Where law meets data: a practical guide to expert coding in legal research

Michal Ovádek,[1] Phillip Schroeder,[2] Jan Zglinski[3]

Abstract

The rise of empirical methods has had a polarizing effect on legal studies in Europe. On the one hand, quantitative empiricists have frequently dismissed traditional doctrinal scholarship as unscientific and its insights as unreliable. On the other hand, many doctrinal scholars are apprehensive about the perceived displacement of domain expertise from legal research caused by the empirical turn. To bridge the gap between the two camps and address their respective concerns, we propose a wider adoption of expert coding as a methodology for legal research. Expert coding is a method for systematic parsing and representation of phenomena such as legal principles in a structured form, using researchers' subject matter expertise. To facilitate the uptake of expert coding, we provide a step-by-step guide that addresses not only the coding process but also fundamental prerequisites such as conceptualization, operationalization and document selection. We argue that this methodological framework leverages legal scholars' expertise in a more impactful way than traditional doctrinal analyses. We illustrate each step and methodological principle with examples from European Union law.

1. Introduction

In recent years, the landscape of legal research has witnessed a significant transformation, marked by the rising prominence of empirical methods. This shift, fuelled by an increasing interest in systematic data-driven analysis, has had a polarizing effect on the field. Originating in the United States,[4] empirical legal studies have begun to globalise, with both their geographical and substantive reach broadening.[5] In Europe, the integration of these 'new' methodologies has been slower, mainly due to the more limited availability of specialised training, but also due to the lower openness of European universities to accommodate the integration of law into mainstream social science.[6] The ascendancy of empirical methods has, in many places, sparked unease among traditional doctrinal

[1] University College London.

[2] LMU Munich.

[3] LSE Law School.

[4] Although the term 'empirical legal studies' has been popularised only in the 2000s, a variety of research stream had been making use of empirical methods to study legal phenomena, including the law and society and law and economics movements. Earlier precursors can be found among the American Legal Realists, see HM Kritzer, 'Empirical Legal Studies Before 1940: A Bibliographic Essay' 6 (2009) Journal of Empirical Legal Studies 925.

[5] T Eisenberg, T Fisher and I Rosen-Zvi, 'Israel's Supreme Court Appellate Jurisdiction: An Empirical Study' 96 (2010) Cornell Law Review 693; A Meuwese and M Versteeg, 'Quantitative methods for comparative constitutional law' in M Adams and J Bomhoff (eds), *Practice and Theory in Comparative Law* (Cambridge University Press 2012); G Shaffer and T Ginsburg, 'The Empirical Turn in International Legal Scholarship' 106 (2012) American Journal of International Law 1.

[6] A Dyevre, W Wijtvliet and N Lampach, 'The future of European legal scholarship: Empirical Jurisprudence' 26 (2019) Maastricht Journal of European and Comparative Law 348.

scholars, who harbour apprehension over the displacement of domain expertise from the heart of legal scholarship.[7]

In this article, we seek to show that empirical and doctrinal approaches to studying law can be productively combined and chart a guide to a method called 'expert coding' which we believe integrates the best of both worlds. Our starting premise is that knowledge about the law and the institutions implementing it, or legal expertise, holds inherent value and needs to be leveraged. At the same time, we think that traditional doctrinal scholarship, which remains the dominant force in European academia, does not make the most out of the legal expertise it generates. Frequently, legal researchers work without a clear methodological framework, not laying out the parameters and assumptions behind how their assessments of the law are made. This weakens the strength of their findings. We argue that the situation can be improved upon. In proposing a practical guide to expert coding, we build on previous work in law and social sciences,[8] with an eye on presenting the key tenets of the methodology in a manner that should feel particularly familiar to scholars in (and of) Europe. To avoid potential misunderstandings, our argument is not that export coding is the only way of conducting legal research or that it does not come with its own limitations and pitfalls; in fact, we identify the most important challenges that can arise and provide advice on how to navigate them. We do, however, believe that it is a tool that can be useful for many legal researchers and research projects.

Expert coding is, in short, a method for systematically parsing and representing phenomena in structured form by leveraging researchers' knowledge of the studied domain. The outputs of expert coding are 'codes' (sometimes referred to as 'labels' or 'scores') which form the basis of subsequent qualitative or quantitative analysis. We believe that legal scholars possess precisely the kind of domain-level expertise in their subject area that is ideally suited for the adoption of expert coding as a common methodology. Moreover, unlike much of empirical legal scholarship, expert coding does not make unrealistic demands on the types of knowledge legal scholars need to acquire to produce rigorous work. The methodological guide which we propose below espouses principles which augment the skills legal scholars already possess. While our main focus is on how to do expert coding well, we also cover two fundamental and often overlooked pre-requisites of that process: how to think about concepts in a more systematic way and how to select documents for legal research. All along we illustrate each point with examples from European Union (EU) law and in particular the decision-making of the Court of Justice of the European Union (CJEU).

The article follows the lifecycle of an expert coding project in an abbreviated form. The first section introduces a structured approach to thinking about concepts, a crucial part of academic research that is frequently overlooked not only in the legal domain. Second, we consider the role of the input

[7] See, among others, JM Balkin, 'Interdisciplinarity as Colonization' 53 (1986) Washington and Lee Law Review 949; J Goldsmith and A Vermeule, 'Empirical Methodology and Legal Scholarship' 69 (2002) The University of Chicago Law Review 153; I Augsberg, 'Some Realism About New Legal Realism: What's New, What's Legal, What's Real?' 28 (2015) Leiden Journal of International Law 457; N Petersen and K Chatziathanasiou, 'Empirical research in comparative constitutional law: The cool kid on the block or all smoke and mirrors?' 19 (2021) International Journal of Constitutional Law 1810. In the EU context, see G Davies, 'Taming Law: The Risks of Making Doctrinal Analysis the Servant of Empirical Legal Research' in M Bartl and JC Lawrence (eds), *The Politics of European Legal Research* (Edward Elgar 2022).

[8] MA Hall and RF Wright, 'Systematic Content Analysis of Judicial Opinions' 96 (2008) California Law Review 63; JD Clinton and DE Lewis, 'Expert opinion, agency characteristics, and agency preferences' 16 (2008) Political Analysis 3.

data (typically legal texts), how to choose them and where to get them in the realm of EU law. Third, we give advice on how to do expert coding in a way that produces transparent and consistent results. We finish by discussing potential analytical avenues for using the expert-coded output.

2. Making and measuring concepts

As in any other discipline, concepts are the essential building blocks of legal research on EU law. Whatever it is that we are trying to describe or explain in our research, concepts allow us to grasp the phenomena we are interested in and communicate our thoughts succinctly to our audiences. Has the CJEU become more deferential to national authorities over time? How are national lawmakers responding to Europe's rule of law crisis? Does EU law facilitate precarious working conditions for seasonal workers? We cannot start thinking about the answers to these questions without a good sense of what it means for a court to be deferential, what exactly we refer to when we speak of the rule of law, or who would qualify as a seasonal worker let alone the attributes that make work precarious.

The latter example highlights that lawyers and judges are well-accustomed to working with concepts – in fact, they use them all the time. Frequently, a case outcome in court hinges on whether a particular subject falls within the domain of a concept (e.g. whether a Member State's action qualifies as a measure with effects equivalent to quantitative trade restrictions), and legal training serves to interrogate relevant sources of law and scour the facts of a case at hand to argue in favour or against such a qualification in a court of law. Our objective for this contribution is to nurture this talent among legal scholars and transfer it to empirically minded research that makes use of qualitative and quantitative research methods that are commonly applied in the social sciences. We establish a baseline of what is in a concept and share our thoughts on how we might evaluate them – that is, how we would answer the question of what makes a concept 'good'. Further, we discuss how researchers studying EU law can link their concepts to empirical facts, a process referred to as operationalization in the social sciences. Ultimately, embracing advice by the late Giovanni Sartori, we aim to help readers become *conscious* users and developers of concepts.[9]

A. What are concepts?

Before we can discuss what makes a 'good' concept, we need to establish what concepts are. We follow guidance by Munck et al,[10] who themselves draw on seminal work by Odgen and Richards,[11] describing concepts as a set of three related elements: a concept's *term*, its *sense,* and its *reference*. A concept's sense is possibly the most intuitive and arguably the most important element here: it comprises the attributes and, where applicable, the interrelationships between these attributes (more on this later) which lead us to say that something falls within the domain of a concept. For instance, for a court's action to be considered 'deferential', the court's behaviour vis-à-vis its interlocutors needs to be characterised by certain attributes. Studying changing patterns of 'deference' shown in the CJEU's interpretation of EU free movement law, Zglinski argues that over time, the Court has increasingly refrained from making its own legal and regulatory assessments when considering national measures on free movement, and instead allowed national authorities to

[9] G Sartori, 'Concept Misformation in Comparative Politics' 64 (1970) American Political Science Review 4.

[10] GL Munck, J Møller and SE Skaaning, 'Conceptualization and Measurement: Basic Distinctions and Guidelines' in L Curini and RJ Franzese (eds), *The SAGE Handbook of Research Methods in Political Science and International Relations* (Sage 2020).

[11] CK Ogden and IA Richards, *The meaning of meaning: A study of the influence of thought and of the science of symbolism* (Routledge 1923).

determine whether a particular regulatory measure can be justified and is proportionate.[12] The Court's choice not to make such assessments itself and instead leave 'the decision on points of justification and proportionality in the hands of [national authorities]'[13] captures what it means to show 'deference' – the concept's sense.

A concept's term on the other hand can best be understood as a sign or the name that we give to a concept. In our example here, 'deference' serves as the concept's term, it allows the researcher to signify that they are talking about the concept without having to refer to or explain its sense every time it comes up in their discussion. Picking a concept's term is far from trivial and, ideally, we choose a term that is univocal, in other words a term that serves as a unique sign for the concept with the objective to prevent confusion over what researchers are talking about.[14] The term 'deference' allows Zglinski to distinguish his concept from other, ostensibly similar CJEU decisions that Member States might perceive favourably, such as reviewing a measure in line with preferences expressed by Member States in observations to the Court. Finally, a concept's reference captures the actual object it refers to. In our example, the CJEU's actions – more specifically, its judgments that leave the justification and proportionality assessments of regulatory acts in the hands of national authorities – are the reference of the concept.

While all three of these elements – term, sense and reference – collectively constitute a concept, it is generally a concept's sense that takes centre stage in our work. Some concepts are relatively simple, and their sense is conveyed by a single attribute, whereas others comprise a long list of conceptual attributes. Often, however, a concept's sense is not merely an enumeration of conceptual attributes. Rather, when developing concepts, we should consciously think about what Munck et al call the *structure* of a concept, namely the relationships between conceptual attributes and their hierarchy.

Think again of our example of deference in jurisprudence. While the concept's sense discussed above allows us to broadly separate deferential from non-deferential actions, Zglinski makes further distinctions between different types of deference. He calls deference towards national legislatures and executives 'political', while deference towards national courts is called 'judicial'. Political deference is further divided in 'partial deference' on the one hand and 'complete deference' on the other. When granting 'complete deference', the Court not only allows national executives and legislatures to 'take the policy decision they want, they can also choose *how* to reach their decision'.[15] In contrast, when granting 'partial deference', the CJEU gives national authorities 'the freedom to make a certain regulatory assessment but stipulates how this assessment ought to be made'.[16] With these two variants of political deference in hand, we can identify a hierarchy among the attributes that define what it means to be deferential. Leaving the justification and proportionality assessment of a regulatory decision in the hands of national authorities is an attribute that distinguishes deferential actions from non-deferential actions, and conceptually sits at a higher level than the attributes that characterise either 'partial' or 'complete' deference, i.e.

[12] J Zglinski, 'The Rise of Deference: The Margin of Appreciation and Decentralized Judicial Review in EU Free Movement Law' 55 (2018) Common Market Law Review 1341.

[13] Ibid, 1345.

