NIPFP Working Paper Series

Problems with the e-Courts data

No. 314
29-July-2020
Devendra Damle and Tushar Anand

National Institute of Public Finance and Policy
New Delhi

Problems with the e-Courts data

Devendra Damle Tushar Anand*

Abstract

The creation of the e-Courts platform for disseminating data from the subordinate judiciary was an important step in making Indian courts more transparent. This platform has also prompted an interest in data-driven research on courts. While the e-Courts platform is a major reform in itself, there are numerous obstacles in successfully using this data for research. Previous work has pointed out that the data has standardisation issues, particularly in case-type nomenclature. It has also been shown that other fields, such as statute names and section numbers, are missing in some cases. In this paper, we quantify these error rates, which have so far only been known to exist anecdotally. We also identify new issues with the data, notably issues with wrong data being entered in certain fields. We report and quantify problems with mismatches between case-types and statute names, missing and malformed data in the statute name, section number, and date-time fields. We also show variations in error rates across states. The Indian Supreme Court eCommittee has taken cognisance of and initiated interventions to address some of these issues. However, the fundamental cause of bad quality data, viz. the lack of systematic data quality reviews and capacity building for the same does not seem to be part of the committee's plans. Until these quality issues are addressed, the use of this data for research will be limited.

* The authors are researchers at National Institute of Public Finance and Policy.

Acknowledgments

The authors thank Omidyar Network India for their financial support for this research. The authors would like to thank Prof Ila Patnaik, and Prof Ajay Shah from the National Institute of Public Finance and Policy for guiding this work, and for their feedback. The authors are grateful to Surya Prakash B S, from DAKSH, for useful discussions and insightful comments. The authors thank Sudhir Krishnaswamy of National Law University, Delhi and Varsha Aithala of Azim Premji University for useful discussions. The authors also thank all members of the Omidyar Network's Property Rights Research Consortium for comments and feedback.

All errors remain ours.

Contents

Acronyms

AC-Act	Arbitration and Conciliation Act, 1996.
CIS 3.0	Case Information System 3.0.
Commercial Suits Dataset	.
Contractual Disputes Dataset	.
CPC	Code of Civil Procedure.
CrPC	Code of Criminal Procedure.
EODB	Ease of Doing Business.
IC-Act	Indian Contract Act, 1872.
NI-Act	Negotiable Instruments Act, 1881.
NJDG	National Judicial Data Grid.
SR-Act	Specific Relief Act, 1963.
TP-Act	Transfer of Property Act, 1882.

List of Tables

1 Introduction

The judicial system performs the critical role of enforcing the rule of law and protecting the fundamental rights of citizens. Judicial functions must be effective and efficient to maintain order and security, reduce chances of corruption, ensure effective regulatory enforcement, and uphold civil and criminal justice principles. According to the National Judicial Data Grid (NJDG), 32.7 million cases were pending across Indian district courts as of June 2020. Close to 40% of these cases have been pending for more than 3 years. Out of the 98.6 million cases recorded as disposed, 26% were disposed after 3 years.

Court efficiency in India has been a topic of great interest in recent years. The *Economic Survey 2017–18* notes that the next challenge in Ease of Doing Business (EODB) in India is to improve the country's contract enforcement regime. This is also reflected in India's poor ranking in the *Enforcing Contracts* index of the World Bank's EODB rankings. While this measures the efficiency of the contract enforcement regime, it does so only for two cities — Delhi and Mumbai. Moreover, since the measure is meant to compare countries, it does not give useful insights on where each state falls short and what it can do to improve. We began this exercise with the idea of studying the contract enforcement landscape across the states in India. Our focus was on analysing litigation in the subordinate judiciary since that is where most contract disputes originate.

To that end, we planned on using data from the e-Courts platform. E-Courts is a pan India project to integrate technology in the judicial system. Among other things, it aims to provide district wise information about court proceedings. Under phase-II of the project, Rs. 9.21 billion have been released for this purpose and over 16000 courts have been ICT enabled.[1] Using the e-Courts platform, we collected data for cases filed under the following Acts:

1. Arbitration and Conciliation Act, 1996 (AC-Act);
2. Indian Contract Act, 1872 (IC-Act)
3. Negotiable Instruments Act, 1881 (NI-Act);
4. Specific Relief Act, 1963 (SR-Act); and
5. Transfer of Property Act, 1882 (TP-Act).

We assumed that cases filed under these Acts would represent a significant majority of contract disputes in subordinate courts. We scraped case-level data from two sample districts in each state. We picked the districts which have high court benches. If there was only one high court bench in the state, we picked the two most populous districts. The rationale behind this sampling strategy was to avoid data availability problems that might affect more remote districts. These districts are also likely to have the greatest business activity in the state and, as a consequence, the bulk of contract disputes. Our objective was to understand the caseload under each Act, the duration required for resolving cases, the time taken at each stage, and the pendency for each case type.

