European Law Open (2024), page 1 of 7
doi:10.1017/elo.2024.17

DIALOGUE AND DEBATE: ESSAY

A note of caution on CJEU databases

Michal Ovádek

Department of Political Science, University College London, London, UK
Email: m.ovadek@ucl.ac.uk

(Received 8 November 2023; revised 13 April 2024; accepted 19 April 2024)

Abstract

The purpose of this short research note is to draw attention to two major pitfalls of working with databases of decisions of the Court of Justice of the European Union. The first one is technical in nature and relates to the discrepant coverage of the Curia and Eur-Lex databases. The second one is linguistic in nature and relates to the fact that most scholars using these databases work in English. New work on this front is capable of addressing the first issue but a change to research practices would be required to address the second.

Keywords: European Court of Justice; database; case law; EU law; text-as-data

1. Introduction

Databases of decisions of the Court of Justice of the European Union (CJEU) are at the heart of academic and practitioners' work on EU law. Two official databases feature most prominently in this line of work: Curia and Eur-Lex. Those more familiar with them will be aware of some differences between them, but despite the centrality of these databases in research on the EU, there has so far been no systematic attempt at clarifying the degree to which they overlap. Relatedly, one may regard as particularly concerning the fact that to date neither database contains fully digitised text of landmark rulings handed down before 1990 that are considered to belong to the 'Pantheon' of EU law such as *Costa v ENEL* and *Van Gend en Loos*. This short research note shows concretely the limitations of both databases and how new work in this area contributes to overcoming them.

Nonetheless, in addition to the varying temporal and site-specific coverage of the main CJEU databases, much of EU law scholarship overlooks the linguistic dependence of CJEU decisions. This research note quantifies the perhaps surprising extent to which decisions in French (the working language of the CJEU) outnumber decisions translated into English (the language used by most scholars and practitioners). While we do not expect users to switch to French *en masse*, it may be useful to understand the risks associated with working with CJEU decisions in English.

2. Extant work

The use of Curia and Eur-Lex is so widespread in EU law that it is not necessary to dwell on their importance. Both doctrinal and social science scholars source their information from these databases. Some of the extant research attempted to assemble collections of CJEU decisions, but they all suffer from drawbacks. Databases are either focused on just one of the

sources[1] or they do not address the discrepancies between them.[2] The only database that has so far successfully tackled problems with the official sources does not contain the texts of CJEU decisions,[3] which limits its usefulness for legal scholars or researchers in the burgeoning text-analysis domain.

In the absence of a unified database of CJEU decisions, the vast majority of researchers in this area defaults to either Curia or Eur-Lex (or a combination of the two). A typical piece of research will rely at least in part on a keyword search attempting to identify the relevant line of the case law whereupon the coherence of a new judgement with preceding cases is examined. The alternative to conducting one's own search for relevant cases is to surrender to the citations selected by the Court. While capable of saving time, this latter strategy restricts the scope of independent, critical inquiry. The CJEU's citations offer a biased[4] picture of its own case law and should be approached as a statement of what the Court wants the reader to see rather than what its case law actually said. An analysis of the coherence of a given line of case law needs to start from a mapping of all relevant cases,[5] not all of which have likely been cited by the Court.

3. Comparison of Curia and Eur-Lex

Assuming a researcher is interested in systematically identifying a set of cases satisfying some criteria (such as the presence of keywords) – or even all the decisions produced by the CJEU – the question remains whether they can rely on Curia and Eur-Lex for the job. In order to answer this question, we need to systematically examine the coverage of both databases. This is now possible thanks to our ongoing work on the IUROPA Text Corpus, a new database of the texts of all CJEU decisions that consolidates and improves the content of both Curia and Eur-Lex.[6] In this research note, we use this database merely to benchmark the coverage of Curia and Eur-Lex to underscore the importance of completeness rather than explore all the research possibilities offered by a more complete text corpus, a topic deserving of separate attention. The reason is that even users of Curia and Eur-Lex who will not use our new database should be aware of the pitfalls of working with the official sources.

Both keyword searches and more advanced text analytical techniques require textual input in the plain text format.[7] Thus, even though it is possible to access the PDF of a ruling on Curia or Eur-Lex (although even here the coverage differs), we want to know how many decisions (including Advocate-General (AG) opinions) are available in a fully digitised, plain-text format. Because French is the working language of the Court but English the most used language in academia and legal practice, we look at database coverage in both languages.