[14] Researchers are frequently tempted to refer to their concept of interest by a term that appears widely in public or academic discourse to make their work more broadly appealing. The risk this practice carries is that of conceptual confusion, whereby the same term comes to denote an increasing array of different senses. Conversely, picking an esoteric term (e.g. an ancient Latin or Greek word) largely disconnected from ordinary language might prohibitively raise the barrier to its re-use by other researchers.

[15] Zglinski (n 12) 1345.

[16] Zglinski (n 12) 1346.

whether the Court specifies a decision-making process that is to be followed. Figure 1 visualises the concept of deference with its two sub-parts, partial and complete deference, highlighting the concept's structure, i.e. the hierarchy and relationship between the different conceptual attributes.

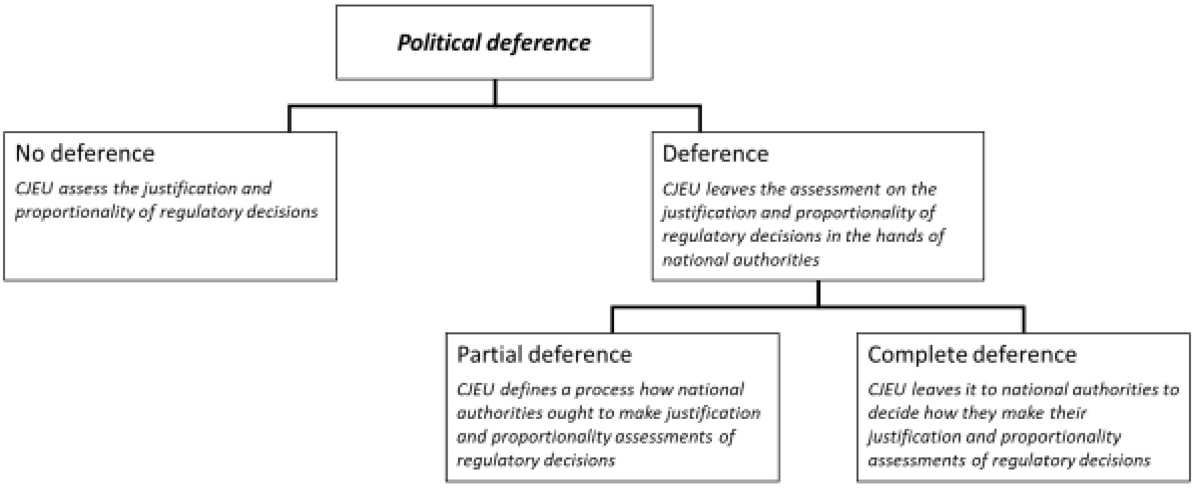
Political deference
No deference
CJEU assess the justification and proportionality of regulatory decisions
Deference
CJEU leaves the assessment on the justification and proportionality of regulatory decisions in the hands of national authorities
Partial deference
CJEU defines a process how national authorities ought to make justification and proportionality assessments of regulatory decisions
Complete deference
CJEU leaves it to national authorities to decide how they make their justification and proportionality assessments of regulatory decisions

Figure 1: The conceptual structure of 'Political Deference'

B. How should we evaluate concepts?

Having clarified what we mean when we refer to concepts and the elements that constitute them, we can now consider what makes 'good' concepts. This may feel unintuitive to some readers, but we would advise researchers not to think about concepts as either true or false. It is futile to argue over whether someone's concept of the 'rule of law', 'judicial independence' or 'landmark rulings' is correct, or that there is a definitive, true sense of a concept. Concepts are constructed by researchers, typically with their own research projects in mind, and reasonable people can disagree over the meaning of a concept – such debates are part and parcel of our work. Instead, we would encourage researchers to consider whether a particular concept, in particular its meaning, is *useful* for their (and ideally others') work. Does a concept help a researcher to achieve their objective for their research project? Where similar concepts already exist, does the research community stand to appreciably benefit from the creation of a new one? Is a concept useful for us when we are trying to collect empirical evidence to test theoretical conjectures about how a legal system works?

There are several criteria that help us gauge the usefulness of concepts, all of which urge us to think carefully about our concepts' attributes. First, useful concepts are characterised by attributes that are *mutually exclusive*. Conceptual attributes are mutually exclusive if individual attributes that sit at the same level do not overlap with each other. To illustrate, consider the distinction between 'partial' and 'complete' deference that we sketched out earlier. In essence, we looked at three attributes, namely 'leaving the outcome of a regulatory assessment in the hands of national governments and legislatures', 'setting rules for the decision-making process of a regulatory assessment', and 'setting no rules for the decision-making process of a regulatory assessment'. While both partial and complete deference share the first higher-level attribute, the two lower-level attributes do not overlap with each other and are thus mutually exclusive – in fact, they are complements: if one applies the other logically must not. This example also highlights why we want to have a good understanding of our concepts' structure. Without a clear picture of a concept's structure, we would find it difficult to evaluate whether conceptual attributes at the same level are mutually exclusive.

Aside from working with mutually exclusive conceptual attributes, we also need to make sure that these attributes are *collectively exhaustive*. We want to avoid forgetting or excluding any essential attributes from our conceptualization. A way to achieve this is to think of our conceptual attributes as individually necessary and collectively sufficient conditions. Each attribute we include in our concept is integral to the latter, and an object that does not possess all our listed attributes consequently does not fall within the domain of our concept. Strictly thinking of attributes in terms of individually necessary and collectively sufficient conditions also pushes us to develop parsimonious concepts. We are forced to cut attributes from our concept that are not central to the concept's meaning. To illustrate, we may initially conjecture that a 'deferential' judgment is also characterised by certain semantic patterns, such as vague language. There is a well-established literature that argues that a vaguely written judgment offers its addressees discretion in how to respond to the judgment, an attribute that we may associate with deference.[17] However, while a judgment may in fact be written in vague language to show deference to governments and legislatures, we are convinced that not every deferential judgment must be vague. Vagueness in language can but does not have to be an attribute of deference. Consequently, we would be well advised to avoid including semantic vagueness as a conceptual attribute – after all, it simply is not a necessary condition.

Figuring out whether our chosen attributes are collectively exhaustive is not easy, even when we are sticking to the advice outlined above. It is ultimately the researcher's call to say that no essential attributes are excluded from a concept. Here, we generally rely on the advice and input from our colleagues and peers to gauge whether we have reached this point. Yet, it remains the case that concepts and their attributes are open for contestation, reinforcing our earlier point that concepts are neither true nor false, only more or less useful.

This brings us to our last criterion for the evaluation of concepts: *conceptual validity*. Following Munck et al.,[18] each decision a researcher makes concerning the exclusion and inclusion of attributes in their conceptualization, as well as the structure of their concept 'can and should be assessed in terms of the extent to which the decision is theoretically justified'. We should have an explanation ready that justifies why we pick a particular attribute for our concept. As authors and users of concepts we should be able to explain how each attribute and their relationship to other attributes contributes to the meaning of our concept. For example, when a researcher decides to include the setting of procedural parameters for a decision-making process as an attribute of the concept 'partial deference', they must have an explanation for doing so. The researcher is working off a theoretically grounded belief that the CJEU's definition of a decision-making process that national authorities ought to follow in their regulatory assessments leads to a qualitatively distinct type of deference and thus warrants inclusion in the concept.

We may think of concepts as building blocks of theory – our expectations of how the world works. A theory may state that one concept is a cause of another concept. For instance, we may conjecture that a higher workload causes a court to show (partial or complete) deference in its jurisprudence more frequently than courts with a less crowded docket. Yet, concepts are more than that, they themselves require careful theorizing and, as Gary Goertz contends, 'it is not possible to easily separate the causal hypotheses within the concept and the causal hypotheses that use the

[17] See e.g. JK Staton and G Vanberg, 'The value of vagueness: delegation, defiance, and judicial opinions' 52 (2008) American Journal of Political Science 504; JF Spriggs, 'Explaining federal bureaucratic compliance with Supreme Court opinions' 50 (1997) Political Research Quarterly 567.

[18] Munck (n 10) 336.

concept'.[19] The attributes of our concepts often exercise a causal force on another concept, and the criterion of conceptual validity demands that we should have an explanation for why this is the case. While we do not have the space here to elaborate on the role of theory in legal research, the interlinkage between concepts and theory demands that their crafting is a joined-up exercise.

The preceding paragraphs on what makes a concept useful carry special importance in legal research. The ubiquity of concepts in law – be they legislative or judge-made in source – tempts researchers to replace their own conceptualization with references to definitions appearing in legislation or case law. If academic legal research intends to be independent from the practice of law, we believe, similarly to Van Gestel and Micklitz,[20] that legal researchers must be in charge of the concepts that underpin their discipline. The standards and motivations of judicial or legislative conceptualization are, as outlined above, inherently and substantively different from scholarly conceptualization. Judges' primary objective when constructing a legal concept is not its usefulness in answering research questions and illuminating theories. Judicial conceptualization is influenced by arguments of the parties, case law precedents, considerations of efficiency and cognitive biases rather than academic usefulness. In addition, judicial and legislative concepts are frequently imprecise to facilitate compromises between litigants or politicians. Concepts employed by scholars can benefit from the luxury of being crafted purely for the purposes of research, which should give them an analytical and critical edge compared to concepts from legal practice.[21]

C. Measuring concepts

For many of our research projects, we are not satisfied with merely developing concepts that allow us to grasp a phenomenon or establish theories about the relationship between concepts. Often, we are interested in finding out whether there is any empirical evidence that supports our conjectures, and we want to identify empirical patterns linked to our concept(s) of interest. This pushes us to think about how we can – in social science parlance – *measure* our concepts. Such measurement, however, is tricky: our concepts typically capture phenomena that are not directly observable. For instance, we may be interested in identifying 'landmark rulings' published by the CJEU in the past two decades. To do so, we have to come up with a concept of a 'landmark ruling', discussed its conceptual attributes and structure. Yet, even with a good conceptualization in hand, CJEU judgments do not come with a tag characterising themselves as a 'landmark rulings', and even if the Court did describe its judgment as such, we would have reasons to be sceptical.[22]

A researcher interested in measuring their concepts then must develop what is referred to as 'measurement instruments'. Put simply, we need to identify indicators that link our ideas on unobservable concepts to observable facts.[23] Once we engage in measurement and select indicators,

[19] G Goertz, *Social science concepts: A user's guide* (Princeton University Press 2006) 55.

[20] R Van Gestel and HW Micklitz, 'Why Methods Matter in European Legal Scholarship' 20 (2014) European Law Journal 292.

[21] This is not to say that concepts developed by judges (or legislators, who have their own set of diverging motivations) cannot inspire scholarly conceptualization. But instead of taking them at face value, we encourage researchers to think critically about what purpose a concept – proportionality, subsidiarity and so on – was intended to serve in a line of case law or a piece of legislation, and what factors are likely to inhibit its analytical leverage.

[22] U Šadl and MR Madsen, 'A Selfie from Luxembourg: The Court of Justice's Self-Image and the Fabrication of Pre-Accession Case-Law Dossiers' 22 (2015) Columbia Journal of European Law 327.

[23] R Adcock and D Collier, 'Measurement validity: A shared standard for qualitative and quantitative research' 95 (2001) American political science review 529.

the question of validity rears its head again. However, at this stage in our work, we are no longer concerned with the validity of our concept and its attributes. If we had doubts about conceptual validity, we should not have proceeded to measurement. Instead, we care about what Adcock and Collier refer to as 'content validity', that is whether the indicators we have chosen capture collectively the sense of our concept.[24] Intuitively, our indicators of choice should thus be closely linked to a concept's attributes.

