2024

The Scales Project: Making Federal Court Records Free

Christopher A. Cotropia

 Printed in U.S.A.
Vol. 119, No. 1

Essays

THE SCALES PROJECT: MAKING FEDERAL COURT RECORDS FREE

David L. Schwartz, Kat M. Albrecht, Adam R. Pah, Christopher A. Cotropia, Amy Kristin Sanders, Sarath Sanga, Charlotte S. Alexander, Luís A.N. Amaral, Zachary D. Clopton, Anne M. Tucker, Thomas W. Gaylord, Scott G. Daniel & Nathan Dahlberg

ABSTRACT—Federal court records have been available online for nearly a quarter century, yet they remain frustratingly inaccessible to the public. This is due to two primary barriers: (1) the federal government's prohibitively high fees to access the records at scale and (2) the unwieldy state of the records themselves, which are mostly text documents scattered across numerous systems. Official datasets produced by the judiciary, as well as third-party data collection efforts, are incomplete, inaccurate, and similarly inaccessible to the public. The result is a de facto data blackout that leaves an entire branch of the federal government shielded from empirical scrutiny.

In this Essay, we introduce the SCALES project: a new data-gathering and data-organizing initiative to right this wrong. SCALES is an online platform that we built to assemble federal court records, systematically organize them and extract key information, and—most importantly—make them freely available to the public. The database currently covers all federal cases initiated in 2016 and 2017, and we intend to expand this coverage to all years. This Essay explains the shortcomings of existing systems (such as the federal government's PACER platform), how we built SCALES to overcome these inadequacies, and how anyone can use SCALES to empirically analyze the operations of the federal courts. We offer a series of exploratory findings to showcase the depth and breadth of the SCALES platform. Our goal is for SCALES to serve as a public resource where practitioners, policymakers, and scholars can conduct empirical legal research and improve the operations of the federal courts. For more information, visit www.scales-okn.org.

AUTHORS—The authors would like to thank the Symposium organizers and editors of the *Northwestern University Law Review* for their tireless work and support. David Schwartz is the Frederic P. Vose Professor of Law at Northwestern Pritzker School of Law. Kat Albrecht is an Assistant Professor of Criminal Justice and Criminology at Georgia State University. Adam Pah is a Policy Analytics Lead and Clinical Associate Professor at Georgia State University. Christopher Cotropia is the Dennis I. Belcher Professor of Law at University of Richmond School of Law. Amy Kristin Sanders is an Associate Professor of Journalism and Media at University of Texas at Austin. Sarath Sanga is a Professor of Law at Yale Law School and the corresponding author (e-mail: sarath.sanga@yale.edu). Charlotte Alexander is a Professor of Law and Ethics at Georgia Institute of Technology. Luís Amaral is the Erastus Otis Haven Professor of Chemical and Biology Engineering at Northwestern University. Zachary Clopton is a Professor of Law at Northwestern Pritzker School of Law. Anne Tucker is a Professor of Law at Georgia State University College of Law. Thomas Gaylord is an Associate Law Librarian for Scholarly Communications at Northwestern Pritzker School of Law. Scott Daniel is a software engineer. Nathan Dahlberg is a data scientist.

INTRODUCTION

The federal judiciary charges $0.10 per page to view PDFs of court documents online.[1] Accessing a single case can cost $10 or more, while accessing all cases in a given year could easily exceed millions of dollars.[2] Consequently, we—lawyers, scholars, journalists, and citizens—lack the means to empirically and systematically evaluate the judiciary. Although one can purchase an individual case, that only allows one to examine individual episodes of justice. Without access to all records, there is no way to analyze the operations of the system as a whole or search for patterns, biases, or inefficiencies across cases, litigants, judges, and districts.[3] Without comprehensive access to the records, it becomes impossible to rigorously measure even the most fundamental aspects of the judicial system, such as the average duration of cases or the share of cases that proceed to trial.

The argument for increasing access to judicial records is clear and compelling. Democracies do not work without public access to government records.[4] Indeed, in a well-functioning democracy, the government should do more than merely *allow* access; it should affirmatively *enable* the public to access government records and scrutinize its operations. The other two branches of the federal government—Congress and the Executive—both do this by providing the public with extensive and free online access to records in many domains, including congressional voting records, regulatory notice-and-comment rulemaking, public company disclosures, and more.[5] The federal judiciary stands in stark contrast to this. It openly acknowledges and even valorizes the principle of public access to court *proceedings*—while at the same time all but foreclosing meaningful access to court *records*.[6]

---

[1] *Find a Case (PACER)*, U.S. CTS., https://www.uscourts.gov/court-records/find-case-pacer [https://perma.cc/W34J-Z9J5].

[2] Adam R. Pah, David L. Schwartz, Sarath Sanga, Zachary D. Clopton, Peter DiCola, Rachel Davis Mersey, Charlotte S. Alexander, Kristian J. Hammond & Luís A. Nunes Amaral, *How to Build a More Open Justice System*, 369 SCIENCE 134 (2020).

[3] *See* Fred H. Cate, D. Annette Fields & James K. McBain, *The Right to Privacy and the Public's Right to Know: The "Central Purpose" of the Freedom of Information Act*, 46 ADMIN. L. REV. 41, 42–43, 65 (1994) (using the Freedom of Information Act (FOIA) to show the importance of access to government documents to evaluate the conduct of government officials).

[4] *Id.* at 42.

[5] Some forms of public records are particularly designed for this purpose, like FOIA, whose "first and most important" goal is to "ensure public access to the information necessary to evaluate the conduct of government officials." *Id.* at 65; *see also How to Find Congressional Votes*, U.S. SENATE, https://www.senate.gov/legislative/HowTo/how_to_votes.htm [https://perma.cc/3ES5-M32A]; *Rulemaking Process*, FCC, https://www.fcc.gov/about-fcc/rulemaking-process [https://perma.cc/7Y9A-A2BK]; *Exchange Act Reporting and Registration*, SEC, http://www.sec.gov/education/smallbusiness/goingpublic/exchangeactreporting [https://perma.cc/S7LC-4W7N].

[6] *See infra* Section I.A.

Enter the SCALES project. The mission of SCALES is to make court records freely accessible to the public.[7] To do this, we established an online data repository for court records, augmented with a set of powerful AI-backed tools to enable the public to analyze the operations of the federal judiciary.[8]

The development of SCALES was a multi-step and multi-year project. We began by acquiring unprocessed (or "raw") civil and criminal litigation data. The raw data is a collection of loosely connected documents. Engaging with them is like reading scattered pages from multiple books with missing or mislabeled chapters and headings. Even basic information from the raw data—such as parties, lawyers, law firms, and judges—are rife with inconsistencies.

Our vision was to not just clean the data, but to create a platform where the narrative of federal litigation—both at the individual case level and throughout the system as a whole—is clear and interpretable by legal experts and nonexperts alike. To achieve this vision, we implemented several processing steps. We constructed an ontology of litigation events: complaints, motions, decisions, and other outcomes or filings that define a lawsuit's lifecycle, whether a civil or criminal matter. We used this ontology to train AI models to recognize and categorize litigation events and apply standardized labels to the raw docket entries. These labels serve as a higher-level architecture that enables users to efficiently search and analyze court data. The SCALES platform now provides a complete pipeline to take unstructured court records, automatically organize and index them, disambiguate names and entities, and apply the labels of litigation ontologies. The database currently covers all federal cases initiated in 2016 and 2017, and we intend to expand this coverage to all years.

In this Essay, we introduce the SCALES project and present new foundational descriptive statistics about the federal courts. This Essay is part of *Northwestern University Law Review*'s Symposium *Data Justice: How Innovative Data Is Transforming the Law*. We hope that this Symposium Issue—in addition to this Essay and the dataset it introduces—will inspire new avenues of empirical legal research and enhance scholarly and public engagement with the federal courts.[9]

---

[7] SCALES stands for "Systematic Content Analysis of Litigation EventS."

[8] To visit SCALES, go to: https://scales-okn.org [https://perma.cc/7DZ5-SJ7B].

[9] There has been a recent uptick in scholarly research using court records at scale, in part thanks to the rise in computing power and digital data available. *See, e.g.*, Maria-Veronica Ciocanel, Chad M. Topaz, Rebecca Santorella, Shilad Sen, Christian Michael Smith & Adam Hufstetler, *JUSTFAIR: Judicial System Transparency Through Federal Archive Inferred Records*, 15 PLOS ONE 1, 6 (2020) (introducing

We begin in Part I with a foundational claim that motivates the SCALES project: Increasing access to court records will increase public trust and confidence in the judiciary. In Part II, we survey the limitations of extant sources of court data and analytical tools. Part III introduces the SCALES Open Knowledge Network and describes how SCALES makes court data accessible. We provide a snapshot of the types of data insights available through the SCALES Data Explorer, as well as a case study on access to justice. We close the Essay with a brief conclusion and a call to action.

I. USING DATA TO BUILD PUBLIC TRUST IN THE JUDICIARY

A. The Judiciary's Theory of Transparency

Courts believe in transparency. But when they engage with the principle of transparency, they typically operationalize it as an individual's ability to exercise *direct* oversight, such as by personally attending court proceedings. The jurisprudence around this public right to attend trials is where we find the judiciary's most compelling arguments for judicial transparency. Yet these very same arguments, in our view, also make the case for why individuals must additionally have the right to freely access public court records.

The public right to attend trials was articulated in the 1980 Supreme Court case of *Richmond Newspapers, Inc. v. Virginia*.[10] There, Richmond Newspapers argued that a judge's decision to close a murder trial to the public and press violated the First Amendment.[11] The Court drew upon the historical precedent of open trials in this country's criminal justice system to justify its decision. Chief Justice Warren Burger, on behalf of a plurality of the Court, wrote:

> The crucial prophylactic aspects of the administration of justice cannot function in the dark; no community catharsis can occur if justice is "done in a corner [or] in any covert manner."[12]

JUSTFAIR or the Judicial System Transparency through Federal Archive Inferred Records, a large scale, crosswalked, free public database of 600,000 records); Daniel Martin Katz & M.J. Bommarito II, *Measuring the Complexity of the Law: The United States Code*, 22 A.I. & L. 337, 344–45 (2014) (offering a new framework for measuring legal complexity); Michael Evans, Wayne McIntosh, Jimmy Lin & Cynthia Cates, *Recounting the Courts? Applying Automated Content Analysis to Enhance Empirical Legal Research*, 4 J. EMPIRICAL LEGAL STUD. 1007, 1024–27 (2007) (testing various text classification models for determining content in Supreme Court advocacy briefs).

[10] 448 U.S. 555, 580, 581 (1980).

[11] *Id.* at 563–64.

[12] *Id.* at 571 (citing Concessions and Agreements of West New Jersey (1677), *reprinted in* SOURCES OF OUR LIBERTIES 188 (Richard L. Perry ed., 1959)).

He went on to further identify transparency as the source of the public's trust in the judiciary:

> A result considered untoward may undermine public confidence, and where the trial has been concealed from public view an unexpected outcome can cause a reaction that the system at best has failed and at worst has been corrupted. To work effectively, it is important that society's criminal process "satisfy the appearance of justice," and the appearance of justice can best be provided by allowing people to observe it.[13]

Throughout the 1980s, the Court applied similar logic to other parts of the judicial process. In *Press-Enterprise Co. v. Superior Court*, the Court returned to historical accounts indicating that jury selection was presumptively public to hold in favor of a public right to attend voir dire.[14] Chief Justice Burger, this time writing for the majority, again argued that an open process benefits the public:

> The value of openness lies in the fact that people not actually attending trials can have confidence that standards of fairness are being observed; the sure knowledge that *anyone* is free to attend gives assurance that established procedures are being followed and that deviations will become known. Openness thus enhances both the basic fairness of the criminal trial and the appearance of fairness so essential to public confidence in the system.[15]

Chief Justice Burger again upheld the public's right of access in 1986 in *Press-Enterprise II*, which held that the First Amendment guarantees the public a presumptive right to attend pretrial hearings.[16]

The Court's theory of its own legitimacy is premised on the public's ability to directly observe judicial processes. These and other Supreme Court cases gave the press—through the rights granted to the public—an essential presumption of access to nearly all court proceedings.[17] The Court's theory of open access to judicial process is also the basis of its corollary theory of the legitimacy of *closed* proceedings. The Court has repeatedly emphasized that closure is appropriate only in the narrowest of circumstances, such as situations in which open proceedings might infringe upon a criminal defendant's right to a fair trial.[18]

---

[13] *Id.* at 571–72 (citing Offutt v. United States, 348 U.S. 11, 14 (1954)).

