affordable and original data, which guarantee certainty and precision of the results.

The main objective of our research is extracting and mapping the precedents cited in the CJEU case law in the area of fiscal state aid. The extracted data allows to offer a graphical representation that can be used to support legal studies, and may, in the future, enable further analysis and elaborations, for instance by way of NLP and other AI techniques [24]. The selected methodology is based on the annotated corpus of cases made available (in XML when possible) by the CJEU. This initial data is then further analyzed through the use of regular expressions to extract precise references to other CJEU cases. Regular expressions are used relying on heuristic rules, when the case is not yet available in XML form. This extracted data, saved in JSON and graph format, can then be analyzed through network analysis tools such as neo4j.

This experiment represents the first step for further comprehensive linguistic, legal and computer-science analysis.

Following this introduction, in Sect. 2 we briefly analyze the available works in the field of legal citation analysis. In Sect. 3 we outline the methodology proposed in our research. Section 4 then presents the main results achieved through the abovementioned methodology. Finally in Sect. 5 we sum up the initial results of our research, and outline possible future developments.

2 Related Work

Research in legal citation analysis has attracted interest over the last decades and is now becoming a well-established research area. The work of Fowler et al. on the case law citations by the Supreme Court of the United States [10] had a key role in identifying the benefits of citation analysis. The authors applied network analysis to address judicial citations, namely to explore the functioning of *stare decisis* and other issues related to the use of precedents. In the continental legal system similar approaches were followed to analyse the decisions of the Dutch Supreme Court [26], the CJEU [9] and the European Court of Human Rights (ECHR) [16].

Previous studies focused mainly on three directions: extraction of references, ranking of references and references labelling.

The automatic extraction of references has been developed mainly with regard to their different formats. To this end, Adedjouma et al. [1] used gazzettiers and concept markers; Palmirani et al. [17] used regular expressions; Harasta [12] used Conditional Random Files (CRF); Leitner et al. [15] used BiLSTM neural networks. Other authors adopted methods beyond natural language processing, such as ML, deep neural networks and CRF to extract citations [2].

As shown in Langone [14], CRF and neural networks achieved similar performance in the extraction of citations, taking into account the number of references effectively identified in the legal sources at hand. As concerns other methods such as regular expressions, performances are dependent upon specific research objectives, e.g. whether authors aim to extract explicit or implicit references. Although effective, these experiments are not able to extract the entirety of references without gaps, differently from our methodology that affords to extract and map judicial citations contained in well-structured and annotated judgments published on CURIA, the official website of the CJEU.

Some studies are devoted to rank legal sources as a result of citation analysis based essentially on the number of codes' mutual citations [3].

In 2014, Derlén and Lindholm [8] argued that in-degree centrality is misleading when assessing the role of judicial decisions, because the relevance of a rarely cited judgment can increase if citations appear in important future decisions. Some domain-specific methods took into account other features, including the judicial instance [25].

Instead of ranking judgments, Sadl et al. [21] proposed to directly rank cited paragraphs to avoid confusion and inaccuracy, considering the peculiar structure

of the CJEU system of references. For the same reason our extraction system allows to create an accurate, fine-grained, and reliable dataset (see Sect. 4).

In conclusion, multiple works address the task of identifying and labelling citations. To this end, several studies made use of citations' surrounding text [11], while others took into account different features, including signal words and patterns [7].

Network analysis in legislation found application in Sadeghian et al. [19], in which the authors focused on predicates, namely words that highlight the purpose or feature of citations without referring to the subject-matter at hand. They classified edges with 9 labels predicates were extracted through Conditional Random Fields and k-means classification with embeddings afforded to automatically classify the edges.

In 2018, Sadeghian automatically detected the purpose of cross references with a detailed semantic label set [20]. In a different way, Sulis et al. identified implicit citations by building a network of single portions of a text and undertaking binary classification tasks by way of co-occurrence network analysis [23].

While previous efforts in this field have been limited by the lack of well formed XML representations of case law, in our case the data made available by the European Union have a common structure for references in their XML schema. For this reason, the methodology presented in this paper may be easily extended to other legal domains of the European jurisdiction. In addition, our analysis allows a recursive regression that support the identification of pioneering interpretation statements, thus enabling a historical analysis of case law.