If we consider, for instance, a landmark ruling to be characterised by the attribute that it 'establishes a legal rule or principle that is employed to resolve future issues',[25] we should find indicators that capture the presence (or degree) of this attribute. Naturally, scholars seeking to identify landmark judgments in the CJEU's jurisprudence have turned to citation patterns between judgments as indicators for landmark rulings.[26] This choice of indicator is based on the theoretically informed assumption that CJEU judgments which are cited extensively are the most authoritative judgments – the kind of judgments establishing precedent and employed to resolve subsequent cases.[27]

We also need to decide on an appropriate measurement scale for each indicator we have chosen. Measurement scales define the values that the researcher can assign to an indicator. We distinguish between four types of measurement scales, which differ with respect to their precision. At the bottom of this hierarchy of measurement scales are *nominal scales*. Nominal scales comprise two or more mutually exclusive categories that cannot be ranked. For instance, think of the type of applicant in court ('private individual', 'commercial enterprise', 'public authority' or 'other'). O*rdinal scales* also comprise mutually exclusive categories, but with categories following a rank order. Stone Sweet and Brunell, for example, define an ordinal scale that captures the position of a national court in its domestic judicial hierarchy.[28] Their scale ranges from lower courts (i.e. courts that ordinarily decide cases in the first instance) at the bottom to higher courts (i.e. courts whose decisions cannot be appealed) at the top and intermediate courts in the middle.

Next are *interval scales*, which comprise numeric values with equal intervals between neighbouring values. Consider for example the grading scale used in German state examinations for law students, which ranges from 0 to 18 points. Here, we find the same interval between the scores 5 and 8 on the one hand, and 13 and 16 on the other. Note also that the value 0 is an arbitrary choice for the lower end of the scale. State examiners could have let the scale range from -5 to 13, and the scale would have conveyed the same information about students' aptitude. This latter point distinguishes interval scales from *ratio scales*, the final and most precise measurement scale. Ratio scales also

[24] Ibid.

[25] M Derlén and J Lindholm, 'Goodbye *van Gend en Loos*, hello *Bosman*? Using network analysis to measure the importance of individual CJEU judgments' 20 (2014) European Law Journal 667, 668.

[26] See, for example, U Šadl and S Hink, 'Precedent in the Sui Generis Legal Order: A Mine Run Approach' 20 (2014) European Law Journal 544. A more process-oriented conceptualization of 'landmark rulings' relies on the Court formation rendering the judgment. See M Ovádek, 'The making of landmark rulings in the European Union: the case of national judicial independence' 30 (2023) Journal of European Public Policy 1119; M Ovádek, W Wijtvliet and M Glavina, 'Which Courts Matter Most? Measuring Importance in the EU Preliminary Reference System' 12 (2020) European Journal of Legal Studies 121.

[27] Measurement instruments are rarely perfect, but some are more defensible than others. We seek indicators that map as closely onto our concepts as possible, but there are limitations to what we can observe (and even more limitations to what we can measure). The important thing is therefore to always justify the choice of an indicator.

[28] A Stone Sweet and TL Brunell, 'The European Court and the national courts: a statistical analysis of preliminary references, 1961–95' 5 (1998) Journal of European Public Policy 66.

comprise numeric values with equal spacing between neighbouring values yet are defined by a natural zero reference point. To illustrate, the number of citations to sources of primary EU law in the CJEU's is measured on a ratio scale; we could count seven or twenty citations but never less than zero, which indicates the absence of the measured phenomenon.

We generally advise to seek measurement scales that are higher in precision, meaning we would favour an indicator measured on an interval scale over an ordinal scale. Following Epstein and Martin, we believe that 'more detail rather than less detail is a principle worth remembering',[29] not least because – if deemed necessary – we can transform a ratio scale into an ordinal scale at a later stage in our work but not vice versa. Likewise, we prefer scales with more over fewer possible values for the researcher to choose from. Starting out with a more fine-grained measurement scale, we can easily collapse several values into a single value should we at some stage in our work realise that it is necessary to do so. In contrast, splitting an existing value into two or more fine-grained values is not possible without revisiting our empirical material and investing additional effort in re-redoing at least some of our coding.

However, adding values to our measurement scale without second thought or favouring higher precision at all costs is not a silver bullet. The same kind of guiding principles we had highlighted in our discussion of selecting conceptual attributes above are equally relevant when thinking about the choice of values on our measurement scales. First, we want our values to be *mutually exclusive*, which is particularly important for nominal and ordinal scales. To illustrate, adding the category 'appellate courts' to an ordinal scale capturing courts' position in the domestic judicial hierarchy might be unhelpful. Such a category would overlap with both courts at the top of the hierarchy and in the middle, and thus adds confusion rather than precision. Further, we want the values on our measurement scales to be *exhaustive*. Consider a nominal scale that measures the type of applicant in court, comprising the categories 'private individual', 'commercial enterprise' and 'public authority'. While these values certainly capture applicant types in a wide range of cases, we would not know what to do if we ever came across a case in which an applicant is a non-profit organization as such an organization does not fit any of the aforementioned labels. A simple remedy to ensure that the values on our (nominal) scales are exhaustive is to add a value 'other', a catch-all category for objects that do not fit into any other category.

Finally, decisions such as whether a non-profit organization deserves its own category as an applicant type on our measurement scale should be guided by theory. Do we have a theoretically justified reason to include a particular category in our measurement scale (e.g. in the context of our explanation of a particular phenomenon, does it matter whether an applicant is a for- or non-profit organization)? Do we need a particular value on our scale to test a theoretical expectation in our empirical analysis (e.g. if our theory states that non-profit organizations are more likely than private individuals to argue in a national court that their case should be referred to the CJEU, then we most certainly want to include this category in our scale).

To conclude, when we are developing our concepts and the instruments that measure them, we should make *conscious* choices and have explanations ready for why we pick a particular conceptual attribute or select a particular indicator to measure the concept. These explanations are in essence theories of how the world – or at least the objects that we study – works, which is part of why developing good concepts is difficult. After all, we need concepts to formulate theories but at the same time we need theories to develop concepts, a conundrum known as Kaplan's paradox of

[29] L Epstein and AD Martin, *An Introduction to Empirical Legal Research* (Oxford University Press 2014) 92.

conceptualization.[30] Luckily, we rarely have to start from scratch as there is no shortage of existing theories of how courts, individual judges, litigants, lawmakers and others behave. We can draw on these theories to develop and measure our concepts, carry out our empirical studies and reflect on the uncovered evidence, which in turn improves future theories.

3. Data collection

Once the conceptual work is done, we typically look for data. The widely used term 'data' may feel intimidating to legal scholars at first, but it need not be. Data denotes information about the world, which can be captured by numerical or non-numerical values.[31] Legal academics, some without realising it, are used to collecting data in their day-to-day research. They will, for instance, read a number of competition law judgments to find out how the CJEU has defined the concept of an undertaking, or check the Court's latest annual report to find out how many fundamental rights cases were brought over the past year. All of this material constitutes data that, in principle, can form the basis of an empirical analysis.

In legal research in general, and in EU law research in particular, legal texts are the main source of data.[32] This notably includes court rulings, legislative measures (as well as constitutional materials[33]), and administrative acts. Most empirical legal research focuses on extracting and analysing information contained in these documents. But there is a variety of data beyond those classical sources which can be productively made use of. In EU law, for instance, studies have appeared that analyse the role of the legal services of the EU institutions by conducting interviews with their members,[34] determine the salience of CJEU rulings by looking at their media coverage,[35] or examine compliance with EU consumer law by trawling through the contractual terms and conditions of online sellers.[36] In this section, however, we focus on examples of traditional sources of data – notably court decisions – which can most obviously serve as input for expert coding.

A. Population and sampling

When thinking about how much and what kind of data needs to be collected for a study, the starting point should be the intended breadth of the research. Are we trying to make claims about a single line of case law (e.g. case law on direct effect), the decision-making practice of a court (e.g. the General Court) or an entire legal system (e.g. EU law)? The group of rulings (or measures, rules, etc.) that our research conclusions intend to speak to is called the target population or population of interest. When we want to draw conclusions about a single line of case law, the size of our target population – as expressed by the number of decisions that we want to draw conclusions about – is going to be a lot smaller than if we are trying to make broad statements about CJEU decision-making in general.

[30] A Kaplan, *The conduct of inquiry: Methodology for behavioural science* (Chandler Publishing 1964).

[31] For one among many, see L Epstein and G King, 'The Rules of Inference' 69 (2002) The University of Chicago Law Review 1.

[32] MA Livermore, *Law as Data: Computation, Text, & the Future of Legal Analysis* (SFI Press 2018).

[33] See e.g. A Chilton and M Versteeg, *How Constitutional Rights Matter* (Oxford University Press 2020).

[34] P Leino-Sandberg, *The Politics of Legal Expertise in EU Policy-Making* (Cambridge University Press 2021).

[35] J Dederke, 'CJEU Judgments in the News – Capturing the Public Salience of Decisions of the EU's Highest Court' 29 (2022) Journal of European Public Policy 609.

[36] F Pflücke, *Compliance with European Consumer Law: The Case of E-Commerce* (Oxford University Press 2024).

When the target population is relatively small, we might be able to devote a lot of time to each observation (e.g. a ruling) and still cover the entire population of interest. However, covering all observations with labour-intensive methods of research such as expert coding would prove more difficult when our population of interest is very large. When examining the entire population is not possible, we collect data on a so-called 'sample', a subset of the population. Take, for instance, the question whether the Court of Justice engages in judicial activism, which has long preoccupied EU lawyers.[37] Scholars have defined the concept of judicial activism differently over the years, but under one common definition the CJEU is activist when it applies teleological methods of interpretation like the *effet utile* doctrine.[38] Given that the Court has rendered over 22,000 decisions by now, and assuming we cannot rely solely on automation to code them, we will have to choose a smaller, manageable number of decisions, say 1,000, to read and analyse. In this example, the 1,000 cases are our sample and the 22,000 decisions our population.[39]

It is important to keep in mind that even when we are working with a sample, we do so to draw inferences (conclusions) about the population, not merely the sample. In other words, we want to *generalise* from the sample to the population. We can only generalise, in the sense of drawing valid conclusions from the sample about the population, when the characteristics of the sample are *representative* of the population. By far the most and arguably only reliable way of ensuring that the sample is representative is to draw it randomly from the population.[40] Because in random sampling every member of the population (e.g. a court decision) has an equal chance of being selected, well-established theorems of probability theory[41] have shown that a growing sample increasingly resembles the population.[42]

If we are interested in judicial activism writ large, examining the CJEU's reasoning in five or six judgments will not be enough to paint a representative picture of the entire case law, given its size. But how many random samples do we need to draw before the sample can be considered representative? While the mathematical equation required to calculate how large a sample needs to

[37] H Rasmussen, *On Law and Policy in the European Court of Justice* (Nijhoff 1986); M Dawson, B de Witte and E Muir (eds), *Judicial Activism at the European Court of Justice* (Edward Elgar 2013); A Arnull, 'Judicial Activism and the Court of Justice: How Should Academics Respond?' (2013) Maastricht Faculty of Law Working Paper 2012/3.

[38] T Tridimas, 'The Court of Justice and Judicial Activism' 21 (1996) European Law Review 199; A Grimmel, 'Judicial Interpretation or Judicial Activism? The Legacy of Rationalism in the Studies of the European Court of Justice' 18 (2012) European Law Journal 518.

[39] When faced with a numerous target population, legal scholars sometimes try to make their life easier by excluding a certain subset of cases where the phenomenon of interest is unlikely to be present. However, this reduces the target population and therefore the breadth of the subsequent analytical inferences. For example, someone researching judicial activism might exclude all staff cases from their analysis. While such a choice can be justified on grounds of resource constraints, it alters the scope of the conclusions that we can draw from an analysis of a sample. When a segment of the population was excluded from the analysis, it is important to transparently communicate what the new population of interest consists of.

[40] All popular software environments for working with data (such as R and Python) include simple functions to make random draws. While we recommend using a computer, blindly picking shuffled pieces of paper from a bag would work just as well.

[41] Specifically, the law of large numbers.