[14] 464 U.S. 501, 505, 511 (1984).

[15] *Id.* at 508.

[16] Press-Enterprise Co. v. Superior Ct. (*Press-Enterprise II*), 478 U.S. 1, 13 (1986).

[17] *See, e.g.*, Presley v. Georgia, 558 U.S. 209, 211–12 (2010) (detailing the public's right of access to nearly all court proceedings).

[18] *See* Globe Newspaper Co. v. Superior Ct., 457 U.S. 596, 606–07, 610–11 (1982) (holding unconstitutional a Massachusetts statute that automatically closed rape trials during the testimony of minor victims); *see also Press-Enterprise Co.*, 464 U.S. at 510.

The Court's jurisprudence on transparency and its own legitimacy is fundamentally concerned with court *proceedings*. But what about the *records* of those proceedings? Here, the judiciary has been less enthusiastic about openness and transparency. Access to court records, it seems, is not foundational to public trust and confidence in the courts. The jurisprudence on this is defined by the exceptions to the presumption that courts are not required to provide affirmative access to their records. In the Ninth Circuit, for example, the 1983 case of *Associated Press v. United States District Court* affirmed that courts are required under the First Amendment to provide pretrial records.[19] But this has not extended to a universal ideal of public access to all public court records.[20] Courts have instead retained discretion over their own record management, with power to redact or limit access to large swaths of court records.[21]

Meanwhile, Congress has shown moderate interest in improving access to court records, as evidenced by legislation such as the E-Government Act, which aims to enhance the management and promotion of electronic government services and processes.[22] But this is not nearly enough, and congressional oversight on open access to court records has been notably minimal. Furthermore, despite clear legislative intent to increase public access to government information—including court records—the actual implementation and enforcement of these laws have not fully realized the potential for widespread access. Most promisingly, the House passed the Open Courts Act in 2020, which would have eliminated fees for access to federal court records, but the Senate failed to act.[23] The Act has not been brought back for a vote in the years since.[24]

Congress's efforts at fostering greater access stands in stark contrast with the judiciary's approach. This legislative push towards transparency

---

[19] Associated Press v. U.S. Dist. Ct., 705 F.2d 1143, 1145 (9th Cir. 1983).

[20] Federal courts publish and make publicly available some decisions. However, even decisions which rule on substantive motions are not universally available. *See* Christina L. Boyd, Pauline T. Kim & Margo Schlanger, *Mapping the Iceberg: The Impact of Data Sources on the Study of District Courts,* 17 J. EMPIRICAL LEGAL STUD. 466, 467–69 (2020) ("Which district court opinions are published in the *Federal Supplement* or *Federal Rules Decisions* involves an additional nonrandom selection process.").

[21] Ronald D. May, *Public Access to Civil Court Records: A Common Law Approach*, 39 VAND. L. REV. 1465, 1469 (1986); JAMES M. CHADWICK, ACCESS TO ELECTRONIC COURT RECORDS: AN OUTLINE OF ISSUES AND LEGAL ANALYSIS 1 (2001), https://bja.ojp.gov/sites/g/files/xyckuh186/files/media/document/legal-issues.pdf [https://perma.cc/72EU-FL3B].

[22] E-Government Act of 2002, 44 U.S.C. §§ 3601–3616.

[23] Sarath Sanga & David Schwartz, Opinion, *Tear Down This Judicial Paywall*, WALL ST. J. (Dec. 13, 2020, 6:00 PM), https://www.wsj.com/articles/tear-down-this-judicial-paywall-11607900423 [https://perma.cc/WS7Z-HXXD].

[24] *The Courts and Congress—Annual Report 2022*, U.S. CTS., https://www.uscourts.gov/statistics-reports/courts-and-congress-annual-report-2022 [https://perma.cc/9332-UM39]; *see also* Tanina Rostain, *Access to Justice as Access to Data*, 119 NW. U. L. REV. 5, 18–19 (2024).

underscores the gap between the ideal of open access championed by one branch of government and the guarded stance of another. The Court's adamant protection of the right to attend court proceedings is in considerable tension with its claim that the public does not have a right to access free records of those proceedings.[25] Chief Justice Burger's arguments, quoted at length above, should apply with equal force to both.

The federal courts' PACER system charges users $0.10 per page to view images of court records online.[26] Imagine if the federal courts also charged $0.10 per minute to attend any public judicial proceeding. Even setting aside the exclusionary effect of such a policy, its expressive effect alone is already repulsive to democratic norms. Financial barriers to transparency such as PACER's paywall undermine the public's trust and confidence in the judiciary.

B. Why Court Records Must Be Free

The inability to freely access public court records constitutes a profound impediment to understanding and improving the judiciary. It prevents researchers, journalists, and the public from studying and uncovering insights about the courts—insights that could in turn improve the administration of justice.[27]

While PACER is not the only way to access federal court records, other methods have similar and significant downsides that inhibit large-scale, systematic analysis of the judiciary. Commercial legal databases such as Westlaw, LexisNexis, and Bloomberg require expensive subscriptions that

---

[25] PACER purports to provide free access to judicial opinions, but even that appears to be vastly incomplete. *See* Peter W. Martin, *District Court Opinions that Remain Hidden Despite a Long-Standing Congressional Mandate of Transparency—The Result of Judicial Autonomy and Systemic Indifference*, 110 L. LIBR. J. 305, 319 (2018).

[26] For a discussion of the limitations of PACER, see *infra* Part II. *See also Find a Case (PACER)*, *supra* note 1.

[27] DAME HAZEL GENN, MARTIN PARTINGTON & SALLEY WHEELER, NUFFIELD INQ. ON EMPIRICAL LEGAL RSCH., LAW IN THE REAL WORLD: IMPROVING OUR UNDERSTANDING OF HOW LAW WORKS 1 (Nov. 2006), https://www.nuffieldfoundation.org/wp-content/uploads/2019/12/Law-in-the-Real-World-full-report.pdf [https://perma.cc/J6CP-MRKD]; *see also* Charlotte S. Alexander & Lauren Sudeall, *Creating a People-First Court Data Framework*, 58 HARV. C.R.-C.L.L. REV. 731, 739 (2023) ("As legal empiricists Kevin Clermont and Ted Eisenberg have argued, the paucity and limitations of existing court data 'restrict[] what one can study about the legal system, and surely make[] risky any behavioral inferences one might draw therefrom.' Legal scholar Lynn LoPucki has observed further, 'By offering selective access to data, the courts have controlled legal scholars' research agendas, . . . discouraging research that focused on the actions of judges and the impacts of those actions on both litigants and the public.'" (alterations in original)); *see also* Kevin M. Clermont & Theodore Eisenberg, *Litigation Realities*, 88 CORNELL L. REV. 119, 129 (2002); Lynn M. LoPucki, *The Politics of Research Access to Federal Court Data*, 80 TEX. L. REV. 2161, 2162, 2171 (2002).

most news organizations and citizens cannot afford.[28] Additionally, not all commercial services have comprehensive records, and they generally prohibit users from bulk downloading records.[29] This severely limits the utility of such services for large-scale empirical research. Journalists and researchers occasionally can get fee waivers or reductions, but this process is cumbersome, limited in time, ad hoc, and often unsuccessful.[30] Moreover, fee waiver arrangements typically prevent the public release of any underlying records that were obtained to conduct the analysis—thereby preventing replication or follow-up studies.[31]

Government transparency advocates have long called for free and open access to all public records because access to public records is the foundation of trust and confidence in the government.[32] As we and countless others have argued, the reasons are self-evident. An active and engaged citizenry deters government officials, including judges, from unethical and illegal

[28] For example, a one-year, individual subscription to Westlaw is $4,012.80. *See Westlaw Edge Plans and Pricing*, THOMSON REUTERS, https://sales.legalsolutions.thomsonreuters.com/en-us/products/westlaw-edge/plans-pricing [https://perma.cc/J2HQ-LS6A].

[29] *Empirical Legal Research Resources*, STAN. L. SCH., https://guides.law.stanford.edu/c.php?g=685018&p=5822311 [https://perma.cc/VJA8-FQ3K].

[30] *See* Stephen J. Schultze, *The Price of Ignorance: The Constitutional Cost of Fees for Access to Electronic Public Court Records*, 106 GEO. L.J. 1197, 1212, 1213, 1215–16 (2018).

[31] *See Electronic Public Access Fee Schedule*, U.S. CTS. (Dec. 31, 2019), https://www.uscourts.gov/services-forms/fees/electronic-public-access-fee-schedule [https://perma.cc/Y4SX-Y9XA].

[32] The literature has shown that oversight affects judicial conduct. For example, Professors Lim, Snyder, and Strömberg studied newspaper coverage of judicial behavior, finding that as coverage of a particular judge increased, the judge—assuming they were selected in a nonpartisan election—was more likely to increase the length of the criminal sentences they imposed, in part because of pressure from the public to avoid lenient sentences. *See* Claire S.H. Lim, James M. Snyder Jr. & David Strömberg, *The Judge, the Politician, and the Press: Newspaper Coverage and Criminal Sentencing Across Electoral Systems*, 7 AM. ECON. J.: APPLIED ECON. 103, 104, 129, 133 (2015).

activities.[33] Without free and open access to public court records, engaged citizens cannot exercise proper oversight over the judiciary.[34]

II. LIMITATIONS OF EXISTING DATA SOURCES

In this Part, we review the currently available types of data search and analytic tools providing access to court data. We also present a brief history of advances in general data search practices.

A. A Brief History of Court Records

Early legal information technology focused on digitizing case law and statutes, as well as providing basic search capabilities.[35] These early databases digitized primary legal sources into plain text. They were also proprietary, meaning the databases themselves were secured behind access controls and not available for bulk download. Instead, the databases were accessible to those who paid for access—primarily libraries and law firms.[36] Those who could not afford to pay per search had to seek other options, perhaps resorting to traditional library research using published search indexes and reporters.

---

[33] Journalists' coverage of federal judges likely spurred the Court to pass its code of ethics. *See, e.g.*, Michael Siconolfi, Coulter Jones, Joe Palazzolo & James V. Grimaldi, *Dozens of Federal Judges Had Financial Conflicts: What You Need to Know*, WALL ST. J. (Apr. 27, 2022, 7:30 PM), https://www.wsj.com/articles/dozens-of-federal-judges-broke-the-law-on-conflicts-what-you-need-to-know-11632922140 [https://perma.cc/CX9S-5VHD]; Jodi Kantor & Jo Becker, *Former Anti-Abortion Leader Alleges Another Supreme Court Breach*, N.Y. TIMES (Nov. 19, 2022), https://www.nytimes.com/2022/11/19/us/supreme-court-leak-abortion-roe-wade.html [https://perma.cc/2ZMQ-2BJE]; Andrew Perez, Andy Kroll & Justin Elliott, *How a Secretive Billionaire Handed His Fortune to the Architect of the Right-Wing Takeover of the Courts*, PROPUBLICA (Aug. 22, 2022, 2:45 PM), https://www.propublica.org/article/dark-money-leonard-leo-barre-seid [https://perma.cc/LK3D-FWA2]; Stephen Engelberg & Jesse Eisinger, *The Origins of Our Investigation into Clarence Thomas' Relationship with Harlan Crow*, PROPUBLICA (May 11, 2023), https://www.propublica.org/article/clarence-thomas-harlan-crow-investigation-origins [https://perma.cc/J8QH-7EKK].

[34] *Justice System Reform: Advancing Fairness and Efficiency*, GRAY GRP. INT'L (Mar. 25, 2024), https://www.graygroupintl.com/blog/justice-system-reform [https://perma.cc/ZLH8-B9GZ] ("Data-driven approaches are indispensable in driving effective reform initiatives. By collecting and analyzing relevant data, we can identify trends, predict challenges, and measure the impact of implemented reforms. This empirical evidence empowers policymakers and stakeholders to make informed decisions and adapt strategies when necessary.").

[35] *See, e.g.*, Bill Voedisch, *Westlaw: An Early History*, LEGAL PUBL'G 1, 13–14, 19–20 (2015) (describing Westlaw offering full-text case searching in the late 1970s).