3 Methodology

In this research we focused on a subset of judgments, available on the EUR-Lex website, in the domain of fiscal state aid. The methodology described in this section is nonetheless expandable to other legal domains, and is only dependent on how citations are expressed in the original text of the decision.

In order to extract the desired information on the references to precedents contained in a judgment we started by downloading the XML representation available on the EUR-Lex platform¹.

The availability of this structured data is a step forward in the practical use of the information contained in the platform within research projects, thus simplifying, in our case, the initial assessment of the source text. The XML format is well documented [18] and includes different tags for citations. The one that we used is *REF.DOC.ECR*, a complete reference to previous cases. In particular, we only included references to a specific paragraph, assuming that they are the most legally-relevant. This could be easily changed in future extensions of the methodology and related tools.

It is worth mentioning that not all judgments are available in this format, and for those documents that are not in fact available we decided to use regular

¹ <https://eur-lex.europa.eu>.

expressions as a fallback. It is thus possible to extract information from both a well-formed XML representation, as well as the legal text, although parsing the XML file gives us information that has already been marked and verified.

In most cases the references in a CJEU case are of the form:

[...] see, inter alia, judgments of 15 November 2011, Commission and Spain v Government of Gibraltar and United Kingdom, C-106/09 P and C-107/09 P, EU:C:2011:732, paragraph 73 [...]

This enables us to save a list of tuples, containing the cited case (the ECLI), and the cited paragraph number.

For each of the cited paragraphs, the relevant case is downloaded and the relevant portion of text is then extracted. On that section the reference extraction described above is carried out once again.

By evaluating citations in the single paragraph, we are able to collect only citations that may have a relation useful in the legal analysis. This can be seen in the outcome of our evaluation, where, for instance, the branches of the network containing the legal precedents are distinct from those about procedural rules. By following this procedure recursively we end up with a nested set of citations, with each citation having a list element, named *references*.

For the proper extraction of the paragraph number and its content, the parsing of the XML text is enhanced through the use of regular expressions. This functionality relies on the fact that the structure of these documents is generally quite uniform. Regular expressions have previously been adopted in NLP applications to extract structured information from legal sources/cases in plain text [17,20].

The extracted data is then formatted in the standard JSON format, making it easily accessible to both non-technical persons and, again, to automated computer programs.

The code below shows an example of the information that is extracted for each reference, as it is saved in the JSON structure. Note that we store all different identifiers (ECLI, CELEX, and the Case Number) as well as the name of the judgment and text of the specific cited paragraph:

```
"ecli": "ECLI:EU:C:2011:732",
"text": "C-106/09 P and C-107/09 P Commission and Spain v Government of
      Gibraltar and United Kingdom [2011] ECR I-111137273",
"par_num": "NP0073",
"celex": "62009CJ0106",
"case_no": "C-106/09 P",
"xml_url": <url to the xml representation>,
"references": [...],
"par_text": "On the other hand, advantages resulting from a general
      measure applicable without distinction to all economic operators do
      not constitute State aid within the meaning of Article 87 EC (see, to
      that effect, Case C-156/98 Germany v Commission [2000] ECR I-6857,
      paragraph 22, and Joined Cases C-393/04 and C-41/05 Air Liquide
      Industries Belgium [2006] ECR I-5293, paragraph 32 and the case-law
      cited).",
```

On the basis of this information it is possible to use different tools to build visualization systems, citations' maps, and a database allowing further analysis of the citation network.

One such tool is neo4j, in which the above mentioned data can be imported and visualized through graphical or tabular representations of the citations. The system will, at a simple level, build a tree from the items it finds in the "references" sections of the database. Further analysis can be carried out on the basis of this representation, since it enables reasoning on the relations between the different nodes of the graph (cases and cited paragraphs). In the generated graph the cases are shown as yellow nodes, while the paragraphs are purple. There are two types of relations, the citation between two paragraphs (REFERSTO), and the case that a paragraph belongs to (BELONGSTO).

In the following section we shall see one of the ways in which we are able to query this database, and how it may assist in the legal analysis of precedents.

4 Results

By reasoning on the data structure defined in the previous section, the system is able to build a graphical representation of the relationships between citations. This representation can be useful to visualize the chronological development of citations and the evolution of legal principles.