[42] The sample is by definition never exactly the same as the population – the discrepancy between the characteristics of the sample and of the population is called sampling error. The larger the sample, the smaller the sampling error.

be for it to be representative of the population may seem daunting,[43] there is an intuitive way of approximately checking for representativeness. Once we have drawn a random sample of a certain size, we can compare the sample and population on a relevant characteristic (variable). For example, if we know that in the population of all rulings produced by the Court of Justice in a year around 7% on average are rendered by the Grand Chamber, we would expect a sufficiently large random sample to contain approximately the same proportion of Grand Chamber decisions. Similarly, because the CJEU produces many more rulings nowadays than it used to decades ago, the skewed ratio between older and more recent cases should be approximately replicated in the random sample (see Figure 2).

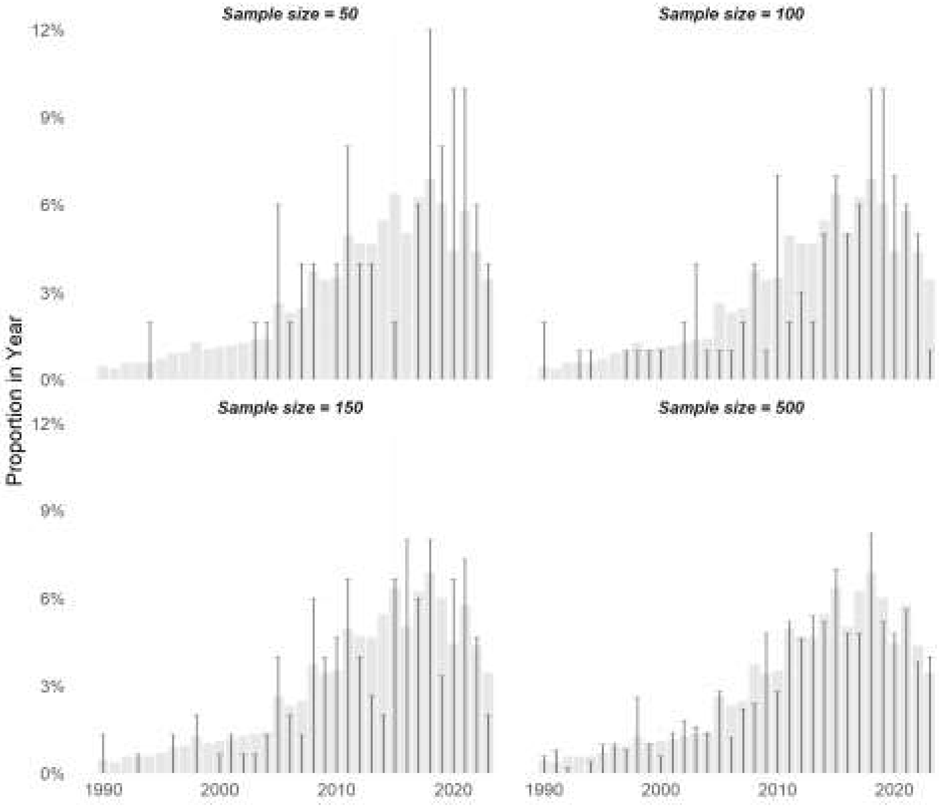

Figure 2: The shaded columns show the (true) over-time distribution of the population of all General Court decisions (N = 14,706) in terms of proportion per year. The vertical lines display the yearly proportions when randomly drawing samples of different sizes (without stratification) from the

[43] Assuming the most common confidence level (95%), the standard equation for calculating the size of a representative sample requires three inputs: the size of the population (N), the margin of error (E) and the variance (V) of a characteristic that matters to the analysis. *E* tells us how close the estimates from a sample are to the population (with a confidence level) and is usually set at 5%. *V* captures the variability in a characteristic of interest. Using the default values for confidence level (which implies a critical value of 1.96) and *E* and assuming we care about the proportion of preliminary rulings in the sample (around 70%), the following equation recovers the necessary sample size to achieve representativeness:

$$\left(\frac{Z(0.95)^2 * V}{E^2}\right) * \frac{N}{N + \left(Z(0.95)^2 * \frac{V}{E^2}\right)} = \left(\frac{1.96^2 * 0.7 * (1 - 0.7)}{0.05^2}\right) * \frac{22000}{22000 + \left(1.96^2 * \frac{0.7 * (1 - 0.7)}{0.05^2}\right)} = 318$$

Given these parameters, we need 318 samples to be able to make conclusions about the population. We only show the calculation steps for illustration here – in practice, there are many tools online or in statistical software that can painlessly recover the required sample size given a set of inputs.

population. We can see that the larger the sample size, the more it resembles the population distribution of rulings over time.

Note that simple random sampling alone does not guarantee[44] that the resulting sample will be balanced with respect to a characteristic of interest. Imagine we want to make sure that our sample contains a proportional number of decisions from each year. For the sample to contain the approximately the same proportion of decisions in each year as the population, we need to adjust our sampling strategy by first dividing the population in so-called *strata*, which in our example are years. Now we randomly draw from each *stratum* (year) a proportional number of decisions. If the number of decisions produced by the Court of Justice in 1987 represents around 1.5% of the 22,000 decisions ever produced, then we sample the same proportion from this stratum and repeat the process for every year to achieve better balance with respect to years.[45]

A key contribution of random sampling, stratified or not, is that it prevents selection bias. Many doctrinal studies suffer from selection bias, because even though they often wish to make claims about a court in general, scholars tend to analyse a sample of cases that is not representative of the population of interest. Typically, cases that receive media attention or are frequently cited by courts are more likely to be analysed than more ‘run-of-the-mill’ cases, which makes the sample suffer from selection bias.[46] Selection bias leads us to over- or underestimate the true incidence of a phenomenon such as judicial activism in the target population.

However, some research designs might require us to select specific cases.[47] When done intentionally, this is called purposive sampling. By itself, purposive sampling does not enable the subsequent analysis to generalise to the population of interest, because the characteristics of the purposive sample do not approximate those of the population, thus creating the risk of selection bias. But we can combine purposive and random sampling to produce a more informative piece of research. We might, for example, study judicial activism in a selection of landmark cases – because these might be particularly important to find out about – and then apply the same research method to a ‘control’ group consisting of the same number of randomly sampled decisions. The latter would give us an idea of how well the findings from the hand-picked observations generalise to the broader population. But note the loss of overall representativeness, as this procedure produces an overrepresentation of landmark rulings in the sample under examination.[48]

When putting together a dataset, it is important to keep in mind the unit of analysis. The unit of analysis indicates the level at which the entities we are interested in are located. In most doctrinal research, the unit of analysis is a court ruling. Each observation in our dataset is therefore of one ruling (or whatever other unit of analysis is being used). Variables, including the actual text of a

[44] But of course, as illustrated in Figure 2, drawing an ever-larger sample will make it resemble the population ever so closely (across every characteristic).

[45] If we are drawing a sample of 1000 decisions in total, then we want to draw 15 decisions from the 1987 group of rulings (1000 * 0.015).

[46] Research has shown that there are large discrepancies between areas of law studied by legal scholars. See A Dyevre, M Glavina and M Ovádek, ‘The Voices of European Law: Legislators, Judges and Law Professors’ 22 (2021) German Law Journal 956.

[47] Especially when the cases we want to include are very rare, pure random sampling might be unlikely to produce a sample that includes enough of them.

[48] The ratio of purposively and randomly sampled cases (in the example it is 1 to 1) can be adjusted to make the sample resemble the population more closely.

decision, have values across all our observations.[49] Even if our observations are often nested in larger units – e.g. a paragraph is nested in a ruling which is nested in a court – we want to make sure that the unit of analysis remains the same throughout a dataset. Retaining such consistency requires clarity regarding what precisely we are intending to study.[50]

B. Data sources

Even with a good sampling strategy in mind, the question remains where to find data for our study. The good news is that thanks to technological advances and the increased interest in empirical legal research a lot of data, including on EU law, is available, and it steadily keeps on growing. The bad news is that it can be difficult to locate it. Attempts have been made to create lists of existing data sources on EU law, but because of the decentralised and dynamic nature of empirical legal research they are necessarily incomplete.[51] The first port of call, especially for projects concerning EU adjudication or legislation, will often be the two official EU databases: Curia and EUR-Lex. Whereas the former is exclusively dedicated to the CJEU's case law, the latter additionally includes data on European as well as, to a more limited extent, national law. Both databases are freely available and, in addition to the raw primary data (court rulings and legal acts), contain useful metadata, for instance on the date, type and result of the legal proceedings, the composition of the Court and the originating Member State of a dispute.[52] They also allow for full text searches, meaning that researchers can look for certain keywords or legal norms relied on, although there are issues here with many older documents which have not been properly digitised.[53] Researchers have developed packages for statistical programmes like R to facilitate the retrieval of data from Eur-Lex.[54]

In addition to those official sources, a number of researcher-compiled databases have emerged over the years. These vary significantly in topic, scope, and quality. On the one end of the spectrum, there are numerous datasets that were created for a single research project only. Some of them are public (and then dispersed across various repositories and personal websites), others are not. These types of bespoke datasets are usually smaller in size and limited to a particular subject-matter. This can range from survey data on 640 judges from four Member States on their age, career and

[49] In practical terms, most scholars use simple rectangular spreadsheets to store their data and we recommend sticking to the rule that each row contains the values of one observation (e.g. a ruling), while each column contains the values for one variable (e.g. the date of decision). For best practices on organizing data in spreadsheets, see KW Broman and KH Woo, 'Data organization in spreadsheets' 72 (2018) The American Statistician 2.

[50] The various units of a legal system – judges, courts, rulings, laws etc. – are to various degrees connected. Clarity in research purpose means that we focus on one issue at a time and keep different units of analysis in separate datasets. For example, if we are interested in the effects of legislation on court decisions, we would probably start by coding the content of the rulings first (decision as unit of analysis). Afterwards, we might want to aggregate the results in a new dataset by for instance counting the number of judgments invoking each law (legislation as unit of analysis).

[51] See https://github.com/michalovadek/eudata.

[52] Although it is worth paying attention to the many omissions and errors in the metadata on both official websites.

[53] See M Ovádek, 'Note of caution on CJEU databases' (2024) European Law Open.

[54] 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 (2021) Political Research Exchange 1870150.

education,[55] to information on plaintiffs, outcomes, procedures and written observations in over 3,000 legal issues decided by the CJEU.[56] There are also more ambitious datasets, commonly compiled as part of larger research projects, which include a higher number of observations and cover more substantive ground. The datasets created as part of EUTHORITY, which investigate various aspects of the relationship between domestic and supranational courts in the EU,[57] and the Berlin Infringement Database, which provides comprehensive data on 13,367 infringement cases from 1978 to 2019,[58] are examples of this. Finally, the first multi-user databases on EU law, such as the IUROPA Project, have surfaced.[59] These are collections that have been developed with the intention of serving not just one, but a multitude of different research projects. At present, IUROPA contains a complete collection of data about CJEU jurisprudence,[60] with information exceeding that provided in Curia, and provides full text corpus data, solving some of the aforementioned problems surrounding digitisation.[61] IUROPA also provides access to data on actors' positions on legal issues,[62] doctrinal outcomes in selected CJEU judgments as well as information on national courts which referred references for preliminary rulings to the CJEU.

Sometimes the data we need for our research is already available, other times it must be gathered. As a general rule, it will be more efficient to use existing data than to collect it from scratch, at least if that data is of a reasonably high quality.[63] Data collection can be a time and resource-consuming process, so if someone else has done the work, it would be a waste of effort to duplicate it. Note, however, that even where relevant data exists, it might not cover everything we need for our research project. This, quite simply, has to do with the fact that it was compiled by another scholar for their own research project, which may have some overlaps but will rarely fully align with the questions we want to ask, the concepts we employ and the hypotheses we formulate. It may, for instance, happen that we are interested in studying the evolution of preliminary references and data on this topic is available, but only until 2015. In such a case, we may make use of that data but update it for proceedings that have taken place since. Or there might be complete data on the outcomes of direct actions before the CJEU, but no information about the type of dispute at stake. We may take over the information on outcomes and supplement it with information on the rights that were invoked by the parties, for example to see whether human rights claims make it more likely for the Court to quash an EU or Member State act.