[36] *See, e.g.*, William G. Harrington, *A Brief History of Computer-Assisted Legal Research*, 77 L. LIBR. J. 543, 553, 555 (1984) (discussing Lexis's targeting of New York law firms to increase subscriptions).

In the 1970s and 1980s, prominent services such as Lexis and Westlaw dominated the computer-aided legal research landscape.[37] Search capabilities for these databases were primarily Boolean in nature, with limited searching of case-digest and headnote information.[38] Over time, these databases began integrating citation information and accompanying functionality. Lexis incorporated its *Shepard's* service into its platform, while Westlaw introduced its KeyCite service in the 1990s (after having previously licensed *Shepard's* from Lexis).[39] These services provided auxiliary information and tracking of judicial treatment, citing references, and historical context for the legal sources in their databases.[40] The databases focused on providing search access to published cases and statutes; apart from citation analysis, they generally lacked analytical abilities.[41]

As Westlaw and Lexis gained prominence in the legal technology space, there was a growing interest in the 1990s in electronic records for federal and state courts. The United States Congress passed legislation instructing the judiciary to implement electronic filing and provide digital access to litigation information.[42] This led the Administrative Office of the U.S. Courts to launch the Public Access to Court Electronic Records (PACER) system.[43] Initially, PACER mainly provided attorneys (particularly government attorneys) access to docket information for federal courts. It also facilitated the electronic filing of pleadings and other court documents by

---

[37] *See id.* at 553–55 (detailing the rise of Lexis and Westlaw); *see also* Robert J. Munro, J.A. Bolanos & Jon May, *LEXIS vs. WESTLAW: An Analysis of Automated Education*, 71 L. LIBR. J. 471, 475 (1978).

[38] *See* Voedisch, *supra* note 35, at 5–7, 13–14 (explaining Westlaw's search capabilities).

[39] *See* James A. Sprowl, *The Latest on Westlaw, Lexis and Dialog*, 70 A.B.A. J. 85, 90 (1984) (noting that the *Shepard*'s citation service was available on both Westlaw and Lexis search platforms); *see also* Paul Hellyer, *Evaluating Shepard's, KeyCite, and BCite for Case Validation Accuracy*, 110 L. LIBR. J. 449, 450 (2018).

[40] *See* Elizabeth M. McKenzie, *New Kid on the Block: KeyCite Compared to Shepard's*, 3 AALL SPECTRUM 8, 8–9 (1998).

[41] Prior to the introduction of KeyCite and *Shepard's*, citation analysis capabilities were also limited to the citing and cited case or statute, along with a rough indication of the cited case's treatment. *See generally* Sprowl, *supra* note 39 (detailing the capabilities of Lexis and Westlaw to provide citation analysis).

[42] Judiciary Appropriations Act of 1991, Pub. L. No. 101-515, § 404(a), 104 Stat. 2101, 2132–33 (1990); *Electronic Public Access Fee Schedule*, U.S. CTS. (Jan. 1, 2020), https://www.uscourts.gov/services-forms/fees/electronic-public-access-fee-schedule [https://perma.cc/NX3V-3NFT]; *see Federal Courts Turn a New Page: Case Management/Electronic Case Files Systems Bring Greater Efficiency/Access*, 35 THIRD BRANCH (Admin. Off. of the U.S. Cts., Washington, D.C.), Nov. 2003, at 11, https://web.archive.org/web/20100412112709/http://www.uscourts.gov/ttb/nov03ttb/page.

[43] *See* Peter W. Martin, *Online Access to Court Records—From Documents to Data, Particulars to Patterns*, 53 VILL. L. REV. 855, 860–65 (2008) (detailing the origins of PACER).

litigants.[44] The system's initial rollout was localized to individual courts.[45] This meant attorneys needed to conduct their searches at the district or circuit court level.[46] Each federal court had its own PACER database.[47]

Beyond these limitations, the search capabilities of PACER were also rudimentary.[48] Users could search by docket number or party name, but broader searches, including keyword searches, have never been available.[49] In its early stages, the platform also only provided docket *information* on cases—the title of documents filed, by whom they were filed, and on what date—it lacked access to the underlying filed documents or opinions themselves.[50]

With time, PACER connected all localized docket databases, enabling users to search for cases filed nationwide from a single login location.[51] PACER also began to link and provide access to the underlying documents associated with the docket entries.[52] These improvements, while necessary, were an insufficient response to the underlying access problems.

Despite being a government-sponsored platform, PACER operates similarly to proprietary commercial case law and statutory databases. The underlying dataset (the docket information) is not publicly downloadable. Users instead access federal court information on an individual-search basis.[53] To get a broader picture of the judiciary, a user would have to piece together this single-search information one case at a time. Yet PACER's fee structure makes this approach infeasible, as users have always had to pay on a per-page basis to access and search through dockets.[54] They have also, with

---

[44] *See id.* at 860–61 ("A large fraction of [PACER's] traffic came from the Justice Department and other governmental units.").

[45] *Id.* at 861.

[46] *See Federal Courts Turn a New Page: Case Management/Electronic Case Files Systems Bring Greater Efficiency/Access*, *supra* note 42.

[47] Martin, *supra* note 43, at 861 ("Initially, those using the system had to retrieve case records on a jurisdiction-by-jurisdiction basis, which meant they had to know which court was involved.").

[48] *See* Lynn M. LoPucki, *Court-System Transparency*, 94 IOWA L. REV. 481, 485–87 (2009) (detailing the limitations of PACER's search capabilities).

[49] *Id.*

[50] *25 Years Later, PACER, Electronic Filing Continue to Change Courts*, U.S. CTS. (Dec. 9, 2013), https://www.uscourts.gov/news/2013/12/09/25-years-later-pacer-electronic-filing-continue-change-courts [https://perma.cc/UW2S-4SF7].

[51] Martin, *supra* note 43, at 861–63.

[52] *Id.*

[53] *See* John L. Moreland, *Is Open Access Equal Access? PACER User Fees and Public Access to Court Information*, 49 DTTP 42, 43 (2021).

[54] *Id.*

the exception of court opinions and certain de minimis searches, always had to pay to download PDFs of the underlying court documents.[55]

PACER's limited search capabilities and associated costs severely restricted access to the underlying docket information.[56] Individual litigants used PACER to keep track of their own cases. But access beyond this by researchers or the public was practically nonexistent.[57]

In the late 1990s and early 2000s, commercial legal research databases began enhancing their search tools and expanding the legal materials they digitized.[58] Services such as Westlaw and LexisNexis introduced natural language searching, enabling users to search legal documents using plain language queries, and improved relevance ranking to enhance search results.[59] The commercial tools also began to incorporate PACER information into their proprietary databases. But the incorporation typically amounted to little more than purchasing the information and essentially replicating PACER itself, with little additional information or synthesis.[60]

Over time, image-based databases of legal materials also emerged. Companies like HeinOnline scanned and provided access to original legal documents such as case opinions, statutes, and regulations.[61] These image-based databases preserved the formatting, layout, and typography of legal documents, providing authenticity and visual context.[62] However, the full-text searching was challenging and developers frequently changed the search architecture.[63]

In the late 2000s and early 2010s, new companies developed software to perform litigation analytics. These tools performed basic analysis about a particular case, court, or judge.[64] One of the first of these technologies was

---

[55] *See id.* at 42–43, 46 (detailing access issues with PACER); Boyd et al., *supra* note 20, at 467–69.

[56] *See* Schultze, *supra* note 30, at 1212, 1221, 1223 ("PACER fees both hinder the press from reporting on cases to the public and erect barriers for formal reporters of decisions.").

[57] *Id.* at 1212, 1221.

[58] Lynn Foster & Bruce Kennedy, *Technological Developments in Legal Research*, 2 J. APP. PRAC. & PROCESS 275, 280–81 (2000) (detailing improvements in Westlaw's and Lexis's search capabilities).

[59] *Id.* at 281.

[60] *See The LexisNexis Timeline Celebrating Innovation . . . and 30 Years of Online Legal Research,* LEXISNEXIS, https://www.lexisnexis.com/anniversary/30th_timeline_fulltxt.pdf [https://perma.cc/TG68-TQBN]; *see also Court Briefs, Records, and Dockets*, JEROME HALL L. LIBR., MAURER SCH. L., https://law.indiana.libguides.com/c.php?g=19814&p=112422 [https://perma.cc/H45G-RYGM].

[61] Joe Gerken, *The Invention of HeinOnline*, 18 AALL SPECTRUM 17, 19–20 (2014) (detailing the development of HeinOnline and its scanning and application of optical character recognition (OCR) to legal material).

[62] *Id.* at 18–19.

[63] *Id.* at 19.

[64] Peter A. Hook, *A Framework for Understanding, Using & Teaching Litigation Analytics*, 26 AALL SPECTRUM 20, 21 (2021).

Lex Machina, launched in 2010 by a group of legal practitioners, software engineers, and academics.[65] Lex Machina's primary focus was utilizing PACER court records to facilitate access to and analysis of patent litigation in federal courts.[66] Lex Machina went beyond merely redistributing PACER data by adding search capabilities for other fields and descriptive analytics, and coding litigation dockets to identify various stages of litigation.[67] To facilitate litigation searching, Lex Machina identified various litigation events, such as summary judgments or jury trials, by analyzing the dockets retrieved from PACER.[68] Users could then conduct data-driven analytics, such as determining how frequently a judge granted summary judgments in patent cases or the average time to trial at the district court level. Lex Machina was eventually acquired by LexisNexis in 2015.[69]

Following Lex Machina's lead, other legal analytical software emerged, including Docket Navigator, again mainly focused on patent litigation, and Bloomberg Law, which covered all types of federal litigation.[70] These tools offer similar functionalities to Lex Machina: PACER information with some additional searching and analysis.[71]

B. Free Data Sources

The common thread throughout the history of court records is that the public generally did not have access. Even for nominally "public" sources like the government's PACER website, the underlying databases were closed. Users could not download the basic data in bulk from PACER and were only able to inspect the database and its methodology via individual searches and the returned results, while subscribers to third-party services could not access PACER data in any kind of database format.

---

[65] *See Lex Machina: 10 Years of Legal Analytics*, CIOREVIEW, https://legal.cioreview.com/vendor/2020/lex_machina [https://perma.cc/HX48-WVGF] (discussing the beginnings of Lex Machina and its early capabilities).

[66] Daniel McKenzie, *Know Your Enemy: Lex Machina Raises $2 Million for IP Litigation Analytics*, TECHCRUNCH (July 26, 2012), https://techcrunch.com/2012/07/26/know-your-enemy-lex-machina-raises-2-million-for-ip-litigation-analytics/ [https://perma.cc/9RQS-C9MX].

[67] *See, e.g.*, Mark A. Lemley, *Where to File Your Patent Case*, 38 AIPLA Q.J. 401, 404, 404 n.11 (2010) (using Lex Machina to perform descriptive analysis on patent-litigant success in various federal district courts).

[68] *Lex Machina*, *supra* note 65.

[69] *Id.*

[70] *See Don't Guess. Know. Better Litigation Outcomes with Data-Driven Insights.*, DOCKETNAVIGATOR, https://brochure.docketnavigator.com [https://perma.cc/5NCU-3RYY]; *Court Dockets Search*, BLOOMBERG L., https://pro.bloomberglaw.com/products/court-dockets-search/ [https://perma.cc/KB3R-JYCG].

[71] *See* Ashley A. Ahlbrand, *Analyzing Analytics: Litigation Analytics in Bloomberg Law, Westlaw Edge, and Lexis Advance*, 42 CRIV SHEET 9, 10–11 (2020) (discussing Bloomberg Law's functionality); *Don't Guess*, *supra* note 70.

In recent years, however, there has been a rise in free and accessible digital legal research tools. One example is the Free Law Project, which aims to increase access to public court records and reduce the cost of accessing legal documents.[72] Yet such crowdsourced records present limitations due to their uneven representation of the underlying cases. For example, the Free Law Project obtains records from users who purchase them directly from PACER and then upload them to the Free Law Project (the uploading occurs automatically via RECAP, a browser extension installed by the user).[73] These records are thus limited to the individual documents purchased by users who have downloaded the extension. Most extant cases are thus not covered, and those that do appear are often incomplete.