[1]JC Fjelstul, 'The Evolution of European Union Law: A New Data Set on the Acquis Communautaire' 20 (2019) European Union Politics 670; M Ovádek, 'Facilitating access to data on European Union laws' 3 (1) (2021) Political Research Exchange, available at https://doi.org/10.1080/2474736X.2020.1870150.

[2]A Dyevre and N Lampach, 'Issue attention on international courts: Evidence from the European Court of Justice' 16 (2021) Review of International Organizations 793; A Dyevre, M Glavina, and M Ovádek, 'The Voices of European Law: Legislators, Judges and Law Professors' 22 (6) (2021) German Law Journal 956; V Mattioli and K McAuliffe, 'A corpus-based study on opinions of advocates general of the Court of Justice of the European Union: Changes in language and style' 6 (1) (2021) International Journal of Legal Discourse 87.

[3]SA Brekke, JC Fjelstul, SSL Hermansen, and D Naurin, 'The CJEU database platform: Decisions and decision-makers' 11 (2) (2023) Journal of Law and Courts 389.

[4]J Frankenreiter, 'The politics of citations at the ECJ: policy preferences of EU member state governments and the citation behavior of judges at the European Court of Justice' 14 (4) (2017) Journal of Empirical Legal Studies 813.

[5]MA Hall and RF Wright, 'Systematic content analysis of judicial opinions' 96 (2008) California Law Review 63.

[6]M Ovádek, JC Fjelstul, D Naurin, and J Lindholm, 'The IUROPA Text Corpus' in J Lindholm, D Naurin, U Sadl, A Wallerman Ghavanini, SA Brekke, JC Fjelstul, SSL Hermansen, O Larsson, A Moberg, M Nässtrom, M Ovádek, T Pavone, and P Schroeder (eds), *The Court of Justice of the European Union (CJEU) Database* (IUROPA 2023).

[7]Plain text can be embedded in HTML code (as is the case on websites) but differs from text in PDFs. Extracting plain text from PDFs is typically done using optical-character recognition (OCR) technologies.

https://doi.org/10.1017/elo.2024.17 Published online by Cambridge University Press

Table 1. Number of decisions with plain text

	English	French
Curia	23,439	35,957
Eur-Lex	32,034	35,060
IUROPA	**36,251**	**46,530**

Table 1 compares the number of decisions available in plain text on Curia, Eur-Lex and in the IUROPA database across English and French.[8] The IUROPA database uniquely identifies decisions[9] and sources their texts from both Curia and Eur-Lex, depending on which offers the higher quality text.[10] In addition, it digitises PDFs where no or only poor plain text exists. EU law experts will be familiar with the subpar and incomplete digitization of older decisions – including the most coveted ones such as *Costa v ENEL* – on Eur-Lex.[11] These documents are only available in capitalised letters (with punctuation issues) and miss the Court's exposition of the facts, law and arguments of the parties.

Even from the overview in Table 1 we can glean the scale of discrepancies between the official databases and the more comprehensive IUROPA corpus. The combined database contains a staggering 10,000 French decisions in plain text more compared to either Curia or Eur-Lex. Moreover, as not all decisions are translated from French to English, there is a similarly large difference between the two language versions. There are many more plain-text documents in English on Eur-Lex than Curia but similar numbers of decisions in French. However, this masks the different coverage of the two databases, as shown in the Venn diagram in Figure 1.

Figure 1 reveals that even though the number of documents in French is similar on both Curia and Eur-Lex, the two databases are to a significant degree non-overlapping. Two factors primarily influence the discrepant coverage. First, Curia only contains plain-text documents from June 1997 onwards. Decisions before this date are only available in PDF files on Curia. Second, many decisions never make it to Eur-Lex. The process by which files are transmitted from the Court (and its database, Curia) to Eur-Lex (maintained by the EU Publications Office, an independent agency) is not at all transparent. But in general terms, Curia is the more comprehensive database of the two official sources for decisions rendered after June 1997.

To save resources, the CJEU does not translate all its decisions into English from French, its working language.[12] As a result, the majority of research conducted in English is liable to miss a non-negligible portion of rulings.[13] The scale of the discrepancy comes to the fore especially when considering all decisions handed down by the CJEU (as captured by the IUROPA Text Corpus)

[8]The figures are current as of 4 November 2023.