[55] JA Mayoral, U Jaremba and T Nowak, 'Creating EU law judges: the role of generational differences, legal education and judicial career paths in national judges' assessment regarding EU law knowledge' 21 (2014) Journal of European Public Policy 1120.

[56] CJ Carrubba, M Gabel and C Hankla, 'Judicial Behavior under Political Constraints: Evidence from the European Court of Justice' 102 (2008) American Political Science Review 435.

[57] https://www.law.kuleuven.be/euthority/EN.

[58] https://www.polsoz.fu-berlin.de/en/polwiss/forschung/international/europa/bid/index.html.

[59] https://www.iuropa.pol.gu.se/.

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

[61] M Ovádek, JC Fjelstul, D Naurin and J Lindholm, 'The IUROPA Text Corpus', in J Lindholm et al, The Court of Justice of the European Union (CJEU) Database, 2023, IUROPA, https://iuropa.pol.gu.se/.

[62] P Schroeder and J Lindholm, 'From One to Many: Identifying Issues in CJEU Jurisprudence' 11 Journal of Law and Courts 163.

[63] This can be verified, for instance, by checking the completeness of a dataset, assessing the plausibility of the conceptual definitions, going through the codebook, and recoding a small subset of cases to see whether the coding is satisfactory.

Obtaining new data can be challenging at times. Certain types of data may not be available at all or only hard to get. This may, in some cases, simply be due to difficulties with finding or getting access to it, for instance when certain records have not been digitised and require travelling to an archive which is located far away from where the researcher is based.[64] But there can also be harder, including legal, limitations affecting data collection. Consider, for example, the behaviour of CJEU judges during the so-called 'délibéré', the part of a proceeding during which the members of the Court come together to discuss a case's merits and determine the outcome. There is a number of fascinating questions that could be asked about this process: how often do judges agree on the outcome? Are certain judges more persuasive in advocating their viewpoints than others? And which characteristics (seniority, gender, professional background etc.) affect that? It would, in theory, be possible to get to the bottom of all of these issues by placing a camera inside the 'room where it happens' or interviewing judges afterwards about the content of their discussions. Alas, the CJEU's principle of secrecy prohibits the revealing of this type of information. It is legally impossible to gather (direct[65]) data on this aspect of judicial behaviour.

Where a researcher is refused access to certain data, but believes they are entitled to it, they can bring legal action or administrative proceedings. There are some prominent recent examples of such 'data activism' in EU law. Laurent Pech tried to obtain the opinion of the Council's Legal Service concerning a proposal for a regulation on the protection of the EU's budget in case of generalised rule of law deficiencies, a precursor to the Rule of Law Conditionality Regulation. The request was denied based on a lack of an overriding public interest in disclosure. In response, Pech initiated *De Capitani*-style proceedings before the General Court and successfully challenged the decision.[66] Others have pursued a similar strategy when trying to obtain emails sent by EU institutions that are automatically deleted after a certain period.[67] Freedom of information requests of this kind can be useful, even essential, to obtaining the data needed for one's research. However, it is important to factor in access issues when embarking on an empirical project. The inherent uncertainties and potential delays with obtaining data may significantly affect the feasibility and timeline of the project.

A final peculiarity of EU law data should be noted: its multi-lingual nature. The EU currently has 24 official languages. Consequently, many legal documents are translated and appear in different versions. This creates two potential problems for empirical legal research. The first is that not all documents are translated, and those that are will not always be available in all official languages. The CJEU's case law is a point in case. The Court's working language is French, an outlier among EU institutions. Therefore, decisions are initially written in French and only then translated into other languages. This, of course, constitutes a gargantuan task for the language services. To protect resources, the Court does not translate all of its decisions. As a result, the body of case law available in French exceeds that in other languages – in the case of English by over 10,000 documents. The

[64] The historical archives of the Court of Justice are, for instance, based at the European University Institute in Florence.

[65] Scholars like Vera Fritz and Morten Rasmussen have tried to unveil the dynamics inside the Court of Justice by relying on notes, diaries, and biographies of former judges; see V Fritz, 'Judge Biographies as a Methodology to Grasp the Dynamics inside the CJEU and Its Relationship with EU Member States' in MR Madsen, F Nicola and A Vauchez (eds), *Researching the European Court of Justice* (Cambridge University Press 2022).

[66] Case T-252/19 *Laurent Pech v Council of the European Union* [2021] ECLI:EU:T:2021:203; Case C-408/21 P *Council v Pech* [2023] ECLI:EU:C:2023:461.

[67] See https://euobserver.com/rule-of-law/157076.

choice of language thus significantly affects the possibilities for research and potentially the findings as well.[68]

Translation presents challenges in its own right. Certain terms may mean different things in different languages. Others, while superficially different, may, in substance, denote the same phenomenon. This can become particularly relevant when studying doctrinal concepts. To come back to the example of judicial deference: whereas the Court predominantly uses the term 'marge d'appréciation' to signal the use of the doctrine in the French versions of its judgments, there is far more heterogeneity in the English translations, with options ranging from 'degree of latitude', to 'power of assessment', to 'area of discretion' or simply 'discretion'.[69] As scholars tend to work in English and, consequently, more often than not with translated Court documents, any discrepancies need to be accounted for in the data collection and coding process.

4. Doing expert coding

Armed with a well-defined concept on the one hand, and the necessary data on the other, we are ready to 'code' (or 'label', 'classify', 'score'[70]) our observations. Expert coding is the process of assigning labels (which may or may not be numerical) to data observations (such as texts of court rulings) according to a pre-determined scheme (the codebook) to create new (and robust) variables for subsequent (qualitative or quantitative) analysis.[71] It is particularly relevant for legal research, as many concepts of interest are not directly observable and require evaluation by researchers with knowledge of the law (experts). As outlined above, what data we choose to code and how will be closely linked with ideas on how to operationalise our concepts and, more broadly, the overarching research question that we are aiming to answer.

How to conduct expert coding in practice depends on available human and technical resources. In an ideal world, the project leader has at their disposal enough coders that each case can be scored independently by at least two well-trained individuals. In practice, the majority of legal research is undertaken by a single scholar with no budget to hire research assistants. We offer general as well as more specific guidelines for both scenarios.

A. Choices in expert coding

Regardless of the specific setup, we contend that the overriding concern in any expert coding task is understanding and mitigating the factors affecting the *difficulty* of making consistent coding choices. There are likely to be many such factors, some of which are easier to navigate than others.

The easiest factors to rein in are typically technical in nature. Even seemingly innocuous and mundane choices are worth reflecting on. Does the coding take place in a spreadsheet or another software? How are the values (labels) recorded? Is the source text easy to read? How many manual actions – mouse clicks, keyboard strokes and so on – does the coder need to perform for each observation? Does the coder score texts on a single or several variables at a time?

These concerns are not merely about the efficiency of the coding process, which is important in its own right. The technical setup can influence the coding outcomes as well. A text that is difficult to read or labels impractical to assign are liable to frustrate and tire the coder. A frustrated or tired

[68] Ovádek 2024 (n 53).

[69] J Zglinski, *Europe's Passive Virtues: Deference to National Authorities in EU Free Movement Law* (Oxford University Press 2020) 19.

[70] Adcock and Collier (n 23).

[71] KL Marquardt and D Pemstein, 'IRT models for expert-coded panel data' 26 (2018) Political Analysis 431.

coder is more likely to produce labels that are less consistent. In general, we recommend making coding tasks as simple and easy to carry out as possible. As every researcher knows from experience, human concentration is fickle;[72] the more cognitive effort required from coders, the more likely they make inconsistent decisions or mistakes. In practical terms, a technical mitigation measure can mean, for example, splitting a long piece of text into several easier-to-read chunks.

Nonetheless, the major part of coding difficulty resides in the actual cognitive process of coding a piece of text. Here, again, we recommend making the coding task as simple as possible. At one end of the spectrum, coding decisions can be so straightforward as to be better served by some form of automation. A typical example is coding texts for presence of keywords. We might be interested in measuring the prevalence of the proportionality test. In an ideal world, this doctrine would be consistently identifiable by a set of keywords, such as 'proportionality', 'test' and 'necessity'. In reality, legal writing might be too inconsistent for such an approach to yield satisfactory results. The CJEU sometimes addresses proportionality without using the usual keywords.[73] Or perhaps we might come to the realization that there is not one proportionality test but several shades of the underlying principle, complicating our original, simple measurement scheme.

Still, if a coding task can be adequately served – even if only in part – by an automated process such as a keyword search, we recommend opting for this approach. Not only does the machine not suffer fatigue and as such is less prone to making errors, but an automated coding process is both easier and cheaper to replicate. Another researcher can obtain the same results by simply running our program, rather than hiring an army of research assistants or individually revisiting every coding decision. Moreover, labels obtained through deterministic (as opposed to probabilistic) automation systems are very certain. There is no uncertainty around coding decisions made by exact matching – a keyword is either present or it is not.

Automating a part or all of the coding task should not be viewed as diminishing the importance of the expert in charge. If a keyword search is a sufficient tool for the coding process, the expert's knowledge is critical in determining which keywords are optimal for the task and what the potential pitfalls are. For example, only someone with a high familiarity with EU law will know that the presence of the phrase 'legal basis' in a paragraph of a CJEU decision is not sufficient to establish that the decision addresses the issue of the choice of legal basis of EU legislation.[74] Ensuring that an automated process – even as simple as a keyword search – retrieves the desired information is a necessary step often referred to as validation.[75] Validation is typically carried out against a 'gold standard' (or 'ground truth') consisting of a manually labelled and unbiased sub-sample. The

[72] AJ Berinsky, MF Margolis and MW Sances, 'Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys' 58 (2014) American Journal of Political Science 739; S Zorowitz et al, 'Inattentive responding can induce spurious associations between task behaviour and symptom measures' 7 (2023) Nature Human Behaviour 1667.

[73] This is notably (but not exclusively) the case with the CJEU's early proportionality rulings, which refer to the need for measures to be 'reasonable', not 'out of proportion', or 'objectively justified by the need' to achieve a certain policy objective; see Zglinski (n 64) 127 et seq.

[74] The term 'legal basis' ('base juridique' in French) is frequently used by the CJEU when discussing the legality of EU sanctions. Most scholarship, however, associates it with the doctrine of the choice of Treaty articles enabling the making of EU legislation; see M Ovádek, 'Procedural Politics Revisited: Institutional Incentives and Jurisdictional Ambiguity in EU Competence Disputes' 59 (2021) Journal of Common Market Studies 1381.

[75] W Lowe and K Benoit, 'Validating estimates of latent traits from textual data using human judgment as a benchmark' 21 (2013) Political analysis 298.

objective is for the automated process to yield results that match as closely as possible the gold-standard coding. In general, the more complicated the automation, the more validation is needed to confirm the machine is not feeding us bogus results.[76]

As useful and recommended as automating some of the coding process might be, chances are your desired measure requires deeper interpretation or background knowledge to implement. Broadly speaking, the more complex the concept of interest, the more likely we might want to employ humans to do the coding.[77] There is no combination of keywords that would reliably tell us how much discretion the CJEU granted a national court in a preliminary ruling. Nevertheless, we underline the importance of thinking through the coding project before diving into it. In particular, we invite legal scholars to reflect on sources of variation in coding decisions. What about the data (e.g. paragraph of a CJEU decision) would make reasonable persons (e.g. EU law experts) disagree about the correct coding value (e.g. the amount of deference granted by the Court)? What aspects of the target concepts are likely to lead to discrepant coding? How do expertise, personal beliefs and contextual knowledge factor in the coding decisions and can they be equalised through a training session?[78]

Legal issue: When a public authority holds environmental information and is requested to disclose this information, must the public authority consider the grounds of refusal for releasing the information listed in Article 4(2) of Directive 2003/4 *cumulatively*, or must the authority weigh each ground of refusal *separately* against the public interests served by disclosure?	
Coder A	***Coder B***
Identified position: Where a public authority holds environmental information, it may, when weighing the public interests served by disclosure against the interests served by refusal to disclose, in order to assess a request for that information to be made available to a natural or legal person, take into account cumulatively a number of the grounds for refusal set out in that provision.	***Identified position:*** While the Directive does not require a cumulative weighing the Member States may perform such an exercise.
Identified code: would not restrict autonomy	***Identified code:*** competing effect on autonomy The position cannot be easily categorised as *would not restrict autonomy* or *would restrict*

[76] We recommend implementing some form of validation against gold-standard coding anytime automation is involved. Even seemingly innocuous tools such as keyword searches are susceptible to misspecification that produces erroneous results. No matter how convinced we may be that the word 'proportionality' (or a sophisticated machine learning algorithm) correctly identifies CJEU decisions applying the proportionality test, it is better to manually code a sub-sample of decisions to check that this is indeed so.