Similarly, Cornell's Legal Information Institute offers free searchable digital access to statutes and regulations as well as other primary legal materials.[74] Harvard's Caselaw Access Project provides free access to all official, book-published judicial decisions through 2020.[75] These are extremely valuable sources, but they do not have access to the comprehensive judicial records available on PACER or other commercial databases. They also lack advanced analytical tools and the data enrichment needed to answer specific legal questions or synthesize underlying data.

III. INTRODUCING SCALES

In this Part, we introduce the SCALES dataset and preview the types of insights that it can generate. We begin with a discussion of the sources of the data. We then describe the protocols we developed to standardize and organize the data. Finally, we present new foundational descriptive statistics about litigation in the federal courts. We close with a case study on access to justice.

---

[72] *See About Free Law Project*, FREE L. PROJECT, https://free.law/about [https://perma.cc/TXX7-XS7A]. Other projects do as well, such as Stanford Law School's Intellectual Property Litigation Clearinghouse (now Lex Machina), Stanford Law School's Securities Class Action Clearinghouse, and University of Michigan Law School's Civil Rights Litigation Clearinghouse. *See Intellectual Property Litigation Clearinghouse*, STAN. L. SCH., https://law.stanford.edu/publications/intellectual-property-litigation-clearinghouse-data-overview/ [https://perma.cc/Q734-GTR2]; *Securities Class Action Clearinghouse*, STAN. L. SCH., https://securities.stanford.edu/about-the-scac.html [https://perma.cc/7V3F-K8BM]; C.R. LITIG. CLEARINGHOUSE, https://clearinghouse.net/about [https://perma.cc/2Y2T-G8V8].

[73] *RECAP Suite—Turning PACER Around Since 2009*, FREE L. PROJECT, https://free.law/recap [https://perma.cc/K6K8-RZM5].

[74] *See Who We Are*, CORNELL L. SCH., LEGAL INFO. INST., https://www.law.cornell.edu/lii/about/who_we_are [https://perma.cc/K9JX-Z5VQ].

[75] The Caselaw Access Project is online at *Our Data*, CASELAW ACCESS PROJECT, https://case.law [https://perma.cc/FPL8-DH4K]. *See also About*, CASELAW ACCESS PROJECT, https://case.law/about/ [https://perma.cc/QKM4-DMU9].

A. Overview of SCALES

The SCALES Open Knowledge Network, an organization funded by the National Science Foundation, is dedicated to transforming the accessibility and transparency of federal courts.[76] One of the primary goals, as the name suggests, is to establish to an open knowledge network (OKN). By definition, an OKN is freely available to all stakeholders, including the researchers who will help push this technology further. It is a nonproprietary public–private development effort that spans the entire data science community. The result of an OKN is an open, shared infrastructure. The formation of the SCALES OKN was driven by a clear need: the unavailability of raw data from the U.S. federal courts for comprehensive research purposes. SCALES brought together an interdisciplinary team with expertise in law, social science, journalism, and data and computer science, with the goal of making federal court data freely and easily accessible.[77]

SCALES uses AI tools to create a platform that enables systematic analysis of court records.[78] This platform is made publicly available via a data explorer.[79] Crucially, users can take full advantage of the data explorer without any computer programming knowledge. The data explorer is designed to accept common-language queries and questions. The underlying data powering the data explorer is drawn from PACER using software that automatically downloads queries, dockets, case summaries, and documents.[80] Importantly, the goal of SCALES is not to serve as a financial intermediary to PACER. Instead, SCALES extracts, transforms, and enriches PACER data to make it amenable to nuanced analysis and accessible to everyone.

Through the remainder of Part III, we detail the origins and development processes of the SCALES database. The computational methods developed and employed by SCALES for data acquisition, processing, and organization will be elaborated upon in a separate article.[81]

---

[76] For an overview of the most relevant sources, refer to the SCALES site: *About Scales*, SCALES, https://scales-okn.org/about-the-project/ [https://perma.cc/9APD-PKDP]; this piece in SCIENCE: Pah et al., *supra* note 2; and the SCALES documentation site: *SCALES OKN Documentation*, SCALES, https://docs.scales-okn.org/ [https://perma.cc/2WJU-D76W].

[77] *Our Team*, SCALES, https://scales-okn.org/team-2/ [https://perma.cc/8QVV-66X8].

[78] For a detailed description of the specific computational methods used to create this platform, see *SCALES OKN Documentation*, *supra* note 76.

[79] To access the data explorer, visit *Transforming the Accessibility and Transparency of Federal Courts*, SCALES, https://scales-okn.org [https://perma.cc/7JV3-JFR7]. See *infra* Section III.F for a more detailed description of the data explorer.

[80] All SCALES software falls under a GPL license and is available for use at our GitHub repository. *SCALES*, GITHUB, https://github.com/scales-okn [https://perma.cc/6842-5D6L] (hosting the SCALES software along with a full suite of documentation).

[81] *See SCALES OKN Documentation*, *supra* note 76.

Additionally, the code developed by the SCALES team is freely accessible for both review and use by the public.[82] Thus, the following Sections do not provide a technical explanation of the processes we developed. Instead, they offer a general overview of the SCALES organization and data platform.

B. Acquiring and Processing Court Data

The SCALES team initially focused on extracting case information from the federal courts' Case Management/Electronic Case Files (CM/ECF) system via the PACER interface. Due to the unique CM/ECF system maintained by each of the 94 judicial districts (which each have minor operational variances), the code was tailored for each district. Instead of sampling, SCALES opted to comprehensively download case details filed in the years 2016 and 2017 across all U.S. district courts. This approach allows for a complete view of the federal litigation landscape during these years. We intend to eventually expand this coverage to all years.

Our first step involved downloading the docket report, also known as a "docket sheet," for each case, which serves as a comprehensive, real-time chronological index of all events in a case. These dockets are distinct from the underlying case filings themselves. To grasp the scale of the underlying documents, we downloaded a sample and used it to estimate the total cost of acquiring all documents from PACER. Based on a projected expense of $0.10 per PDF page, we estimated the total expenditure for one year's worth of documents to be between $5.3 million and $5.5 million.[83] Given these significant financial considerations, we deferred the acquisition of the complete set of underlying documents to a later stage.

Our process commenced with the raw docket reports downloaded in HTML format, followed by extensive cleanup efforts.[84] The docket reports include case header information (such as the nature of the suit, presiding judge, and filing dates), the parties (including addresses), lawyers (firm name, lawyer name, address, phone number, and pro hac vice status), and docket entries for each litigation event. While the information on case headers, parties, and legal representation is structured, docket entries consist of unstructured text; they are essentially an enumerated list of case activities.

---

[82] For a guide, see Scott Daniel, *PACER Parser: Observations, Warnings, and Advice*, SCALES (Mar. 30, 2023), https://docs.scales-okn.org/guide/parserguide/ [https://perma.cc/3F6Z-3DNM]. The code is available at *Scales-okn/PACER-tools*, GITHUB, https://github.com/scales-okn/PACER-tools [https://perma.cc/R8N2-BU5X] under a GPL license, giving users the ability to use and alter the code.

[83] *Modelling PACER Costs: A Technical Review*, SCALES (Dec. 21, 2020), https://scales-okn.org/2020/12/21/modelling-pacer-costs-a-technical-review/ [https://perma.cc/6VZ2-FM3V] ("Our final modelled document cost was on average somewhere between $5.3 million and $5.5 million.").

[84] Scott Daniel, *Notes on Our Internal Data Pipeline*, SCALES (Mar. 26, 2024), https://docs.scales-okn.org/guide/pipeline/ [https://perma.cc/3SS9-T6GQ].

These entries, which are generated by both the court and the parties involved in the litigation, contain a rich narrative of the legal proceedings. To systematically analyze and enrich this unstructured data, we employed AI and other advanced computational techniques to annotate the docket information.

One of the principal challenges was to disambiguate entities such as litigants, lawyers, judges, and third parties, and to map the evolution of the intricate relationships among these entities over the course of a case. We developed two methods to address this: (1) sophisticated techniques for entity disambiguation and (2) a set of event ontologies. In lay terms, we developed methods to determine whether, for example, the lawyer "Bill Taft" in one case was the same as a "William Taft" in another, as well as a set of methods and category labels to determine whether a given docket entry was, for example, a complaint or a motion to dismiss. The next two Sections describe each in turn.

C. Entity Disambiguation

There were many challenges to distinguishing names. For example, judges could be referenced in a variety of ways: some entries might list the full name including middle name, others might use only a middle initial or exclude the middle name entirely, and still others might only mention the last name. Additionally, titles such as "The Honorable," "Judge," or "District Court Judge" might precede the judge's name in some cases, while in others, a title is used without any name. Spelling errors and the existence of multiple judges sharing the same first and last names posed additional challenges. Given the finite number of federal judges, we were able to construct a model that correctly identifies judges from the docket entry text in nearly all cases. We also established a disambiguation pipeline that links these identified judge entities to their official biographical records, thereby enriching the dataset and expanding the analytical possibilities for users.

Disambiguation of parties, lawyers, and law firms presented even more complex challenges. For example, the prevalence of common names or familial naming conventions (e.g., John Smith or John Smith Jr.) leads to confusion. Corporations also introduced disambiguation difficulties, with a single company potentially being referred to in multiple variations (e.g., "IBM," "IBM, Inc.," "IBM Corp.," "IBM Corporation," "International Business Machines," and so on). We addressed these varied challenges through the development of specialized algorithms.

To further refine our understanding of entity relationships, we implemented a custom Named Entity Recognition (NER) pipeline.[85] This tool identifies parties within docket sheets and tracks their appearances across different cases. While the concept may seem straightforward, the execution is complex due to factors such as judicial reassignments, name changes, and title variations. Our NER pipeline and disambiguation processes manage these factors to distinguish and track the identities of judges, lawyers, and law firms.

Processing the text of the docket entries presented additional challenges. This task was particularly difficult because each court, judge, clerk, lawyer, and party can use their own idiosyncratic linguistic methods to refer to litigation motions, notices, events, and case progression.[86] The root source of this problem lies in PACER's limited mission. PACER was primarily designed to make case management by the court easier—not to provide open access to the public or to enable third parties to interpret or analyze court records.[87] The PACER system is fundamentally inward-looking and ad hoc. Our multi-step annotation process remedies this. The result is that judges, parties, and the litigation events they involve are indexed, searchable, and amenable to detailed analysis.

D. Event Ontology

We developed a system of "event ontologies" to label and categorize legal events. We use these labels to construct a narrative map of each case. An individual docket may be filled with entries of minor importance, such as a request to increase the page limit of a brief, or a notice of a party's change in address. The goal of the event ontology labels is to sift through these details to identify the critical case milestones. Such milestones include events such as complaints, answers, indictments, motions, arrests, orders, extensions, dismissals, probations, and judgments.

We developed an extensive classification of granular litigation events to identify the pathway of each case. The classification scheme is hierarchical, including major types of events (entries, motions, notices, etc.) and then distinct subfilings for each type (Figure 1). We use the text of the docket entry to classify each litigation event. The prediction models are the

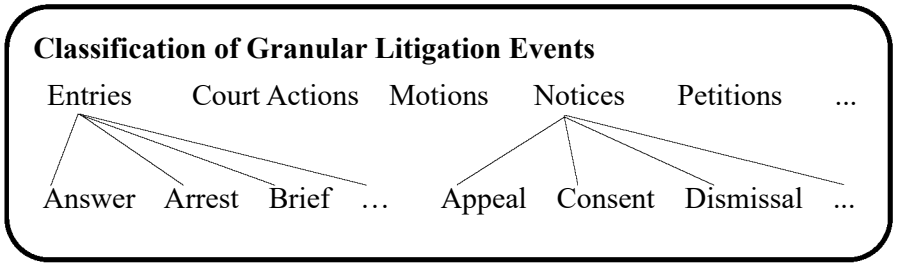
[85] Chris Rozolis, *Entity Disambiguation*, SCALES (Mar. 30, 2023), https://docs.scales-okn.org/guide/disambiguation/ [https://perma.cc/64RT-W5EZ].

[86] *See generally* Adam R. Pah, Christian J. Rozolis, David L. Schwartz & Charlotte S. Alexander, *PRESIDE: A Judge Entity Recognition and Disambiguation Model for US District Court Records*, *in* 2021 IEEE INTERNATIONAL CONFERENCE ON BIG DATA 2721, 2722 (2021), https://ieeexplore.ieee.org/document/9671351 [https://perma.cc/J5RA-W98C] (developing a model to disambiguate judges in court records to enable study of judicial decision-making variations).