[9]The European Case-Law Identifier (ECLI) combined with the date of decision identifies decisions uniquely. On the contrary, the more commonly used case numbers do not achieve this, as a case can be associated with several decisions (for example when there is an interim ruling).

[10]Higher quality means here that the text is longer (if they are different) and the paragraphs are correctly separated and numbered.

[11]Curia offers no plain-text representation of these decisions, further limiting their searchability.

[12]There is a much smaller subset of decisions for which the reverse holds true. Based on the records from the IUROPA Text Corpus, there are roughly 500 General Court decisions which are available in English but not in French. The exact number depends on whether one counts summaries of decisions as well.

[13]This is without considering the additional possible issue of a researcher conducting an analysis in English while a decision is being translated from French but is not yet available publicly.

https://doi.org/10.1017/elo.2024.17 Published online by Cambridge University Press

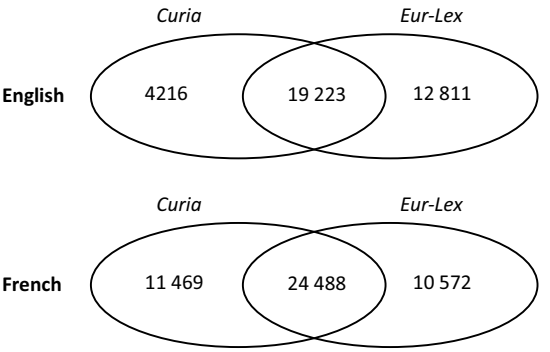

Figure 1. Overlap in plain-text decisions between Curia and Eur-Lex.

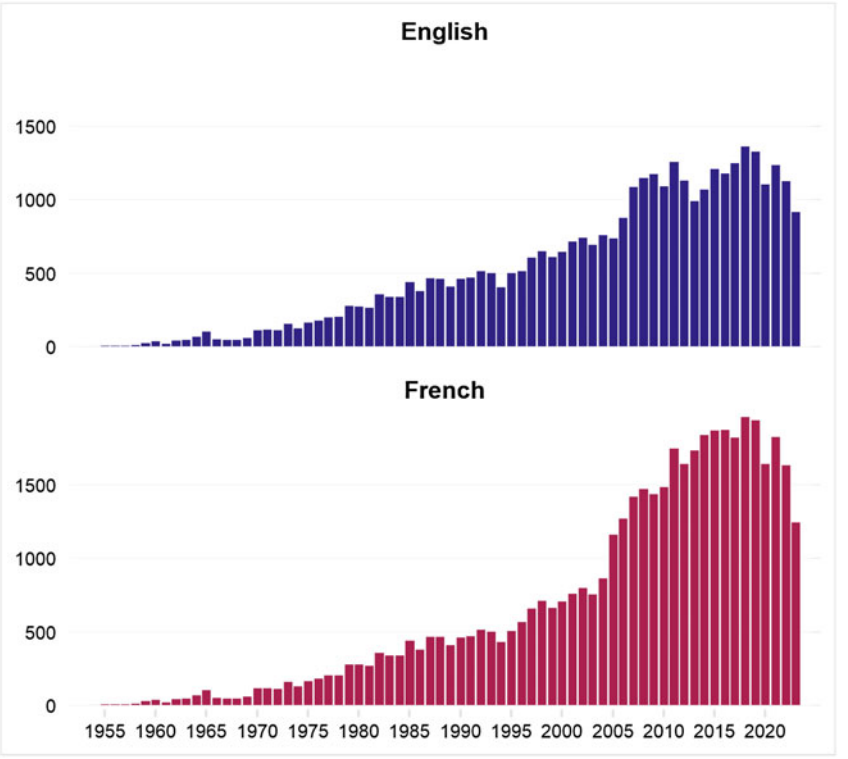

Figure 2. Number of decisions in the IUROPA Text Corpus by language and year.

and is more significant from 2005 onwards (see Figure 2).[14] Although the Court prioritises the translation of what it considers the most important decisions,[15] the exact size of the gap between

[14]The decline in the proportion of cases translated into English after 2004 is likely related to an increase in the backlog of cases around this time. The CJEU was forced to seek efficiency gains in light of rising delays. See more generally JC Fjelstul, M Gabel, and CJ Carrubba, 'The timely administration of justice: using computational simulations to evaluate institutional reforms at the CJEU' 12 (2023) Journal of European Public Policy 2643; TYC Yeung, M Ovádek, and N Lampach, 'Time efficiency as a measure of court performance: evidence from the Court of Justice of the European Union' 53 (2) (2022) European Journal of Law and Economics 209.