[77] Although the rise of large language models capable of identifying complex textual patterns has already started shifting efforts away from human coding. If the performance of language models continues to improve, artificial intelligence will provide a cost-effective alternative to human expert coding. Even then, the need for human validation is likely to persist, however.

[78] There are complex statistical models that try to address coders having different levels of expertise, but we do not deal with these here due to space constraints. See KL Marquardt and D Pemstein, 'IRT models for expert-coded panel data' 26 (2018) Political Analysis 431.

If this position would be the judgment, the autonomy of the Member State would not be more restricted than before.	*autonomy*, because it implies that autonomy would be restricted in one aspect and not restricted in another aspect, or because it would not affect the autonomy of the Member State.

Table 1: Example table of two coders disagreeing about national autonomy implications of a position voiced in preliminary reference proceedings at the CJEU

To illustrate, consider an example of coder disagreement from IUROPA's *Issues and Positions* dataset (Table 1). The dataset centres on the legal questions the CJEU was asked to resolve in preliminary reference proceedings lodged between 1995 and 2011, as well as the positions that EU institutions and Member States took on these questions. Based on written reports of the hearings held at the Court and the texts of CJEU judgments, coders were asked to identify the implied effects of actors' position on Member States' legislative and/or executive autonomy. In essence, the lead researchers of the project were interested in finding out which actors advocated for interpretations of EU law that would further restrict Member States' autonomy and which opposed such restrictions, along with several other categories of positions falling between these two poles.

The example given in Table 1 shows how two coders arrived at different conclusions based on their understanding of the coding instructions and their assessment of an actor's position (here, the European Commission) in light of these instructions. Although expressing the identified position in different terms, the meaning of the position is in essence the same for both coders. However, while one coder – correctly – thought the position means that Member States' autonomy would not be further restricted, the other identifies a more ambivalent, competing effect on autonomy. It is likely that Coder B arrived at their conclusion because the position makes no clear statement what a public authority must or must not do in a situation described by the legal issue, hence leading them to pick a coding category that makes no clear statement concerning the extent of autonomy restrictions. Examples like this show that adding clarifications to the codebook and offering training can help coders to discriminate between categories.

B. Coding in teams

Manual coding tasks should optimally be carried out by multiple persons. The reason for this practice being standard is that we want to have confidence that the results are representative of more than a single expert's opinion. In other words, given roughly the same context and level of expertise, would another coder come to the same conclusion? We want our coded variable to be *reliable* by which we mean that 'it produces the same results regardless of who or what is actually doing the measuring'.[79] When multiple coders generally agree on the labels attached to observations on a given variable, we say the variable has high *inter-coder reliability* or high *inter-coder agreement* (we expand on this below). This is a minimal standard by which to evaluate the reliability of the coding. More demandingly, we might ask: are the assigned labels representative of a broader expert consensus?

To clarify the various standards of inter-coder reliability, consider the following stylised example. A group of researchers at the European University Institute in Florence decides to assess the compatibility with EU law of a variety of unilateral Member State external actions.[80] There is a high degree of agreement among the researchers, and they find that on average 60 per cent of the

[79] Epstein and Martin (n 29) 52.

[80] Assume the unit of analysis are newspaper excerpts, each describing an international action by a Member State government.

evaluated actions infringe on EU prerogatives in external action. However, when a group of Danish law professors decides to reproduce the original measurement, they find that only around 10 per cent of the actions should be considered incompatible with EU law, despite using the same data and following meticulously the codebook and instructions prepared by the Florence team.

High inter-coder reliability does not by default mean that our measure is free of bias. In the illustrative example, both research teams were internally homogenous – but their different underlying beliefs about European integration and law systematically shaped the strictness with which they evaluated Member State actions.[81] Perhaps the true expert consensus is somewhere between the scoring arrived at by the two teams in which case they both failed to be representative in this regard. Although the mitigation measure against this scenario is fairly obvious – employ a heterogenous and sufficiently large group of coders – this may be hard to carry out in practice due to strong selection effects at both institutional and individual level. In addition, if coding is meant to be representative beyond the research team, we would need to define the population of relevant experts for our problem. That might seem a daunting task, but it should at least be possible for the researchers to survey the range of relevant schools of thought that could systematically affect coding.[82] More importantly, coding tasks should be defined in a way that minimises the relationship between expert judgment and personal beliefs. This can often be achieved by making assumptions explicit and coding instructions more specific. In the above example, the researchers would be better off specifying the content of the legal rule which the government actions are evaluated against and what standard is being applied to find an infringement. In general, when coders fill in the blanks in coding instructions, more bias seeps in and measurement reliability decreases.

The most important resource to manage when coding in teams is the coders' attention. It is critical to ensure that coders are focusing only on tasks where their attention has maximum added value. They should be shielded from any tasks that can be reliably automated. The coding task should be presented to them on a metaphorical silver platter to avoid attention being wasted on auxiliary chores. For example, if it can be collected in advance, all necessary information for the coding task should appear on their screen at the time of coding. The research lead needs to think carefully about minimizing decision-making friction and potential sources of time waste.

Central to the success of coding in teams is a good codebook. A codebook is a document or a set of documents explaining in the first place how codes relate to variables conceptually and in the second place the exact process of how codes were assigned to observations. Codebooks typically serve a dual purpose. Internally, they serve as a 'manual' for the researchers engaged in coding and allow to maintain consistency when working in teams. Externally, they explain the coding process, so that it can be replicated by other teams or help others decide whether the coded variables are useful for

[81] Note that this source of bias is different from 'groupthink'. The coding took place individually, without team members directly influencing each other. The problem in the example is that the results were systematically affected by the coders sharing an underlying set of beliefs which was not representative of experts' beliefs more broadly. Given the selection mechanisms at play at a place like the European University Institute, our stylised scenario is not completely unrealistic in the context of EU research.

[82] In other words, if a coder's belief system might exert a strong effect on coding decisions, this factor should be considered in the planning stage as part of the general assessment of the possible sources of variation. For example, would the fact that the coder is a convinced constitutional pluralist as opposed to European federalist affect the scoring outcomes? If the answer is yes and our research team only consists of federalists, the resulting data needs to come with the caveat that it is only replicable if the coders subscribe to the CJEU's supremacy doctrine.

their project. A good codebook will maximise both internal and external transparency regarding the purpose and process of the coding exercise.[83] In addition, while a 'coding system … inherently matches the context it is created in',[84] it is worth for the research lead to think in advance how other researchers might want to use the codes and codebook to maximise the impact of their work.[85]

Although there is no one recipe for creating a codebook, there are best practices that can help reduce costs and avoid low inter-coder reliability. We have three guiding principles in mind in particular: keep the coding scheme simple, start small and iterate. The point of keeping it simple is well-captured by Mikhaylov, Laver and Benoit:

> 'Coding schemes must balance the researcher's desire to reflect accurately the complexity of the reality represented by a text, with the practical requirements of keeping coding schemes simple enough that they can be implemented by human coders reliably.'[86]

Whenever a new variable or a new code (value of a variable) is proposed, we recommend testing out the scheme on a small subset of the dataset. In the first step, one might want to begin designing codes on a purposely selected subsample of the data. Indeed, it is common that a researcher has concrete examples in mind when they are proposing a certain code or variable. That is why it is important to subsequently test the scheme on a small random sample of the data, first by the lead researcher and then by at least one other team member. This iterative process of working with small samples has the advantage of quickly trying out different ideas and discarding unworkable ones without wasting too many resources. Few things are more disappointing than rolling out a grand coding scheme to multiple coders and letting them invest dozens of hours into coding only to realise that there is no agreement on the codes, either because the coding task was poorly designed, poorly described in the codebook or the coders received insufficient training. Good codebook-creating processes will seek to minimise the need for large-scale recoding.

As a general rule, the more comprehensive the explanation in the codebook, the better. However, when it comes to preparing the coders for the task, it is important to consider the amount of information coders can effectively digest. As codes are frequently illustrated with concrete examples from the data, the choice and number of examples provided for each code is particularly important. As a rule of thumb, it can be useful to give two examples: one that is clear-cut and one that is difficult (see Table 2). The clear-cut example serves as a kind of anchor – it provides a simple and tangible idea of the context in which the code is likely to appear. In contrast, the difficult example should explain how to adjudicate more complex observations where multiple codes are *prima facie*

[83] While agreement on the final design of concepts, variables and codes needs to be found at some stage in the research project, the broader codebook in which the coding processes are documented can be a 'living' document; see V Reyes, E Bogumil and LE Welch, 'The living codebook: Documenting the process of qualitative data analysis' 53 (2024) Sociological Methods & Research 89.

[84] S Bevan, 'Gone Fishing: The Creation of the Comparative Agendas Project Master Codebook', in FR Baumgartner, C Breunig and E Grossman (eds), *Comparative Policy Agendas: Theory, Tools, Data* (Oxford University Press 2019) 18.

[85] There are software tools that help with creating codebooks, such as the *codebook* and *codebookr* R packages.

[86] S Mikhaylov, M Laver and K Benoit, 'Coder reliability and misclassification in the human coding of party manifestos' 20 (2012) Political Analysis 78, 90.

applicable. If more than two examples are needed to illustrate codes, the total number should be proportional to the expected difficulty of coding the variable.

Variable	direct_effect
Description	This variable captures whether a CJEU paragraph states that a rule of EU law has direct effect.
Values	**Yes**: if the paragraph states that a rule of EU law has direct effect. **No**: if the paragraph states that a rule of EU law does not have direct effect. **Ambiguous**: if the paragraph addresses the question of a rule's direct effect but does not make it clear whether the rule has direct effect or not. **NA**: if the paragraph does not address the question of a rule's direct effect.
Instruction	Decide whether or not the CJEU establishes that a rule of EU law has direct effect.
Example 1 (easy)	It should be stated, in the second place, that neither Framework Decision 2002/584 nor Framework Decision 2008/909 has direct effect. That is because those framework decisions were adopted on the basis of the former third pillar of the European Union, in particular, under Article 34(2)(b) EU. **[No]**
Example 2 (difficult)	Whilst it is true, as observed by the Italian and Danish Governments, that a directive cannot of itself impose obligations on an individual and cannot therefore be relied on as such against an individual (see Case C-91/92 *Faccini Dori* [1994] ECR I-3325, paragraph 20), that case-law does not apply where non-compliance with Article 8 or Article 9 of Directive 83/189, which constitutes a substantial procedural defect, renders a technical regulation adopted in breach of either of those articles inapplicable. **[Yes]**

Table 2: Example of a codebook entry for the variable 'direct_effect'

Once a codebook is ready and coders have received sufficient training to accomplish the task, it is time for the big roll-out. Even if everything was tested and refined beforehand, it is crucial to have a monitoring system in place. Coders should submit their work to the lead researcher on a rolling basis and this should be checked relatively closely especially in the early stages of the exercise. No amount of testing will manage to anticipate all issues once the codebook meets the full force of the data. But it makes a big difference whether a major problem is discovered in the first or last week of coding.