[87] *See* Schultze, *supra* note 30, at 1221.

same for every case, whether it has one docket entry or one hundred. However, when we make predictions on certain pathway events that mark the beginning or end of the case, we allow the model to reason with additional data from the case. This additional data includes nearby docket entries and their classification labels, the entities involved, and the case metadata. By identifying these events and their interconnections within the litigation pathway, we have created a richly detailed dataset of court records that supports in-depth, granular analysis.

FIGURE 1: ILLUSTRATED SCALES LITIGATION EVENT ONTOLOGY

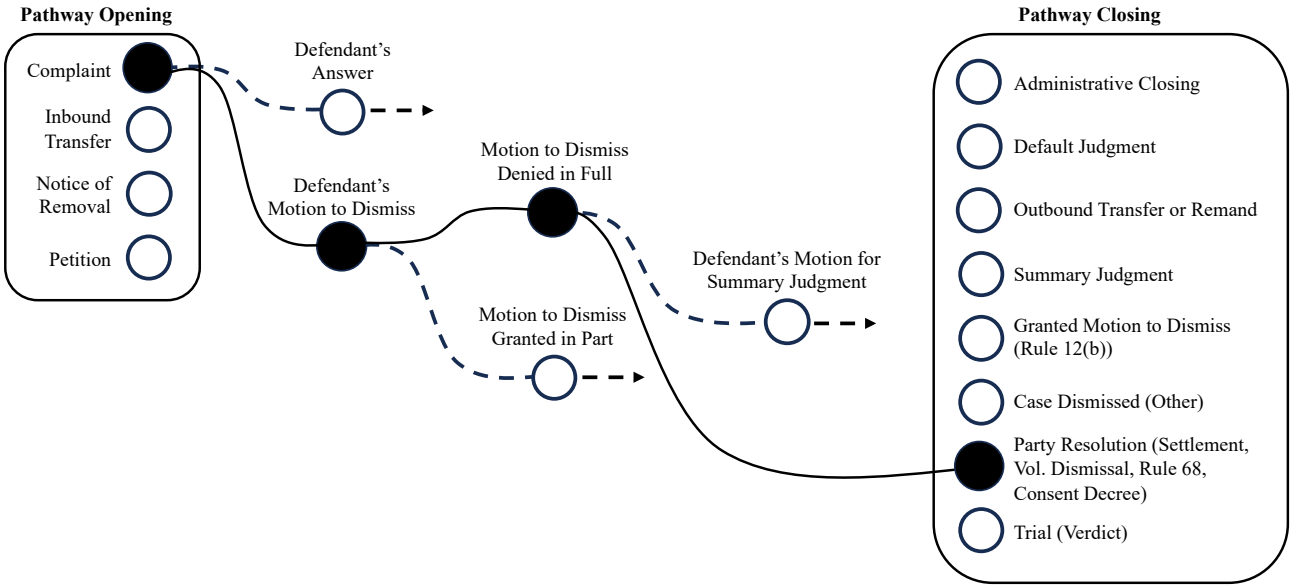


Our AI approach to construct event ontologies is robust as to variation in input methods, much more so compared to a more rudimentary approach that uses, for example, only keyword searches. This is because parties, judges, and courts often use different terminology when referring to the same legal event. This variation means that a precise keyword search for terms like "motion for summary judgment" will not capture all relevant instances. For example, such a search might miss entries labeled "motion for entry of summary judgment" or simply "motion," even when these entries refer to motions for summary judgment.

By integrating the granular classification labels with the notion of key pathway events, we can construct a comprehensive ontology of the litigation process and identify the diverse paths that civil and criminal litigation can follow. This holistic approach creates an optimally simplified and accurate representation of litigation events and their various terminologies. An illustrative example of a complete litigation ontology is depicted in Figure 2 below.[88]

[88] *See, e.g.*, Nathan Dahlberg, *Litigation Ontology*, SCALES (Mar. 30, 2023), https://docs.scales-okn.org/guide/ontology/#pathway-events [https://perma.cc/2VGT-APVZ] (explaining how the litigation events shown in Figure 2 are drawn from the SCALES civil ontology labels).

FIGURE 2: EXAMPLE OF A CASE'S POTENTIAL EVOLUTION VIEWED THROUGH THE LENS OF THE CIVIL LITIGATION EVENT ONTOLOGY



E. Comparing SCALES Data to Other Datasets

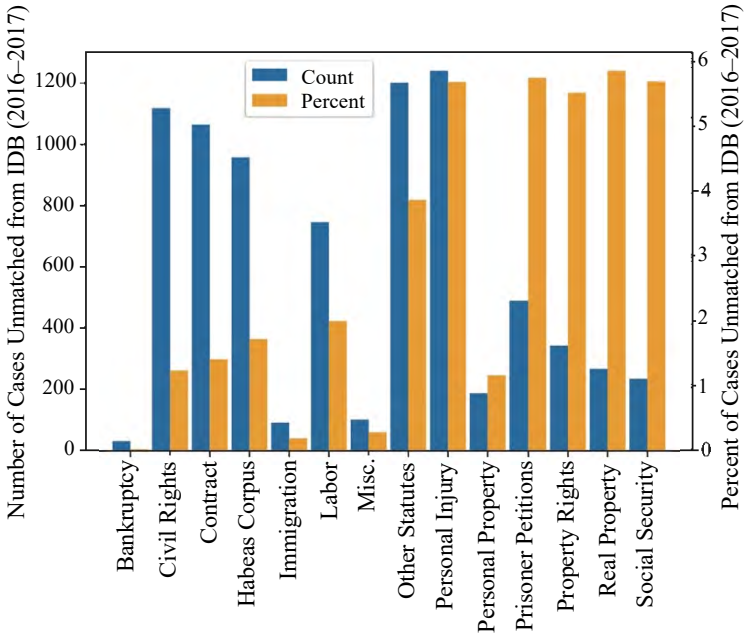
In this Section, we compare the SCALES data to other extant datasets. In general, we find that the SCALES data is more comprehensive and more accurate than both the federal judiciary's official statistics and commercial subscription services.

We begin by comparing SCALES to the judiciary's official case-level database: the Federal Court Integrated Database (IDB), which is maintained by the Federal Judicial Center in coordination with the Administrative Office of the Courts.[89] A common assumption regarding the IDB is that, because it is maintained by the federal courts themselves, the data is comprehensive

[89] *Integrated Database (IDB)*, FED. JUD. CTR., https://www.fjc.gov/research/idb [https://perma.cc/4UVK-SWGZ] ("The IDB contains data on civil case and criminal defendant filings and terminations in the district courts, along with bankruptcy court and appellate court case information from 1970 to the present.").

and accurate.[90] Our analysis indicates otherwise.[91] The SCALES data contains many cases from PACER that cannot be matched to the IDB (Figure 3). This discrepancy underscores the added value of the SCALES dataset in capturing a more complete picture of federal court activity.

FIGURE 3: UNMATCHED CASES IN THE IDB

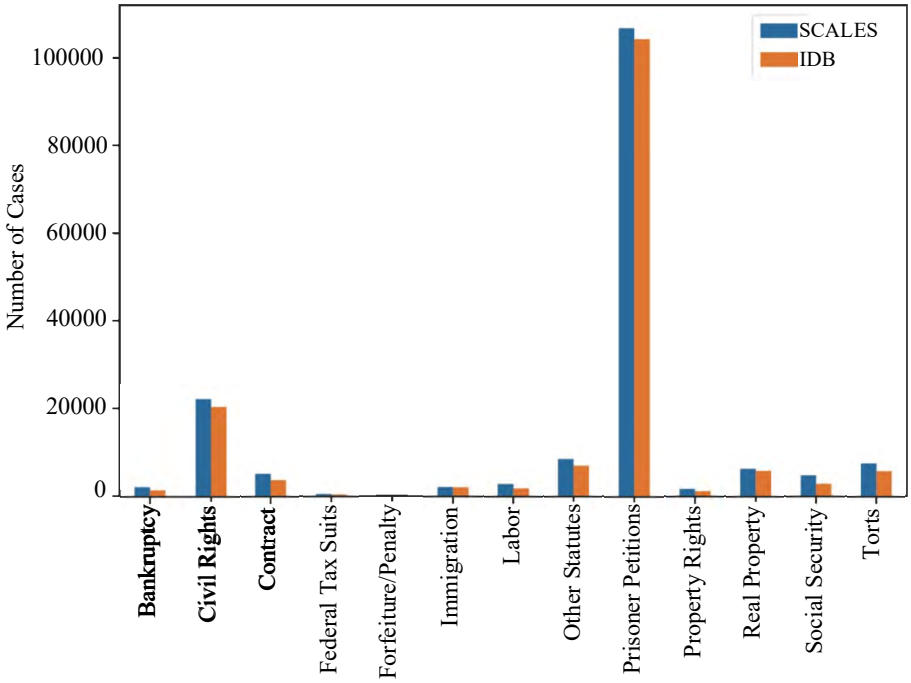


[90] *See* Kyle C. Kopko & Christopher J. Devine, *Home Court Advantage? An Empirical Analysis of Local Bias in U.S. District Court Diversity Jurisdiction Cases*, 125 W. VA. L. REV. 543, 545 (2022) ("Relying on the Integrated Database (IDB) . . . we present an original, empirical analysis of diversity jurisdiction case outcomes in the U.S. district courts from 1988 through 2021. This extensive database contains virtually every civil case filed in the U.S. district courts during this time frame . . . .").

[91] Other scholars have come to similar conclusions. *See, e.g.*, CHARLOTTE S. ALEXANDER & MOHAMMAD JAVAD FEIZOLLAHI, U.S. DEP'T OF LABOR, DECISIONAL SHORTCUTS AND SELECTION EFFECTS: AN EMPIRICAL STUDY OF TEN YEARS OF U.S. DISTRICT COURTS' EMPLOYEE MISCLASSIFICATION DECISIONS 8 (2020), https://www.dol.gov/sites/dolgov/files/OASP/evaluation/pdf/LRE_Alexander-DecisionalShortcutsandSelectionEffects_December2020.pdf [https://perma.cc/74QV-DR7H] ("Nine of [the IDB fields] had missing data for ninety-five percent or more of the lawsuits, including variables meant to capture whether the case was filed as a class action, variables relevant to transferred cases, and variables relevant to arbitrated cases. Other variables were missing values for about half of the records, including those that capture the party in whose favor final judgment was entered, and whether that judgment included a monetary award, injunctive relief, and/or attorneys' fees and costs."). Some scholars have identified error rates ranging from 29% to 69% in IDB data for disposition codes and even judgment amounts. *See* Gillian K. Hadfield, *Where Have All the Trials Gone? Settlements, Nontrial Adjudications, and Statistical Artifacts in the Changing Disposition of Federal Civil Cases*, 1 J. EMPIRICAL LEGAL STUD. 705, 710, 724 (2004); *see also* Theodore Eisenberg & Margo Schlanger, *The Reliability of the Administrative Office of the U.S. Courts Database: An Initial Empirical Analysis*, 78 NOTRE DAME L. REV. 1455, 1475 (2003).

The IDB also consistently fails to identify whether existing cases involve pro se litigants. The extent of these omissions varies significantly; depending on the category, the rate of missing cases ranges from 2% to 40% (Figure 4, showing raw counts). While the Federal Judicial Center recognizes limitations within the IDB, particularly in fields concerning underserved populations, our findings provide concrete documentation of these issues and show that the scope of the IDB's incompleteness is much greater than commonly believed.[92]

FIGURE 4: COMPARISON OF SCALES DATA TO MATCHED IDB DATA FOR SELECTED NATURES OF SUIT



In addition to being more comprehensive, the SCALES platform also outperforms the subscription analytics services offered by Bloomberg Law, Westlaw, and Lexis. To demonstrate this, we designed two benchmark searches: one for "motions to dismiss" and another for "class certification" motions, both for cases filed in 2016 and 2017. The SCALES dataset

[92] FED. JUD. CTR., THE INTEGRATED DATABASE: A RESEARCH GUIDE 4, www.fjc.gov/sites/default/files/IDB-Research-Guide.pdf [https://perma.cc/8CH7-S7BD] ("Nonetheless, there may be some problems with specific fields that are not routinely reported. The two issues with respect to data collection noted above are more likely to affect specific fields related to under-served populations.").

identifies significantly more cases (1,600) than Bloomberg (1,161) and Westlaw (542) with these motions in the Southern District of Indiana, as reflected in Table 1A. The Lexis platform is unable to perform the basic search as it cannot restrict the search by filing year.[93]

Notably, there is a substantial overlap in the cases identified by SCALES and those found by Bloomberg and Westlaw, suggesting that the "motions to dismiss" located by the commercial providers are also captured by SCALES. Because the search algorithms are proprietary, we cannot know for certain why Bloomberg and Westlaw have a lower retrieval rate.[94] However, we would speculate that these algorithms focus on minimizing false positives ("type I" errors) because these are errors that users can see and thus identify as errors. For example, suppose a user searches for "Motion for Summary Judgment" and the algorithm returns a list of results. If one of those results is actually a "Motion to Dismiss" (a false positive), then the user can see the error. If the algorithm fails to identify a true "Motion for Summary Judgment" (a false negative), however, the user is none the wiser because they only see the positive results. Moreover, one simple way to minimize false positives is through exact keyword matching, which will tend to retrieve only documents that exactly fit the search criteria. Again, while we cannot know for sure, this strikes us as one plausible explanation for why the subscription services tend to have a lower retrieval rate.