[15]This policy is most obvious from the fact that virtually all Grand Chamber rulings are translated, but the same is not true of Chamber decisions.

https://doi.org/10.1017/elo.2024.17 Published online by Cambridge University Press

English and French will also vary based on the area of the law and the deciding court. The General Court is more likely than the Court of Justice to see its rulings go untranslated. Similarly, VAT cases, for example, are more likely to remain in French only compared to, say, citizenship cases. The translation policy thus impacts differently on scholars depending on the focus of their work.[16] The upshot is, however, that researchers should be mindful of the risk of missing relevant cases if they choose to work exclusively in English.

Looking merely at the number of decisions available in plain text obscures the fact that many of these documents are incomplete, in particular before records went digital in 1990. The IUROPA Text Corpus has made significant strides towards fully digitizing decisions adopted between 1954 and 1989, but the work remains in progress.[17] Of the document pages 55 per cent in French have so far been processed and incorporated into the database.[18] In addition to the digitised text concerning some of the most important rulings in EU law history, the amount of judicial text recovered in this way is also significant. The partially digitised decisions from this period on Eur-Lex contain on average around 55 paragraphs. In contrast, the fully digitised documents in the IUROPA corpus average some 120 paragraphs.[19]

4. Implications

Overcoming the limitations of the official databases has practical implications for all types of research that rely on the texts of CJEU decisions. Legal scholars stand to benefit from a database that dutifully retrieves the relevant information from the entire universe of CJEU decisions, rather than whatever undefined selection of them happens to live on either Curia and Eur-Lex. By way of example, scholars working on the reception of international law in the EU legal order might want to trace the history of engagement with customary international law and legitimately ask when the CJEU referred to it for the first time. The case that most often comes up in this regard is *Poulsen*, decided in 1992.[20] Konstadinides mentions *Van Duyn v Home Office*[21] (1974) but this ruling in fact does not mention custom explicitly.[22] If we search Eur-Lex for 'customary international law' or 'droit international coutumier',[23] the earliest mention

[16]These discrepancies would become even more acute if we were to consider the full gamut of the official EU languages. See, generally, K McAuliffe, 'Hybrid texts and uniform law? The multilingual case law of the Court of Justice of the European Union' 24 (2011) International Journal for the Semiotics of Law 97.

[17]The process of digitizing the older case law into high-quality plain text is labour intensive. In the first place, the decisions have to be correctly delineated, as they come mostly from the Official Journal, which means that the first or the last page can contain text belonging to more than just a single case. Second, the pages of the decision must be fed to an OCR engine that converts the PDF contents into plain text. Most engines, however, struggle with pages that contain texts in columns. In the third and costliest stage, the output of the engine is checked and cleaned to ensure that paragraphs are correctly separated, and the text is free of any major errors. Page numbers and running headers also need to be removed for the plain text of the decision to resemble that of more recent rulings.

[18]The costs involved in manually correcting individual pages is high. As a result, advances in text-generative artificial intelligence tools might prove necessary to complete the digitization task at a lower cost but lower accuracy. In the long run, the EU law community as a whole could gradually correct errors in the text corpus, including those in the source documents, for the benefit of all EU law users.

[19]In total, the IUROPA Text Corpus contains 10,076,341 quasi-paragraphs (regular paragraphs, lines of text in the presentation part, footnotes, etc.) across both languages, compared to around 8,273,331 quasi-paragraphs on Curia and 8,264,982 on Eur-Lex. In addition, the IUROPA texts are cleaner and corrected for obvious errors in the source databases.

[20]Case C-286/90 *Anklagemyndigheden v Poulsen and Diva Navigation* ECLI:EU:C:1992:453. See, for example, P Craig and G de Búrca, *EU Law: Text, Cases and Materials* (Oxford University Press 2020) 406.

[21]Case 41/74 *Van Duyn v Home Office* ECLI:EU:C:1974:133.

[22]T Konstadinides, 'Customary International Law as a Source of EU Law: A Two-Way Fertilization Route' 35 (2016) Yearbook of European Law 513.