Coding by several coders will inevitably lead to disagreement. Much of the above is intended to restrict coder disagreement to substantive points and keep it within acceptable bounds that the academic audience will find convincing. In line with our previous point about monitoring the coding process, it is advisable to check for reliability of the measurement continuously. In the words of Hayes and Krippendorff, we want to be 'evaluating whether common instructions to different observers of the same set of phenomena yields the same data within a tolerable margin of error'.[87]

When thinking about acceptable levels of inter-coder (dis)agreement, it is helpful to begin from a baseline. The simplest baseline is mere chance. If we were to randomly allocate codes to observations, the chance that any one observation is coded correctly is 1 / C where C stands for the number of possible codes (values) the variable in question can take. Say we are coding CJEU replies to preliminary questions, and we want to know whether the Court agrees with the preferred position of the national court. If the possible codes are (A) agrees; (D) disagrees; and (N) no preference expressed by the national court, then the probability that we assign the correct code by

[87] AF Hayes and K Krippendorff, 'Answering the Call for a Standard Reliability Measure for Coding Data' 1 (2007) Communication Methods and Measures 77, 78.

mere chance is 33%. However, this naïve probability does not take into account that code (N) is much more prevalent in the data than the other two possibilities. If instead of assigning the three codes randomly, we assign only code (N) to all observations, the baseline accuracy would be even higher, probably well over 50%.[88]

There are multiple ways to calculate inter-coder agreement. It might be tempting to simply look at the overlap between the labels assigned by two coders, but this approach does not account for agreement by chance. If two coders independently score 100 observations and agree on the label in 50 cases, the raw rate of agreement is 50%. If the coders were choosing from only two categories, this level of agreement is equivalent to scoring the observations randomly.[89] The more categories the coders are choosing from, the less likely it is they agree purely by chance. Either way, a good measure of inter-coder reliability will adjust the statistic for the influence of chance.

After reviewing the literature on reliability measures, the most straightforward recommendation we can make is to virtually always use Krippendorff's alpha to calculate inter-coder agreement.[90] Krippendorff's alpha is a general metric that applies to every coding setup. It denotes measurement reliability of a coded variable for any number of coders (at least two), any number of observations and for any type of variable (nominal, ordinal, interval or ratio). Without going into the mathematical details of the calculation, alpha is defined at the general level as $1 – D_o / D_e$ where D_o represents observed disagreement between coders and D_e represents disagreement expected by chance.[91] Usefully, functions for calculating Krippendorff's alpha exist in all popular statistical software packages.[92]

The values of alpha have an intuitive interpretation: the value of 1 indicates perfect reliability between coders, while 0 indicates complete absence of reliability.[93] In general, alpha values over 0.8 signal that the measurement (variable) is reliable. Krippendorff suggests, as a guideline, to consider variables with alpha between 0.8 and 0.667 only for drawing 'tentative conclusions' and to discard variables with reliability below this threshold.[94] To illustrate the values of alpha, imagine we employed three coders (c1, ... c3) to score ten preliminary rulings (r1, ... r10) according to whether the CJEU agreed (A) or disagreed (D) with the national court (or not applicable (N),[95] as above) and obtained the data matrix in Table 3.

[88] Or perhaps we are able to deploy a simple machine classification system (such as keyword matching) which produces even higher baseline accuracy. The baseline level of accuracy – whether it results from chance, indiscriminate assigning of dominant values or a simple computer program – gives us a starting point for assessing the acceptable level of disagreement among coders. The rate of agreement between coders should be significantly higher than this baseline for the coding exercise to be considered worthwhile.

[89] This only holds over a large enough sample. By the law of large numbers, the sample average converges to the expected value (mean) of the variable as the sample increases. A common way to think about this principle is to imagine we are assigning categories by flipping a coin. The ratio of the two categories being assigned will converge to approximately 1:1 (50% each) with enough coin flips.

[90] K Krippendorff, *Content Analysis: An Introduction to Its Methodology* (4th ed, Sage 2018) 280 et seq.

[91] The mathematical details are given in Krippendorff (n 85) 285.

[92] See, especially, the *icr* and *irr* packages in R and *krippendorff* in Python.

[93] Values of alpha below 0 indicate systematic disagreement between coders worse than chance.

[94] Krippendorff (n 85) 356.

[95] It is useful for the coding scheme to anticipate different kinds of missing, inapplicable or neutral values and make them explicit. In our example, a CJEU ruling might omit the necessary information from the national court, the national court's preference is expressed too vaguely to enable comparison with the final or it might

	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10
c1	A	A	D	N	D	D	A	A	D	A
c2	A	A	D	N	D	D	A	A	N	A
c3	A	D	D	N	D	N	A	A	N	D

Table 3: Example of a coding matrix. Each row is the work of a single coder (c1, c2, c3), each column is a court ruling (r1, … r10). 'A' stands for CJEU agreement with the national court, 'D' for disagreement and 'N' for not applicable.

The Krippendorff's alpha of this matrix is 0.596. The reliability value for the first five rulings (r1, … r5), where there is only one disagreement (r2), is 0.803. The third coder (c3) was more often in disagreement than the other two. Dropping 'c3' from the measurement would increase the alpha for the matrix to 0.843. This might look tempting but doing so ex post on the sole ground that c3's coding decreases the alpha could be considered fraudulent. Conversely, and less controversially, we could employ another coder to classify the same sample. If they ended up with the same codes as c2, the alpha for the entire data matrix would rise to 0.693, highlighting the greater consensus in the coding of the variable.

We stress that the purpose of coding in teams is to rigorously unearth expert consensus (or the lack thereof) on a given issue, not to validate the principal investigator's pet theories. There might be considerable temptation on the part of the team leaders to temper with the coding process when preliminary results suggest deviation from the expected direction. A carefully designed and theoretically important coding scheme might turn out to yield completely unreliable codes. The possibility of failure is intrinsic to empirical inquiry. Nonetheless, the research team has a responsibility to mitigate bias at every stage of the project. From making sure that coders deliver judgments independently from each other as well as from the team leaders (outside the training and small-scale testing phase) to avoiding cherry-picking and manipulating data, the ethics of empirical inquiry demand commitment to the procedural rules of the journey without knowledge of the substantive destination.

There are generally three modes of failure, defined as obtaining a low reliability coefficient on a coded variable: instructions, coders and concepts. Failure of instruction occurs when the coding instructions (including training) are insufficiently clear. Failure of coders happens when instructions are unproblematic, but one or more coders are failing to apply them correctly and consistently. Failure of concept is the only one of interest to the rest of the academic community and occurs when fundamentally experts disagree about the target concept.[96] These modes of failure are not mutually exclusive, so at any one time it might not be immediately obvious to the research team why inter-coder agreement is low. The only way to reliably find out is to vary each element while keeping the other two constant.

Assuming the research team navigated all the pitfalls of coding to a successful end (alpha > 0.8), there remains the question of what to do about the observations on which the coders did not agree.

be unclear whether the CJEU agreed with the national court's preference or not. We do not know a priori whether the differences between different types of 'third' values will be meaningful or not. The codes can be aggregated but not disaggregated later.

[96] The expert coding framework provides a robust methodology for re-evaluating many claims made by legal scholars. It would be of value to the academic community to know which popular concepts are reliably identifiable by expert coders.

If the number of coders is odd and there is a value chosen by the majority, then this should be the final value. In other instances, there needs to be a conciliation procedure leading to the most reasonable choice of final value. For the sake of transparency and replicability, the research team should publish not only the final coded variable but also data from all individual coders involved in the process. The latter enables running sensitivity analyses which look at the extent to which conclusions are affected by choosing an alternative value for disputed observations. Such sensitivity procedures can also be useful for salvaging insight from variables with below-par reliability (alpha of 0.6-0.8).

C. Coding alone

Despite the many virtues of coding in teams, much of empirical legal research, in EU law and elsewhere, is conducted alone. Partly, this is the result of resource constraints. Hiring research assistants is expensive and not every project will have the budget to explore this route. Partly, it can also be due to the nature of the research involved. PhD regulations often require that the research going into a dissertation has been done individually which may, in some universities, act as an obstacle to joint coding projects.

That coding alone is feasible and widespread should not be misunderstood as dispensing a researcher from following the basic principles outlined above. Quite the opposite: a greater awareness of potential pitfalls, and the implementation of adequate safeguards, is required. Just as with team efforts, the objective here is to produce coding which is reliable and representative of a broader expert consensus. That these conditions are fulfilled cannot be simply assumed, providing evidence to that effect is necessary. Two strategies can help in this context.

In the absence of being able to employ several coders, a researcher should, first, be even more transparent about how coding decisions were reached than when working in a team. The criteria used for determining whether a certain phrase, outcome, or event represents our variable(s) of interest need to be written down as comprehensively and precisely as possible. Likewise, all major difficulties encountered during the coding process (e.g. unclear or borderline cases), as well as how they were resolved (i.e. the final coding decision), must be documented. The information should be compiled in a separate document or added to the codebook. This will allow the researcher themselves to keep track of the coding and increase consistency. At the same time, it will enable the audience to reconstruct the decision-making process, thus increasing the trust in the results. The document can be attached to the publication; in the case of journal articles, it is common to add an online appendix to the main piece.[97]

Second, a researcher coding alone should seek independent feedback from peers. This can be done by asking colleagues to recode smaller sub-sets of the dataset.[98] The idea here is to get a second (and, possibly, third) pair of eyes to verify the coding that has been done. The motivation for this exercise is what was identified earlier as the key challenge for coding: would another person, given the same context and the same instructions, come to the same conclusion as the principal investigator? Inviting researchers to double check parts of the data is an imperfect alternative for coding in a team. However, it ultimately pursues the same goal, namely to assess, demonstrate and

[97] The transparency appendix attached to the qualitative study by Pavone and Stiansen is one example of best practice. See T Pavone and Ø Stiansen, 'The shadow effect of courts: Judicial review and the politics of preemptive reform' 116 (2022) American Political Science Review 322.

[98] See e.g. Pflücke (n 36).

improve the reliability of the findings. Inter-coder reliability scores along the lines of those described above can be calculated and reported here in a similar manner.

Coding 'solo' poses a series of unique challenges. Therefore, our general advice is to try to pursue empirical research projects, if possible, together with other scholars or assistants. Where this proves unfeasible for some reason, it is important to not only heed the general recommendations for expert coding but also, more broadly, to be mindful of the many sources of bias and variance that can affect human decision-making. To pick perhaps the most evident example, when working alone, the desire to corroborate the theoretical expectations articulated at the outset of a research project, the so-called confirmation bias, will have even more room to influence our coding. Consciously or unconsciously, the researcher may veer towards coding the data in a way that supports their hypotheses. Against this backdrop, recognising and striving to eliminate bias to the extent possible is essential. None of this, we should emphasise, is meant to disincentivise scholars from engaging in empirical projects. Sometimes working alone will be the only practicable way to conduct empirical legal research. In this situation, conducting research on one's own – while seriously considering the challenges connected with it – is, of course, preferable to the alternative, which is not conducting research at all.

5. Data presentation and analysis

At this point – after having clarified our concepts, collected the necessary data and coded them for our variables of interest – we are at an advanced stage of the project. But no project is complete without presenting and analysing the expert-coded variables. Data analysis is the process of evaluating data in light of our research objectives and theories and in line with an established analytical methodology, be it quantitative or qualitative. Data presentation and analysis are among the most immediately recognisable features of empirical legal research. They can include excerpts from interviews, figures containing trend lines, or the infamous regression tables with coefficients and p-values. In this section, we provide one concrete example of how to analyse expert coded variables to illustrate this step in the research process. Inevitably, there are many other qualitative and quantitative ways of analysing data. The choice of an analytical method is, just as all other steps in the process , contingent on the questions the research aims to answer.