TABLE 1A: COMPARISON OF MOTION TO DISMISS DATA-QUERY RESULTS ACROSS SYSTEMS

Database	Number of Cases	Overlap with SCALES
Westlaw	542	540
Bloomberg	1161	1132
SCALES	1600	-

[93] *Searching by Date on Lexis+*, LEXISNEXIS, https://supportcenter.lexisnexis.com/app/answers/answer_view/a_id/1088494/~/%20Filters%20on%20Lexis%2B [https://perma.cc/GYK9-AGHX].

[94] To guard against the possibility that misclassified events by SCALES are driving the difference in retrieval rate, we conducted a series of manual validation checks where we confirmed the result of the SCALES classification with a trained legal professional's case classification. We found that SCALES classification yielded 99% precision for motions to certify and 100% precision for motions to dismiss, broadly defined. A fuller robustness check of SCALES classification is available at *SCALES OKN Documentation*, *supra* note 76.

Our analysis of motions for class action certification yields even more striking results, displayed in Table 1B. Bloomberg Law's analytics do not even cover these motions, so our comparison in Table 1B is solely with Westlaw. The data reveals a high degree of overlap between Westlaw and SCALES, indicating a common set of identified cases. However, Westlaw's dataset lacks approximately two-thirds of the motions to certify class actions that SCALES captures in the Northern District of Illinois. Again, because Westlaw's search algorithms are proprietary, one can only speculate as to why their retrieval rate is so low.

TABLE 1B: COMPARISON OF MOTION TO CERTIFY A CLASS DATA-QUERY RESULTS ACROSS SYSTEMS

Database	Number of Cases	Overlap with SCALES
Westlaw	137	135
Bloomberg	-	-
SCALES	414	-

In conclusion, the SCALES dataset demonstrates a more comprehensive capture rate for common legal searches compared to other well-established sources, including the Federal Court's IDB and commercial providers. Compared to the alternatives, SCALES is the most comprehensive and complete. It is also the only source that is freely accessible by the public.

F. The SCALES OKN Data Explorer

The SCALES OKN Data Explorer was designed to make it as easy as possible to access and analyze federal court records. To this end, we built a system that enables users to filter and query the SCALES OKN corpus and generate aggregate statistics and trends across all 94 federal district courts. The design was partly based on user interviews and feedback.[95]

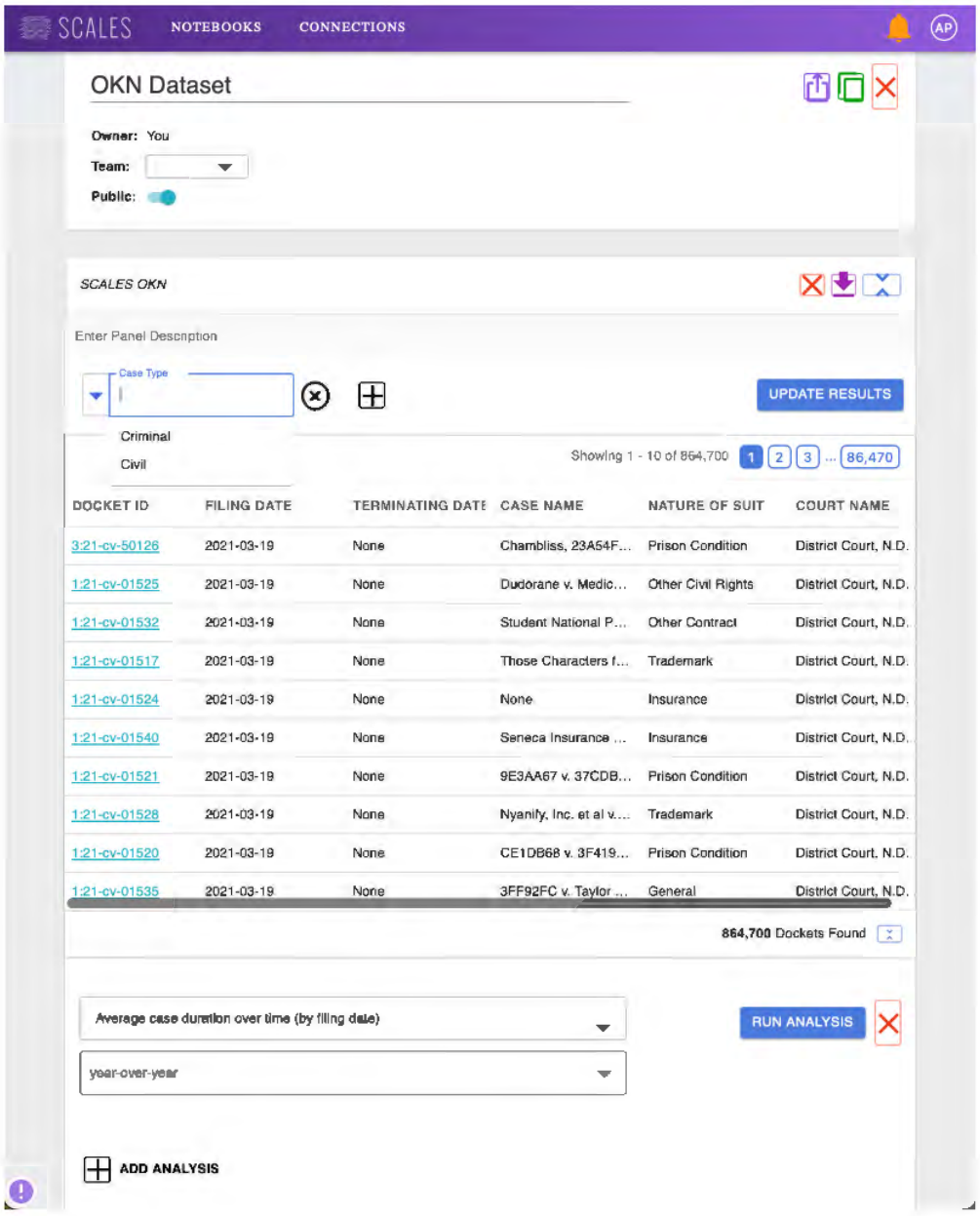
The Data Explorer research notebook is split into two distinct parts: data filtering and data analytics. Data filtering allows the user to set multiple parameters to refine their search, such as restricting the corpus to only civil cases or only cases filed in 2016. Users can then view the matching cases and download the docket reports as a CSV file for further analysis. The data

[95] Rachel F. Adler, Andrew Paley, Andong L. Li Zhao, Harper Pack, Sergio Servantez, Adam R. Pah & Kristian Hammond, *A User-Centered Approach to Developing an AI System Analyzing U.S. Federal Court Data*, 31 A.I. & L. 547, 566–67 (2022).

analytics section also allows users to quickly calculate and plot aggregate statistics about the filtered case results. Currently, the analytics section can analyze the volume, duration, and cost of cases and how these quantities are distributed in time, across geography or nature of suit, and by attributes of the case (e.g., whether a fee waiver was filed in the case, or the name of the judge).

The Data Explorer is designed to allow for collaboration and sharing of analysis to further facilitate research reproducibility. Any user can develop a research notebook with the OKN data and then publicly share it as a read-only notebook. Any other person on the platform or with the link can then view the data filtering and analysis steps. If a group wishes to work collectively on analyzing a collection of cases, then they can alternatively create a team and share a notebook. All members of a shared notebook can make changes to the data filters and generate analyses on the resulting case records.

FIGURE 5: THE SCALES–OKN DATA EXPLORER



G. Descriptive Statistics from the SCALES Data Explorer

In this Section, we use the SCALES Data Explorer to generate new foundational descriptive statistics for federal district court litigation for the years 2016 and 2017 combined. These are the two years for which the SCALES database has dockets for all suits filed in federal court. These statistics cover a variety of dimensions, including nature of suit, grant rates for motions to dismiss and for summary judgment, trial frequency, litigation intensity, and the frequency of appearances by lawyers and law firms.

Table 2 provides the distribution of federal litigation across various case types, detailing the quantity and percentage of cases within each nature of suit, as well as identifying the federal court districts with the highest and lowest frequency of these cases.[96] For the purposes of generating Tables 2 through 8, we include only those districts reporting at least thirty cases that meet the applicable criteria (i.e., thirty motions to dismiss).

The data show that Criminal proceedings, followed by Personal Injury lawsuits and Habeas Corpus petitions are the most common natures of suit. However, the composition of case types exhibits significant variation across districts. For instance, Criminal matters constitute 76% of cases in the District of New Mexico but make up just 1% of cases in the Eastern District of Louisiana. Intriguingly, the district with the highest case volume across all case types is the Eastern District of Louisiana.

TABLE 2: DISTRIBUTION OF NATURE OF SUITS (ALL FEDERAL DISTRICT COURTS, 2016 & 2017)

Rank	Nature of Suit	Number of Cases	Percent of Total	Min District	Max District
1	Criminal	126,371	18.7%	E.D. La. (1.3%)	D.N.M. (76%)
2	Personal Injury	115,924	17.1%	D.N.M. (2.3%)	S.D.W. Va. (89%)
3	Habeas Corpus	90,317	13.3%	D.P.R. (1.5%)	M.D. Tenn. (51%)
4	Civil Rights	75,495	11.2%	S.D.W. Va. (0.8%)	W.D. Pa. (23%)

[96] The nature of suit categorization is developed by the judiciary. We report only the major nature of suit categories. *See Nature of Suit,* PACER, https://pacer.uscourts.gov/sites/default/files/files/nature%20of%20suit%20codes.pdf [https://perma.cc/MQ9M-8HHL].

5	Other Statutes	55,662	8.2%	S.D.W. Va. (1.1%)	D.D.C. (29%)
6	Contract	47,550	7.0%	S.D.W. Va. (1.2%)	M.D. La. (38%)
7	Social Security	37,332	5.5%	E.D. La. (0.3%)	E.D. Okla. (34%)
8	Labor	35,656	5.3%	S.D.W. Va. (0.7%)	E.D.N.Y (14%)
9	Prisoner Petitions	31,056	4.6%	E.D. La. (0.5%)	E.D.N.C. (21%)
10	Property Rights	21,821	3.2%	E.D. La. (0.2%)	D. Del. (38%)
11	Real Property	16,170	2.4%	E.D. La. (0.4%)	D.P.R. (31%)
12	Personal Property	8,522	1.3%	S.D. Ind. (0.5%)	N.D. Cal. (5%)
13	Immigration	6,229	0.9%	E.D. Cal. (0.4%)	M.D. Ga. (6%)
14	Bankruptcy	4,136	0.6%	E.D. La. (0.1%)	D. Del. (5%)
15	Forfeiture/Penalty	2,096	0.3%	D.N.J. (0.1%)	D. Kan. (2%)
16	Federal Tax Suits	1,895	0.3%	N.D. Ill. (0.2%)	D. Utah (1%)
17	Civil Detainee	567	0.1%	M.D. Fla. (0.3%)	C.D. Ill. (5%)
	All Combined	676,799	100%	D.N. Mar. I. (0.0%)	E.D. La. (5%)

Note. "Min District" is the district for which that nature of suit is the lowest share of the district's case load. So in this table, that means Labor cases make up 0.7% of cases in the Southern District of West Virginia (S.D.W. Va.) (and that this is the lowest among all districts with at least 30 cases). Similarly for the "Max District," this means Labor cases make up 14% of the Eastern District of New York (E.D.N.Y.) cases (and that this is the highest among all districts with at least 30 cases). For the "Min District"/"Max District" in the last row (All cases combined) these are the districts with the fewest and most cases. So the District Court for the Northern Mariana Islands (D.N. Mar. I.) is the smallest district in terms of case load (it has 0% of all cases), and the Eastern District of Louisiana (E.D. La.) is the largest with 5% of all cases.