[23]Others have written about the peril of choice of translated terms in a multilingual legal system. See, for example, K McAuliffe, 'Language and Law in the European Union: the Multilingual Jurisprudence of the ECJ' in LM Solan, and PM Tiersma (eds), *The Oxford Handbook of Language and Law* (Oxford University Press 2012) 200.

https://doi.org/10.1017/elo.2024.17 Published online by Cambridge University Press

is traced to AG Slynn's opinion in *Hurd* (1985).[24] If we search Curia for the English term, we do obtain the correct result – a little known competition case *Geigy v Commission* decided in 1972 –[25] but there is only a PDF document to work with.[26] Interestingly, if we search Curia for the French term, the results do not include *Geigy* at all.[27] If all decisions were properly digitised and available in plain text – what the IUROPA Text Corpus is working towards – these discrepancies would not arise.

The implications for scholars using CJEU texts in a quantitative analysis are potentially even more profound. No existing text analysis of the CJEU corpus is derived from the full universe of decisions.[28] Even though a comprehensive and validated database of all decisions should be an obvious starting point for quantitative text analysis, the labour involved and academic publishing culture likely disincentivised the creation of a solid CJEU database so far. There is a compelling case for revisiting many existing quantitative findings once such a comprehensive database is fully available.[29] In addition, the advent of the artificial intelligence age means that a bewildering array of new analytical tools becomes available to researchers. Unlike older computational techniques, however, the most advanced techniques nowadays can take advantage of high-quality texts with correct paragraph segmentation, capitalization and punctuation. At the same time, it should go without saying that a more complete database does not automatically translate into better research. Nonetheless, we hope that collating CJEU texts will enable applied researchers to spend less time on relatively unrewarding, technical work and more on coming up with creative research designs.

5. Conclusion

This research note sheds light on discrepancies between the two official sources of CJEU decisions – Curia and Eur-Lex – and the two most used language versions of these decisions (English and French). The proper functioning of their search engines relies on the availability of documents in plain text. However, our analysis shows that coverage in terms of number of decisions differs widely, both between databases and languages. Combining and adding to the two official databases, the IUROPA Text Corpus achieves coverage that is more than 10,000 decisions complete in French than either Curia and Eur-Lex, demonstrating the risks of using either of them in isolation.

The technical and linguistic discrepancies between Curia and Eur-Lex have practical implications for both doctrinal and quantitative scholars. While legal scholars are right to second-guess the accuracy of the search results on either website, quantitatively minded researchers should look to a new database that remedies not only the discrepant coverage but also the absence of many high-quality digitised texts prior to 1990. Moreover, all scholars researching primarily in

[24]Case 44/84 *Hurd v Jones* ECLI:EU:C:1985:222, Opinion of AG Slynn.

[25]Case 52/69 *Geigy v Commission* ECLI:EU:C:1972:73.

[26]Although the Eur-Lex database also contains the PDF of the *Geigy* ruling, the keywords are not identified by the search engine there. Eur-Lex includes a plain-text, capitalised version, but because it omits the first part of the ruling which mentions customary international law, the case is not found via the search.

[27]The likely reason why the search results on Curia differ across the languages has to do with the way the PDF documents are created. Document scans with OCR may enable effective searches – although mileage may vary – while scans without character recognition cannot be searched properly.

[28]See n 2 above. The lack of full coverage is particularly problematic where fairly concrete claims about, among others, the average length of paragraphs and sentences are concerned. See, for example, J Frankenreiter, 'Writing Style and Legal Traditions' in MA Livermore and DN Rockmore (eds) *Law as Data: Computation, Text, and the Future of Legal Analysis* (SFI Press 2019) 153–191.

[29]The availability of a comprehensive text corpus would have an impact on the quality of related databases, such as the network of CJEU citations, as case citations could be transparently retrieved from the plain texts of the decisions. Currently, there is structured citation data available on Eur-Lex, but its accuracy is difficult to assess without insight into how it was generated (and it clearly contains at least some errors).

https://doi.org/10.1017/elo.2024.17 Published online by Cambridge University Press

English (or another language) must contend with the fact that thousands of decisions are only available in French. At the very least, it is worth pausing to reflect on the extent to which the (non-)translation issue affects one's work.

Cite this article: M Ovádek, 'A note of caution on CJEU databases' (2024) *European Law Open*. https://doi.org/10.1017/elo.2024.17