In many research projects, summarizing and discussing the expert-coded variables will be a suitable end-result. In addition to reporting Krippendorff's alpha for each variable, the audience usually wants to know the proportion of assigned codes (labels) which should convey how prevalent the underlying concept or conceptual attribute is in the sample and by extension the population if the sample is representative. Such information is typically summarised in tables or visual figures with the goal of making it easy to understand for the reader. We are frequently also interested in how the values of one variable change across the different levels of another. For instance, if we have a variable capturing whether a decision conducts a proportionality analysis, we typically want to know how the share of proportionality decisions changes over time or differs across chamber and Grand Chamber rulings.

We exemplify data analysis in more depth using IUROPA's aforementioned *Issues and Positions* component. The dataset was compiled by a team of research associates who sourced information from written reports of individual hearings held at the CJEU as well as the text of CJEU judgments in preliminary reference procedures. The objective of their work was to identify the substance and relevant characteristics of the questions – referred to as legal issues – which the CJEU was asked to resolve in these procedures, along with information on the positions that Member States, EU institutions and other actors voiced on the issues. The dataset includes a variable that indicates whether a legal issue concerned the direct effect of EU law. Specifically, coders were asked to discern whether the question raised by a national court asked whether a rule of EU law fulfils the criteria of direct effect, coding the variable *Legal issue concerns direct effect* as 'yes' if that was the case and 'no' otherwise.[99] To discern whether IUROPA's codebook instructions lead to a reliable coding, a team of three coders scored the variable for 363 legal issues,[100] and a value of 0.80 for Krippendorff's alpha shows a high level of consensus among coders.

With access to this data, a researcher can easily find out that questions concerning the direct effect of EU law are a relatively rare occurrence. In only 161 of the 5,333 legal issues the CJEU considered in preliminary references lodged between 1995 and 2011 did a national court enquire about the direct effect of a rule of EU law. Further, a discussion of the descriptive statistics of a variable may reveal some interesting insights. For instance, IUROPA's *Issues and Positions* data shows that most preliminary references inquiring about the direct effect of EU law were submitted by German courts (29 references between 1995 and 2011), followed by British (25 references) and Austrian courts (17 references). Uncovering such patterns may prompt researchers to ask why preliminary references that concern the direct effect of EU law are predominantly submitted by courts situated in certain Member States, and thus inspire a new research project. Although in the following paragraphs we show how to run more sophisticated analyses, we stress that comprehensively describing the results (coded variables) from a well-executed expert coding project is in itself an important scholarly contribution.

A. Regression analysis

Aside from discussing the descriptive statistics of variables, the arguably most common use of larger datasets such as IUROPA's *Issues and Positions* component lies in regression analysis, the workhorse of quantitative scholarship in the social sciences. Regression analysis allows researchers to draw statistical inferences about the relationship between different variables from the sample data they are working with to the target population. Typically, researchers employing regression analysis are interested in explaining variation on a variable (e.g. the number of sources cited in a judgment text or whether a court shows deference to national authorities in a ruling), the so-called outcome variable. In the context of a regression analysis, this outcome variable is assumed to be a function of one or more so-called explanatory variables.[101] Regression analysis is so popular in the social sciences because it allows researchers to test their hypotheses, i.e. theoretically-grounded statements that explain a particular phenomenon (e.g. why a court tends to cite many sources in some judgments but not in others, or under which conditions a court would feel compelled to show deference to national authorities). However, we want to highlight here that regression analysis only allows researchers to draw inferences about the *causal* relationship between variables if certain conditions are met. Most notably, researchers must ensure that their findings are not affected by selection bias, a common problem in quantitative scholarship that we had briefly discussed above.[102]

[99] In addition, coders were able to opt for a category 'uncertain' if based on the information available to them, they were unable to code the legal issue as either 'yes' or 'no'.

[100] To save resources, each case was scored by two coders.

[101] In applications that seek to establish causal relationships between variables, the outcome variable is also commonly referred to as the dependent variable, and explanatory variables are referred to as independent variables.

[102] A more formal yet nonetheless accessible exposition of what selection bias is and how it can affect the results of our analyses is provided in Chapter 2 in JD Angrist and J-S Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton University Press 2009).

Yet, even without the aim of making statements about causal relationships, regression analysis can provide useful insights into statistical associations between variables of interest to the researcher.

To illustrate, imagine a team of researchers wondering why the Court of Justice manages to resolve some preliminary references in a relatively short period of time but on other occasions takes much longer to deliver an answer to a national court's questions. These researchers may hypothesise that questions concerning the direct effect of EU law are typically complex questions with significant political implications, and hence the variable *Duration of proceedings* at the Court of Justice (measured in the number of months it took the Court to resolve a preliminary reference) should be a function of the variable *Legal issue concerns direct effect* – in other words, if a question concerns EU law's direct effect, the Court should take longer to resolve a legal issue. Similarly, the researchers may assume that the number of observations the Court receives from Member States, the variable *Number of observers*, affect the duration of proceedings. After all, these observations need to be processed by the Court and judges are likely taking them into careful consideration when deliberating over their judgment.

Here, the duration of proceedings serves as the outcome variable, and the researchers hypothesised that variation on this variable can be explained by the two explanatory variables, *Legal issue concerns direct effect* and *Number of observers*. Notice that the outcome variable is measured on a ratio scale, i.e. the values on the variable can be ranked, they have equal distances between each other and there is a natural zero reference point for the scale – the Court cannot take less than zero months to deliberate over a case. We can estimate a linear regression model through a so-called ordinary-least-squares procedure – by far the most common analytical procedure in social sciences – to gauge the effect of our two explanatory variables on the outcome variable.[103]

In Figure 3, we plot the coefficient estimates for the two explanatory variables. These coefficient estimates in essence provide a simple summary of the explanatory variables' effect on the outcome variable. The positive coefficient estimate for the variable *Number of observers* indicates that the CJEU indeed takes longer to resolve a preliminary reference proceeding as the number of observing Member States increases (there is a positive effect). Further, Figure 3 shows that the coefficient's confidence interval – the bars on either side of the dot – do not overlap zero. Hence, the coefficient is distinguishable from zero and the researchers can be reasonably confident that we find this kind of relationship between the *Number of observers* and the *Duration of proceedings* not just in the sample data they are working with (i.e. preliminary references lodged between 1995 and 2011), but also in the target population of their study (i.e. every preliminary reference ever considered by the CJEU).[104]

[103] Although we note that other, more complex types of regressions are more appropriate to model a variable like *Duration of proceedings* which counts number of months it took the CJEU to complete a preliminary reference proceeding.

[104] Notice that due to data availability limitations – reports from hearing were largely discontinued after 2012 – researchers need rely on the assumption that the IUROPA sample is representative of CJEU decision-making for periods outside the sampling period if they want to claim their conclusions apply to preliminary references more generally.

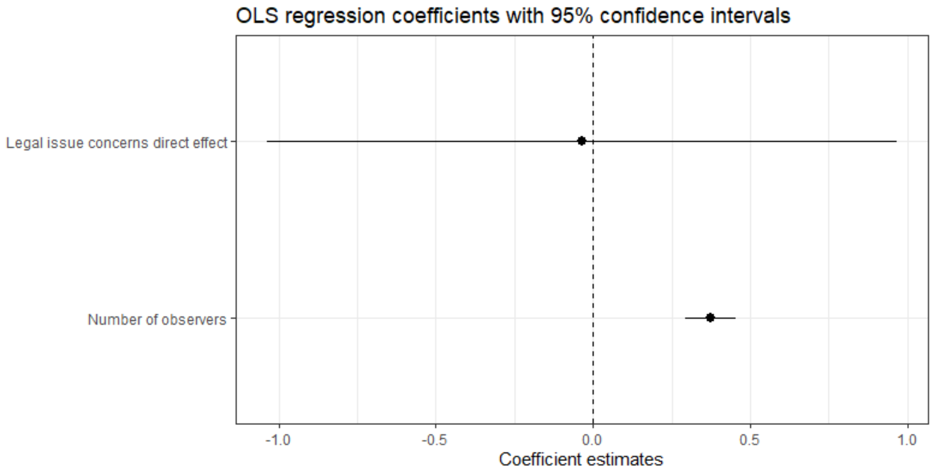

Figure 3: Ordinary least squares regression coefficient estimates with 95% confidence intervals. Outcome variable is the Duration of proceedings for preliminary references submitted to the CJEU between 1995 and 2011 (N = 5,327).

The same cannot be said for the explanatory variable *Legal issue concerns direct effect*. We can clearly see that the estimate falls very close to zero, while its confidence interval clearly overlaps with zero as well. In other words, based on the evidence from this sample there is no indication that the fact that a legal issue revolves around EU law's direct effect has any bearing on how long the CJEU takes to resolve a preliminary reference.

B. Qualitative uses

While information such as the data collected by IUROPA's team of researchers may seemingly be designed for scholars who employ statistical methods such as regression analysis in their studies, it can be of great use for projects that involve the use of qualitative methods as well. For instance, a researcher might be interested in learning how the CJEU reasons when the national court's question specifically concerns the direct effect of directives. Rather than having to scour through every judgment text the CJEU published within a certain timeframe, aided by a keyword search that hopefully picks up every instance of a consideration of a directive's direct effect, the researcher can instead rely on IUROPA's *Issues and Positions* data to get the desired list of cases. Along with the variable *Legal issue concerns direct effect*, IUROPA's research associates also recorded the types of legal acts the national court's question concerned, the variable *List of affected legal acts*. With this information in hand, a researcher can filter the data for legal issues that concerned EU law's direct effect (*Legal issue concerns direct effect* = 'yes') and includes the value 'Directive' in the *List of legal acts affected*. Doing so, the researcher will quickly find the 86 legal issues from the 72 preliminary references lodged between 1995 and 2011 that revolved around the direct effect of an EU directive. The researcher will have saved themselves plenty of time in searching for relevant cases.

A follow-up study might expand on or challenge the already-coded information in a dataset. A qualitative study might dive into the arguments of the litigants and the Court. Or perhaps the coding scheme missed important connections between different elements of a judgment that come to the fore only upon closer and more contextualised reading. Or a researcher might want to relate the already-coded information to documents from other sources, such as legislative debates. One might seek to, for example, understand the extent to which the EU legislator anticipated the CJEU (not) attributing direct effect to various EU measures. Such an analysis would require consulting a wide array of preparatory documents and discussions from the archives of the institutions involved in the making of EU law (the Commission, the Council, the Parliament). The point is that well executed expert coding is likely to produce data of value to more than just the immediate research project within which it took place.

6. Conclusion

This article demonstrates, drawing on examples from EU law, how to execute an expert coding project from start to finish. A successful expert coding project is a holistic endeavour that begins with good concepts and a reflection on how best to operationalise them, considers the population of cases it wants to draw conclusions about based on a carefully chosen sample and maintains a high level of transparency around coding decisions. We see expert coding as highly compatible with the bulk of legal research that already takes place, in the form of doctrinal analyses of individual or smaller sets of rulings. Unlike other empirical approaches, it therefore does not take much for legal scholars to adopt this methodology. At the same time, the framework offered by expert coding and the related conceptual and sampling principles have the potential to significantly boost the impact of legal research by making its findings more reliable and compatible with research in adjacent social science disciplines. A well-executed expert coding project can retain many of the advantages of doctrinal research, notably understanding relevant (case) law and parsing legal language, while remedying the limitations plaguing traditional case commentaries, such as selection bias and reliability.

By emphasizing the potential of expert coding for legal research we want to bridge the gap between empirical and doctrinal scholarship. There is a widespread perception that empirical and doctrinal methods stand in an irresolvable conflict with each other. This perception is, to an important extent, rooted in criticisms of traditional doctrinal scholars, who dismiss empirical approaches as a tool for research into legal phenomena, arguing that they fail to appreciate the unique nature of law. But it is also fuelled by those empirical scholars who present doctrinal approaches as outdated, and the knowledge generated through them useless. We see both positions as unproductively extreme. Legal, including doctrinal, knowledge is important and analyses shirking the legal dimension of social phenomena are bound to be incomplete. At the same time, there is no convincing reason why legal scholarship should stand apart from the methodological paradigms of social science. Expert coding offers a practicable solution for squaring this circle.