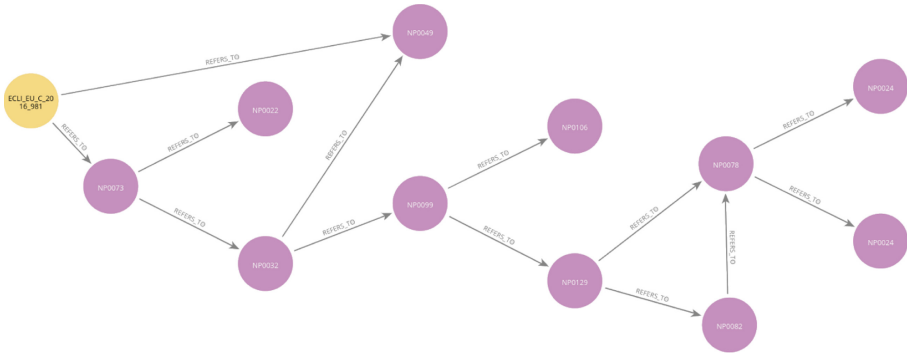


Fig. 1. Single branch of citations starting from a case.

In Fig. 1 we can follow a branch from the original case queried, in yellow, visualizing the sequence of paragraphs cited from previous decisions, in purple. All the cited paragraphs contain information on the original text, and the decision in which they are present. This can be used to analyse the evolution of a legal statement, from its origin in case law to its use as a consolidated principle [5].

This representation can be further enhanced with the ability to easily see the cited text as well as the relations between citations, thus assisting legal experts

to assess the deep impact that the CJEU case law has on the European legal system.

It should be noted that the figure above shows only a single branch of the entire citation network. However, it is possible to expand the visualisation with different queries allowing to extract the needed case-specific information from the database. For instance, it is possible to view the relations between groups of judgments or to only expand the sections we are interested in.

With this information it will be possible, in future developments, to analyze the different citations, with complex algorithms to determine textual similarities and the relative importance of nodes in the network.

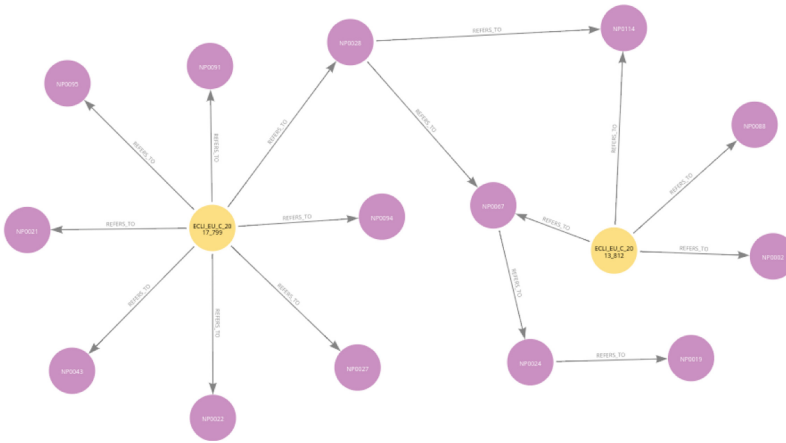


Fig. 2. Relation between multiple cases in the database

Figure 2 shows an excerpt from the database, demonstrating the interconnections between multiple cases, where we can see that multiple cases cite the same paragraph directly (not in recursive citations). With this view it is possible to automatically analyse the relationships with different algorithms (such as PageRank [4]), and obtain those citations that may be more legally relevant.

While at the moment the database is limited to the set of cases on state aid, it can be easily extended to a vast number of judgments, while maintaining the desired functionality. The main hurdle in applying this methodology to a broad set of cases is the consistency of the input data. From a short search on the EUR-Lex platform it is possible to see that cases are not always available in XML format, or where available, they are incomplete. This is not a blocking issue, as we have seen previously, since often the missing information, namely the paragraph number and the identifier, can be found in the text format and can be added after being parsed by the regular expressions.

A greater issue is where the citation is not identified with a unique code, such as ECLI or CELEX. In most cases it may be possible to find the judgment

according to the parties involved in the proceeding, but especially in this legal area there may be several cases involving the same parties, such as the European Commission and a EU member State, e.g. *Italy v. Commission*. Heuristics may be useful to overcome this issue, for example by looking exclusively at earlier cases than the one at hand and supported by the fact that in more recent judgments citations are usually well-formed, at least in the legal text.