We now advance from simple case counts to a more nuanced examination of litigation outcomes, employing the sophisticated civil litigation ontology developed and implemented by SCALES. In Tables 3 and 4, we explore adjudication rates of two common dispositive motions, motions to dismiss and motions for summary judgment. Our findings reveal that, within cases with rulings, motions to dismiss are granted 45% of the time, while motions for summary judgment see a slightly higher grant rate of 52%.[97]

Again, the national averages mask substantial disparities between districts, which are clear both in the aggregate and when divided by nature of suit. For instance, the District of North Dakota has a strikingly high grant rate of motions to dismiss in Habeas Corpus petitions with 89%, in contrast to only 16% in the Western District of Louisiana. Similarly, the Southern District of Florida grants only 12% of motions for summary judgment in Real Property disputes, while the Western District of Washington grants 79%.

TABLE 3: GRANT RATES FOR MOTION TO DISMISS (ALL FEDERAL DISTRICT COURTS, 2016 & 2017)

Rank	Nature of Suit	Grant Rate	Min District	Max District
1	Civil Detainee	58%	M.D. Fla. (62%)	D. Minn. (66%)
2	Criminal	57%	D.N.J. (7%)	E.D. Tex. (95%)
3	Immigration	56%	S.D.N.Y. (42%)	W.D.N.Y. (82%)
4	Prisoner Petitions	55%	M.D. Fla. (51%)	D.S.D. (83%)
5	Social Security	55%	E.D. Ky. (11%)	N.D. Okla. (95%)
6	Habeas Corpus	52%	W.D. La. (16%)	D.N.D. (89%)
7	Civil Rights	49%	D.V.I. (15%)	W.D. Mich. (89%)
8	Real Property	46%	N.D. Ala. (19%)	E.D.N.C. (82%)

[97] A motion to dismiss or for summary judgment was considered granted if it was granted in full or in part. The motion was counted as denied if the entire motion was denied.

9	Federal Tax Suits	46%	No district with ≥ 30 cases	No district with ≥ 30 cases
10	Bankruptcy	45%	S.D. Tex. (39%)	No district with ≥ 30 cases
11	Personal Injury	43%	D. Guam (9%)	D. Utah (67%)
12	Personal Property	40%	S.D. Ill. (19%)	E.D. La. (67%)
13	Forfeiture/Penalty	39%	No district with ≥ 30 cases	No district with ≥ 30 cases
14	Other Statutes	39%	D. Alaska (13%)	N.D.W. Va. (67%)
15	Contract	37%	D.V.I. (8%)	D.N.M. (55%)
16	Labor	34%	W.D. Tenn. (16%)	W.D. Wash. (61%)
17	Property Rights	31%	E.D. Pa. (15%)	D. Ariz. (53%)
	All cases combined	45%	D.V.I. (30%)	W.D. Mich. (71%)

Note. For "Min District," this means that in the Northern District of Alabama (N.D. Ala.), 19% of all motions to dismiss in Real Property cases are granted. It also means that this is the lowest grant rate among all districts that have at least 30 Real Property motions to dismiss. For the "Min District" and "Max District" in the last row (All cases combined) these are the districts with the lowest and highest overall grant rates. So the District Court of the Virgin Islands (D.V.I.) has the lowest grant rate (30%) and the Western District of Michigan (W.D. Mich.) has the highest (71%).

TABLE 4: GRANT RATES FOR MOTION FOR SUMMARY JUDGMENT (ALL FEDERAL DISTRICT COURTS, 2016 & 2017)

Rank	Nature of Suit	Grant Rate	Min District	Max District
1	Social Security	68%	W.D. Wis. (1%)	W.D. Pa. (96%)
2	Civil Detainee	60%	No district with ≥ 30 cases	C.D. Ill. (70%)
3	Federal Tax Suits	57%	No district with ≥ 30 cases	No district with ≥ 30 cases
4	Civil Rights	56%	D. Me. (35%)	D. Wyo. (84%)
5	Habeas Corpus	56%	S.D. Ala. (17%)	W.D.N.C. (83%)
6	Real Property	53%	S.D. Fla. (12%)	W.D. Wash. (79%)
7	Labor	49%	E.D. Tex. (24%)	S.D. Ind. (74%)
8	Other Statutes	48%	E.D. Mo. (29%)	D.N.M. (72%)
9	Contract	46%	D.V.I. (9%)	N.D. Ga. (70%)
10	Forfeiture/Penalty	44%	No district with ≥ 30 cases	No district with ≥ 30 cases
11	Immigration	44%	No district with ≥ 30 cases	C.D. Cal. (52%)
12	Bankruptcy	40%	S.D. Tex. (23%)	No district with ≥ 30 cases
13	Personal Property	38%	E.D. Pa. (20%)	W.D. Wash. (56%)
14	Personal Injury	36%	S.D.W. Va. (8%)	D. Del. (68%)
15	Property Rights	34%	E.D. Tex. (6%)	N.D. Ga. (65%)
16	Prisoner Petitions	11%	S.D. Tex. (8%)	No district with ≥ 30 cases

17	Criminal	10%	No district with ≥ 30 cases	D.S.C. (12%)
	All cases combined	52%	D.V.I. (11%)	E.D. Wash. (83%)

Note. This means that in the Western District of Washington (W.D. Wash.), 79% of all motions for summary judgment in Real Property cases are granted. It also means that this is highest grant rate among all districts that have at least 30 Real Property motions to dismiss. For "Min District" and "Max District" in the last row (All cases combined) these are the districts with the lowest and highest overall grant rates. So the District Court of the Virgin Islands (D.V.I.) has the lowest grant rate (11%) and the Eastern District of Washington (E.D. Wash.) has the highest (83%).

Table 5 shows rates of cases reaching a trial. Consistent with the literature on the "Vanishing Trial," we find that only 1.3% of lawsuits reach trial.[98] Many of the natures of suit do not have any court with at least thirty trials in the two-year window. The variation across districts and natures of suit, while more muted than those from dispositive motions in percentage point terms, is still high in percent terms. For example, the most likely nature of suit to reach a trial is Criminal cases, at 3.2%. This is twice as high as the next-highest category (Contract, at 1.6%) and just shy of three times the average trial rate (1.3%).

TABLE 5: PERCENTAGE OF CASES THAT GO TO TRIAL (ALL FEDERAL DISTRICT COURTS, 2016 & 2017)

Rank	Nature of Suit	Trial Rate	Min District	Max District
1	Criminal	3.2%	W.D. Tex. (1.0%)	E.D. Ky. (12.2%)
2	Contract	1.6%	S.D.N.Y. (2.0%)	C.D. Cal. (3.1%)
3	Civil Rights	1.5%	S.D. Fla. (1.0%)	S.D. Tex. (3.0%)
4	Personal Property	1.4%	No district with ≥ 30 cases	No district with ≥ 30 cases
5	Property Rights	1.3%	No district with ≥ 30 cases	C.D. Cal. (3.0%)
6	Federal Tax Suits	1.3%	No district with ≥ 30 cases	No district with ≥ 30 cases

[98] Marc Galanter, *The Vanishing Trial: An Examination of Trials and Related Matters in Federal and State Courts*, 1 J. EMPIRICAL LEGAL STUD. 459, 461 (2004).

7	Labor	1.1%	S.D. Fla. (1.2%)	C.D. Cal. (3.2%)
8	Real Property	1.0%	No district with ≥ 30 cases	W.D. La. (30.9%)
9	Personal Injury	0.7%	E.D. La. (0.1%)	C.D. Cal. (3.0%)
10	Habeas Corpus	0.7%	E.D. Wis. (3.4%)	S.D. Ohio (4.8%)
11	Forfeiture/Penalty	0.6%	No district with ≥ 30 cases	No district with ≥ 30 cases
12	Civil Detainee	0.5%	No district with ≥ 30 cases	No district with ≥ 30 cases
13	Immigration	0.5%	No district with ≥ 30 cases	No district with ≥ 30 cases
14	Other Statutes	0.4%	No district with ≥ 30 cases	C.D. Cal. (0.9%)
15	Bankruptcy	0.3%	No district with ≥ 30 cases	No district with ≥ 30 cases
16	Prisoner Petitions	0.0%	No district with ≥ 30 cases	No district with ≥ 30 cases
17	Social Security	0.0%	No district with ≥ 30 cases	No district with ≥ 30 cases
	All cases combined	1.3%	S.D.W. Va (0.1%)	D.V.I. (5.6%)

Note. This means that in the Southern District of New York (S.D.N.Y.), 2% of all Contract cases go to trial. It also means that this is the lowest trial rate among all districts that have at least 30 Contract cases that went to trial. For the "Min District" and "Max District" in the last row (All cases combined) these are the districts with the lowest and highest overall trial rates. So the Southern District of West Virginia (S.D.W. Va.) has the lowest trial rate (0.1%) and the District Court of the Virgin Islands (D.V.I.) has the highest (5.6%). The average trial rate (All cases combined) is 1.3%.

Next, we investigate litigation intensity and frequency of lawyer and law firm participation in litigation. We measure litigation intensity by counting the number of docket entries per case.[99] This metric serves as a

[99] Others have used docket entries as a measure of litigation intensity. *See* Jay P. Kesan & Gwendolyn G. Ball, *How Are Patent Cases Resolved? An Empirical Examination of the Adjudication and Settlement*

rough proxy for the complexity and procedural demands of litigation. By this metric, the average case has 34 entries. The most intensely litigated nature of suit, Criminal cases, is 50% above this average (51 entries per case), followed by Property Rights cases (over 40% above average, or 48 per case). Property Rights cases are litigated most intensely in the District of Delaware (D. Del.), with an average of 82 entries per case. Note that the Property Rights category covers three types of intellectual property cases: patent, copyright, and trademark.

TABLE 6: LITIGATION INTENSITY (ALL FEDERAL DISTRICT COURTS, 2016 & 2017)

Rank	Nature of Suit	Litigation Intensity	Min District	Max District
1	Criminal	51	D. Haw. (17)	D. Conn. (104)
2	Property Rights	48	D.D.C. (23)	D. Del. (82)
3	Personal Property	43	D.S.D. (12)	W.D. Mo. (108)
4	Contract	40	M.D. La. (27)	D. Neb. (67)
5	Civil Detainee	40	D.S.C. (28)	C.D. Ill. (49)
6	Civil Rights	39	S.D. Cal. (24)	W.D. Mich. (81)
7	Labor	36	D.S.C. (20)	D.P.R. (60)
8	Real Property	34	S.D. Cal. (17)	N.D. Miss. (116)
9	Federal Tax Suits	34	C.D. Cal. (24)	D. Minn. (45)
10	Other Statutes	32	E.D. La. (6)	N.D.W. Va. (115)
11	Habeas Corpus	27	M.D. Tenn. (9)	S.D. Ill. (58)

of Patent Disputes, 84 WASH. U. L. REV. 237, 284 (2006) (using the number of docket entries in a case as a measure of expenditures).

12	Forfeiture/Penalty	24	D. Md. (14)	D. Colo. (46)
13	Social Security	24	S.D. Iowa (16)	E.D.N.C. (35)
14	Bankruptcy	22	W.D.N.C. (10)	E.D.N.C. (47)
15	Personal Injury	22	E.D. La. (7)	D. Md. (132)
16	Immigration	16	D.N.J. (10)	N.D. Cal. (24)
17	Prisoner Petitions	10	E.D. Cal. (1)	D.N.H. (26)
	All cases combined	34	E.D. La. (10)	D.V.I. (71)

Note. Litigation Intensity is the number of docket entries in a case. This means that the most intensively litigated nature of suit is Criminal cases (which have an average of 51 entries per case), followed by Property Rights cases (48 per case). Property Rights cases are litigated most intensely in the District of Delaware (D. Del.), with an average of 82 entries per Property Rights case. The last row "All cases combined" says there are 34 entries on average (averaged over all cases), and the District Court of the Virgin Islands (D.V.I) has the most entries per case on average (71).