Furthermore, the sporadic unavailability of judgments translated in English must not be underestimated. However, it does not occur often, and it is fixed once the case is translated from the original language.

The system has been briefly tested on other legal fields and, at a first glance, it seems that judgments in fiscal state aid have a more consistent representation of the cited precedents, both in the XML format, and in the natural text. A possible reason for that is the absence of other legal instruments, such as directives or regulations. This fact makes the case law citations ever more important in order to determine the legal framework of fiscal state aid.

5 Conclusions and Future Work

The different representations made available by the system (as described in 4) may provide an interesting tool to support the analysis of case law in the European framework. The vertical representation (Fig. 1) can show branches of citations starting from a specific case and could be explored to better understand the context in which a decision was taken and its ability to affect future cases. This representation could also have practical uses for legal professionals by assisting the search and reasoning on legal principles and their evolution.

This initial work opens to a broad range of future research opportunities. The citation network could be enhanced through NLP and ML technologies, including by enriching the semantic comprehension of the text surrounding citations, possibly extracting meaningful representations and analyzing their role in judicial decision-making. The network could be expanded to show the nodes (i.e., paragraphs/principles) that are more frequently cited and to assess the evolution of nodes' citation frequency over time.

The full network can be used to better identify the most important paragraphs and can be adopted in legal research to compare the text between a citing and a cited paragraph. This relation could enhance the identification of new judicially-established legal principles.

On the legal side, the content of this network can be enriched by including more information extracted from legal cases. For instance, one could match citations to the outcome of the judgment, extract relevant keywords to map cases and their citations by the type of procedure, the specific matter, the originating state, or the legal effect.

Furthermore, the results obtained so far can be worked on with automated NLP solutions to extract more information from the cited paragraphs, without having to parse the whole text. This may be useful to determine whether cited paragraphs, originating from different judgments, are similar enough to share the intended meaning.

Another possibility is to link citations to the keywords extracted in all the available languages and possibly sorted through NLP tools and other automated means. This could be useful to determine how keywords are managed in the different European languages, as well as for the creation of a common, multilingual ontology for each relevant sector.

At first instance, this work relies on the information retrieved from the XML schema adopted by the CJEU itself. In so doing, the results obtained do not require any kind of interpretative bias. Moreover, the extraction is quite rapid and does not require any further manual annotation. Accordingly, it is possible to provide a tool that allows anyone to retrieve the citation network of a decision, just by providing its ECLI. Thus, the output will be immediately useful for the user, even in the absence of any further application of NLP technologies.

References

1. Adedjouma, M., Sabetzadeh, M., Briand, L.C.: Automated detection and resolution of legal cross references: approach and a study of luxembourg's legislation. In: 2014 IEEE 22nd International Requirements Engineering Conference (RE), pp. 63–72. IEEE (2014)
2. Bach, N.X., Thuy, N.T.T., Chien, D.B., Duy, T.K., Hien, T.M., Phuong, T.M.: Reference extraction from vietnamese legal documents. In: Proceedings of the Tenth International Symposium on Information and Communication Technology, pp. 486–493 (2019)
3. Boulet, R., Mazzega, P., Bourcier, D.: Network approach to the French system of legal codes part ii: the role of the weights in a network. *Artif. Intell. Law* **26**(1), 23–47 (2018). <https://doi.org/10.1007/s10506-017-9204-y>
4. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.* **30**(1–7), 107–117 (1998)
5. Brown, L.N., Kennedy, T.: *The Court of Justice of the European Communities*. Sweet & Maxwell, London (2000)
6. Civitarese, S.: A European convergence towards a stare decisis model. *Rev. Digit. de Derecho Admin.* **14**, 173 (2015)
7. De Maat, E., Winkels, R., Van Engers, T.: Automated detection of reference. In: *Legal Knowledge and Information Systems: In: JURIX 2006 the 19th Annual Conference*, vol. 152, p. 41. IOS Press (2006)
8. Derlén, M., Lindholm, J.: Goodbye van gend en loos, hello bosman? using network analysis to measure the importance of individual CJEU judgments. *Eur. Law J.* **20**(5), 667–687 (2014)
9. Derlén, M., Lindholm, J.: Is it good law? network analysis and the CJEU's internal market jurisprudence. *J. Int. Econ. Law* **20**(2), 257–277 (2017)
10. Fowler, J.H., Johnson, T.R., Spriggs, J.F., Jeon, S., Wahlbeck, P.J.: Network analysis and the law: measuring the legal importance of precedents at the U.S. supreme court. *Polit. Anal.* **15**(3), 324–346 (2007). <https://doi.org/10.1093/pan/mpm011>
11. Hamdaqa, M., Hamou-Lhadj, A.: An approach based on citation analysis to support effective handling of regulatory compliance. *Future Gener. Comput. Syst.* **27**(4), 395–410 (2011). <https://doi.org/10.1016/j.future.2010.09.007>
12. Harasta, J., Savelka, J.: Toward linking heterogenous references in czech court decisions to content. In: *JURIX*, pp. 177–182 (2017)