Turning to legal representation in Table 7, an interesting pattern emerges: a larger proportion of lawyers (43%) are involved in 2–5 cases as opposed to a single case (35%), within the two-year span of 2016 and 2017. That proportion is even more skewed for law firms, with 47% appearing in between 2–5 cases compared to only 20% with a single case. Furthermore, 11% of lawyers participate in between 6–10 cases, and an additional 9% represent parties in 11–50 cases during this period.

TABLE 7: DISTRIBUTION OF NUMBER OF APPEARANCES FOR LAWYERS AND LAW FIRMS (ALL FEDERAL DISTRICT COURTS, 2016 & 2017)

Number of Appearances	Number of Lawyers	Percent of Lawyers	Number of Law Firms	Percent of Law Firms
1	112,828	35%	17,362	20%
2–5	139,150	43%	41,939	47%
6–10	35,671	11%	12,528	14%
11–50	28,659	9%	13,416	15%
51–100	3,241	1%	1,703	2%
101–500	1,828	1%	1,192	1%
> 501	210	0%	360	0%
Any appearance	321,587	100%	88,500	100%

Note. "Any appearance" is the grand total row (summing over the column).

Regarding law firms, the landscape varies across different natures of suit, as shown in Table 8. Certain types of litigation align with a more traditional separation between the plaintiff and defense bars. In Property Rights cases, Fish & Richardson, a large law firm, is second to the Liebowitz Law Firm, a tiny law firm. Other types of cases have only the largest law firms being the most common law firms when combining all parties together. For instance, the top three law firms representing parties in litigation involving "Other Statutes" are all behemoths: Jones Day, Kirkland & Ellis, and Arnold & Porter. [100]

[100] "Other Statutes" is a catch-all category for natures of suit that are not common enough to have their own majority category (such as "Contract" or "Real Property"). They include an eclectic mix of case types involving, for example, Antitrust, Agricultural Acts, Freedom of Information Act, Arbitration, and Constitutionality of State Statutes. For a complete list, see *Nature of Suit*, *supra* note 96, at 3–4.

TABLE 8: TOP LAW FIRMS BY NUMBER OF APPEARANCES

Nature of Suit	Rank 1	Rank 2	Rank 3
Personal Injury	Shook, Hardy & Bacon LLP (11,939)	Faegre Drinker Biddle & Reath LLP (8,952)	Thomas Combs & Spann PLLC (7,742)
Civil Rights	Littler Mendelson P.C. (2,429)	Jackson Lewis P.C. (2,363)	Ogletree Deakins (2,332)
Other Statutes	Jones Day (1,616)	Kirkland & Ellis LLP (1,439)	Arnold & Porter Kaye Scholer LLP (1,429)
Labor	Ogletree Deakins (1,569)	Littler Mendelson P.C. (1,309)	Jackson Lewis P.C. (1,283)
Social Security	Law Offices Lawrence D. Rohlfing (1,416)	Olinsky Law Group (1,006)	Law Offices Kenneth Hiller (944)
Contract	Pandit Law Firm (783)	Lewis Brisbois Bisgaard & Smith LLP (607)	Thompson, Coe, Cousins & Irons, LLP (482)
Real Property	Akerman LLP (747)	Locke Lord LLP (674)	Wright, Finlay & Zak, LLP (418)
Property Rights	Liebowitz Law Firm (596)	Fish & Richardson P.C. (503)	Doniger / Burroughs (492)
Habeas Corpus	Cassiday Schade LLP (612)	Hale Law (491)	Jason Owens Law Firm (177)
Criminal	Law Offices Rolando D. Cantu (345)	Jones, Galligan, Key & Lozano (325)	Enoch Tarver Law (278)
Personal Property	Sullivan & Cromwell LLP (234)	Knight Law Group (187)	Cozen O'Connor P.C. (184)
Immigration	Law Offices Taobo Zheng Esq. (169)	Law Offices Aileen Shao (93)	Yerman and Jia (80)

Prisoner Petitions	Cohen Williams LLP (113)	Terpening Law PLLC (45)	The Castaneda Law Firm (26)
Bankruptcy	Blank Rome LLP (59)	Young Conaway Stargatt &Taylor, LLP (55)	Kirkland & Ellis LLP (54)
Civil Detainee	Heyl Royster Voelker & Allen P.C. (63)	Puget Law Group (42)	Cassiday Schade LLP (37)
Forfeiture/Penalty	Venable LLP (27)	Bird, Marella, Boxer, Wolpert, Nessim, Drooks, Lincenberg & Rhow, LLP (15)	Ray and Wood (12)
Federal Tax Suits	McCarthy & Holthus, LLP (13)	Meadows, Collier, Reed, Cousins, Crouch & Ungerman, L.L.P. (12)	Fidelity National Law Group (10)
All cases combined	Shook, Hardy & Bacon LLP (12,393)	Faegre Drinker Biddle & Reath LLP (9,881)	Butler Snow LLP (8,098)

Note. The number of appearances is in parentheses. For example, "Faegre Drinker Biddle & Reath LLP (9,881)" means that the law firm Faegre Drinker Biddle & Reath LLP appeared in 9,881 lawsuits.

H. Case Study on Access to Justice

We close this introduction to the SCALES project with a brief case study on access to justice. This study originally appeared in *Science* in 2020.[101] In that study, we asked a simple question: what are the barriers to accessing the justice system for indigent litigants?

To address this question, we analyzed the rate at which judges granted indigent plaintiffs' requests to waive court filing fees. It costs about $400 to file a federal lawsuit.[102] Indigent plaintiffs can request a waiver of this fee by submitting a petition to appear in forma pauperis (IFP). There is, however, no uniform standard that judges use to determine whether to grant such a

[101] Pah et al., *supra* note 2.

[102] *Id.* at 135.

petition.[103] It is also unknown how many such petitions are filed, how many are granted, and why some are granted while others are denied.

We analyzed dockets to determine whether a party submitted a fee waiver request, and whether that request, if ruled upon, was granted or denied.[104] We then analyzed all petitions that were granted or denied to compute the grant rate of each federal judge in 2016. Because cases are assumed to be assigned to judges at random within a district, if judges used the same standard to assess the merits of IFP petitions, one would expect a roughly uniform grant rate across judges within the same district.

We found substantial variation among judges from the same district in their IFP grant rates, however—much more than would be expected by chance (Figure 6). At the most extreme, in one district, judges' grant rates ranged from only 20% at the low end to 80% at the highest.[105] In such districts, whether an indigent litigant must pay to access the courts seems to mostly come down to the luck of the draw—that is, the chance assignment to a particular judge.

With SCALES data, we were able to uncover such patterns, diagnose problems, and develop effective policy interventions. Indeed, the IFP study described above led at least one federal district court to reexamine its IFP-related procedures, with the aim of developing a uniform standard of review.[106] Without the granular, process-level data from PACER, this sort of analysis and policy change would not have been possible. It is our intention

---

[103] *See* Andrew Hammond, *Pleading Poverty in Federal Court*, 128 YALE L.J. 1478, 1481 (2019); *see, e.g.*, Glenn S. Koppel, *Toward a New Federalism in State Civil Justice: Developing a Uniform Code of State Civil Procedure Through a Collaborative Rule-Making Process*, 58 VAND. L. REV. 1167, 1182–83 (2005) ("Inter-federal district court disuniformity complicates federal litigation, increasing cost and delay in the administration of civil justice. Many legal scholars have criticized inter-federal district court disuniformity in the realm of discovery, which 'is a practice that affects substantive rights and litigation outcomes.' . . . Critics have described contemporary federal procedure as 'impossibly arcane[]' . . . . They also assert that such rules give a tactical advantage to the local 'cognoscenti' over the outside practitioner and to the 'expert litigator over the lawyer making episodic appearances in court.' Other scholars have observed that localism increases the cost of legal services by requiring out-of-district litigants to retain local counsel, restricting competition for legal services. Local procedure has also been criticized for 'complicat[ing] federal practice . . . .'"); Samuel P. Jordan, *Local Rules and the Limits of Trans-Territorial Procedure*, 52 WM. & MARY L. REV. 415, 418 (2010) (noting the court-level procedural variation represented by local rules).

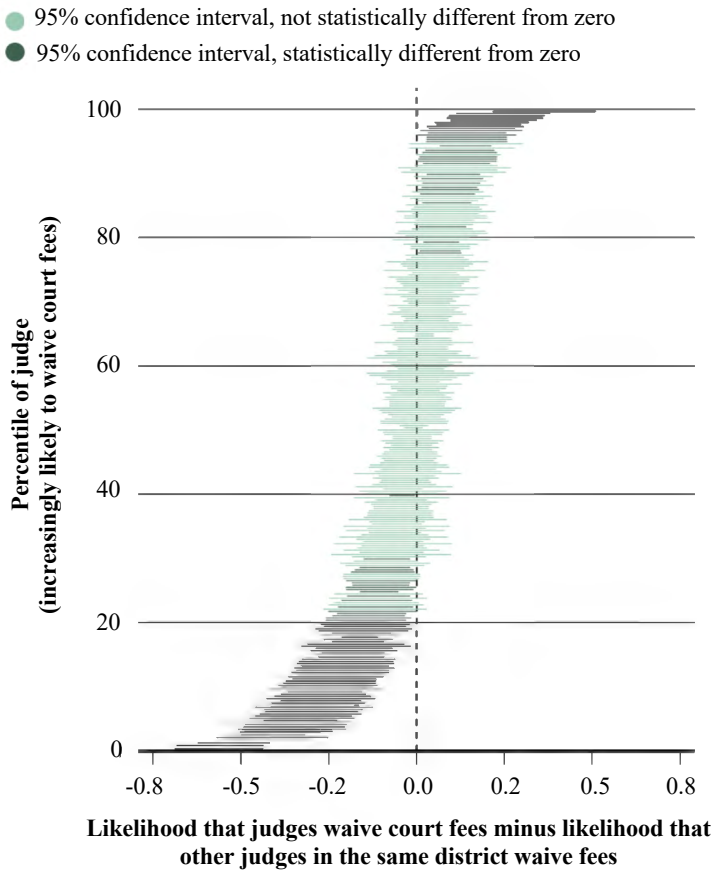
[104] At times, a judge dismissed the case without issuing a ruling on the fee waiver request if the fee waiver request itself was deemed deficient. For example, the District of Connecticut requires prisoner petitioners to include in a fee waiver petition a ledger of their transactions over the previous six months. If the petition lacked this information, the judge provided the prisoner a set period of time to correct the petition. If the prisoner did not respond, the judge dismissed the case without prejudice. *See, e.g.*, Banks v. Song, No. 3:17-cv-01179 (D. Conn. Aug. 25, 2017); Young v. Tarascio, No. 3:17-cv-01481 (D. Conn. May 9, 2024).

[105] Pah et al., *supra* note 2, at 135.

[106] This was communicated to us in a private conversation with a federal judge.

and hope that researchers will use the SCALES platform to produce studies like this one to improve the courts.

FIGURE 6: INCONSISTENCY IN JUDICIAL FEE WAIVER DECISIONS



Note. Litigants filed 34,001 applications to waive court fees in U.S. federal courts in 2016. For visual simplification, we show only the 294 judges (out of 1,742 total) who ruled on at least 35 applications. We would expect 5% of judges to differ from their within-district peers at 95% confidence. Instead, we find that nearly 40% of judges differ.

CONCLUSION

This Essay articulates the mission of SCALES: to enable the public to access and analyze federal court records. Federal court records have been online for nearly a quarter century. Yet they remain outside public reach because the government charges prohibitively high rates to access the data and because the data is composed of unorganized documents that are difficult

to interpret. SCALES eliminates these barriers by providing a comprehensive, organized, and freely accessible database of court records. Our goal is to democratize legal information, promote transparency, and enable empirical research on the judicial system.

SCALES is an ongoing endeavor. We continue to add new records and refine our tools. More importantly, we designed SCALES to enable others to add content and improve its tools. This Essay, therefore, serves as an empirical analysis, a research agenda, and an invitation for collaborative engagement. Our hope is that, together with the insights from *Northwestern University Law Review*'s Symposium—"Data Justice: How Innovative Data Is Transforming the Law"—the SCALES initiative will serve as a cornerstone to enable others to advance rigorous and careful legal research and judicial transparency.