13. Koopmans, T.: Stare decisis in European law. *Essays Eur. Law Integr.*, pp. 1957–1982 (1982)
14. Langone, D., Fulloni, A., Wonsever, D.: A citations network for legal decisions. In: *New Frontiers in Artificial Intelligence. Lecture Notes in Computer Science.* Springer International Publishing (2020)
15. Leitner, E., Rehm, G., Moreno-Schneider, J.: Fine-grained named entity recognition in legal documents. In: Acosta, M., Cudré-Mauroux, P., Maleshkova, M., Pellegrini, T., Sack, H., Sure-Vetter, Y. (eds.) *SEMANTiCS 2019. LNCS*, vol. 11702, pp. 272–287. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33220-4_20
16. Olsen, H.P., Küçüksu, A.: Finding hidden patterns in ecthr’s case law: on how citation network analysis can improve our knowledge of ecthr’s article 14 practice. *Int. J. Discrimination Law* **17**(1), 4–22 (2017)
17. Palmirani, M., Brighi, R., Massini, M.: Automated extraction of normative references in legal texts. In: *Proceedings of the 9th International Conference on Artificial Intelligence and Law*, pp. 105–106 (2003)
18. Publication Office, E.U.: <https://op.europa.eu/en/web/eu-vocabularies/formex/>
19. Sadeghian, A., Sundaram, L., Wang, D., Hamilton, W., Branting, K., Pfeifer, C.: Semantic edge labeling over legal citation graphs. In: *Proceedings of the Workshop on Legal Text, Document, and Corpus Analytics (LTDCA-2016)*, pp. 70–75 (2016)
20. Sadeghian, A., Sundaram, L., Wang, D.Z., Hamilton, W.F., Branting, K., Pfeifer, C.: Automatic semantic edge labeling over legal citation graphs. *Artif. Intell. Law* **26**(2), 127–144 (2018). <https://doi.org/10.1007/s10506-018-9217-1>
21. Sadl, U., Tarissan, F.: The relevance of the network approach to European case law. reflexion and evidence. In: *New Legal Approaches to Studying the Court of Justice* (2020). <https://hal.archives-ouvertes.fr/hal-03098351>
22. Schön, W.: Tax legislation and the notion of fiscal aid: a review of 5 years of European jurisprudence. In: Richelle, I., Schön, W., Traversa, E. (eds.) *State Aid Law and Business Taxation. MSTLPF*, vol. 6, pp. 3–26. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-662-53055-9_1
23. Sulis, E., Humphreys, L., Vernerero, F., Amantea, I.A., Audrito, D., Caro, L.D.: Exploiting co-occurrence networks for classification of implicit inter-relationships in legal texts. *Inf. Syst.* **106**, 101821 (2022)
24. Sulis, E., Humphreys, L.B., Audrito, D., Di Caro, L.: Exploiting textual similarity techniques in harmonization of laws. In: Bandini, S., Gasparini, F., Mascardi, V., Palmonari, M., Vizzari, G. (eds.) *AIxIA 2021 - Advances in Artificial Intelligence*, pp. 185–197. Springer International Publishing, Cham (2022)
25. Van Opijnen, M.: Citation analysis and beyond: in search of indicators measuring case law importance. In: *JURIX*, vol. 250, pp. 95–104 (2012)
26. Winkels, R., de Ruyter, J.: Survival of the fittest: network analysis of dutch supreme court cases. In: Palmirani, M., Pagallo, U., Casanovas, P., Sartor, G. (eds.) *AICOL 2011. LNCS (LNAI)*, vol. 7639, pp. 106–115. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35731-2_7