We were unable to create a data-set due to issues with the data, which rendered it unusable for the analysis we had envisaged. Our finding was not novel. While several judges (Chandrachud (2020) and MB Lokur (2017)) and commentators (Rao (2020), Krishnaswamy, K Sivakumar, and Bail (2014), and Verma (2018)) have applauded and relied on the e-Courts system, others have written about the problems with data. DAKSH (2017) notes that there is a lack of standardisation of case types across states, making a comparison of individual case-types difficult across states. The Law Commission of India (2014) has faced similar challenges while attempting to

[1] DoJ, GoI 2018.

analyse the case load across states. The Supreme Court's eCommittee recognises this problem and mentions it as a motivation for the standard case-type master list of the National Judicial Data Grid in its report on *eCourts Project Phase II: Objectives Accomplishment Report As per Policy Action Plan Document*. However, it notes that this mapping of states' case types to a national master list does not capture all the case-types in each state.[2] Finally, DAKSH (2017) also notes that the case type fields are sometimes incorrectly populated with the names of the statutes. The eCommittee has not explicitly recognised this issue.

Data systems in India continue to be marred by loopholes in collection and dissemination. This starts with variations in the definitions and methods of estimation and includes aggregation errors, lack of verification and validation, and other inconsistencies. This is a cause of worry. Data helps in the process of decision making, setting and prioritising goals, and monitoring progress. It can be used to define needs and plan interventions. Continuous improvement hinges on continuous data-based decision making. This is not possible without a functional data-set.

In addition to these issues, we find several issues with the e-Courts data. First, some states do not report any cases under certain statutes. For example, Karnataka and Tamil Nadu report no cases under SR-Act. This is a result of the variance between what each state considers *a case*. Contract disputes in these states are filed under provisions of the Code of Civil Procedure (CPC), without any reference to the substantive law under which the dispute falls. As an example, an Interlocutory Application or an Injunction Application is treated as a case in itself, without any reference to the Act under which the main dispute lies. Similarly, we also find that the number of cases under the IC-Act, and TP-Act are few in number across states. This is unexpected. We suspect that the reason for this is the same as above, i.e. varying definitions of what constitutes a case. A related issue is that linked matters are not identified in a majority of cases.

Further, we also find cases where the name of the statute entered in the Act Name field contains names that do not make sense or contain the name of a procedure. For example, in our data set of Commercial Suits, we found cases tagged under the Domestic Violence Act filed under Code of Criminal Procedure (CrPC), and cases where the statute name is reported as "Recovery of Money".

While DAKSH (2017) identifies many of the major issues with the e-Courts data, it does not quantify the scale of the problem. The *eCourts Project Phase II: Objectives Accomplishment Report As per Policy Action Plan Document* also does not report the scale of issues with the data. This is particularly salient because while these issues may be addressed for newer cases, there are millions of cases in the existing data set that will need to be remedied. Getting an estimate of the quality of the data is imperative in deciding how it can (or cannot) be used for research. As we show below, e-Courts data should be used cautiously. An over-reliance may lead to false pictures and misguided policies. In this paper, we add to the literature on issues with the e-Courts data by identifying some new issues and quantifying the scale of the known issues. We show the variation in the issues and error rates across states. Finally, we provide recommendations for what the Supreme Court eCommittee can do to improve the data quality and make it more amenable to robust research.

2 Methodology

We have used publicly available case-level data from the e-Courts platform maintained by the National Informatics Centre (NIC). The e-Courts platform is an e-governance portal containing

[2] See Annexure D of eCommittee, Supreme Court of India 2019a.

case-level judicial data for over 6 million cases filed in district courts across most Indian states. Although preparatory work for the digitisation of case records started in 2007, the public portal went live in 2014. Since the e-Courts platform was designed to track cases online, it records several data points pertinent to each case's life-cycle. The date of filing, date of each hearing, and date of disposal are recorded. The purpose of each hearing is also recorded. The interim and final orders can also be accessed through the platform. It also includes case-level meta-data and a 2-3 word summary of the business conducted at each hearing. The meta-data includes details such as the Act name and section under which the suit is filed, the name of the court complex, names of litigants, and the dates of hearing. The platform can be used to extract data for pending cases, as well as those that have been disposed.

2.1 Data description

The platform delineates cases based on the case type and the Act under which they were filed. Information on sections under specific Acts is also available. There are 90 variables for each case in the raw data. These variables consist of unique identifiers, geographic and hearing details, and timestamps for each case's life-cycle. There are also certain variables such as those relating to transfer details of a case or linked matters which are only applicable for a fraction of the cases and are not used otherwise.

2.2 Sampling strategy and data collection

In every state, two districts were selected: (i) the district which contains the capital city, and (ii) one other district where a high court bench is present. If there is no high court bench in the state, we selected the districts with the highest population in the state. After identifying the two districts, we scraped data from all subordinate courts (i.e. Munsif, Metropolitan Magistrates, District Courts, Sessions Courts, etc.), which are present on the e-Courts platform.

We created two sets of case-level data. The first is a data-set of cases related to contract disputes (based on the statute name). This data-set is the one we planned to use for our original study. The second is a random sample of 10 cases tagged as 'Commercial Suits' from each district for each year of cases.

2.2.1 Contract disputes data-set

For our data-set on contract enforcement litigation in district courts, we scraped case-level data for pending and disposed cases filed under the five laws mentioned in section 1 i.e. (i) Arbitration and Conciliation Act, 1996, (ii) Indian Contract Act, 1872, (iii) Negotiable Instruments Act, 1881, (iv) Specific Relief Act, 1963, and (v) Transfer of Property Act, 1882.

The rationale was that these laws cover a majority of contract disputes in the sample districts. This data collection exercise gave us a total of 10.2 lakh cases across all states in the country. We used this to quantify the number of cases filed under each statute. We also and measured how many cases in this set have missing, erroneous, or malformed entries in the fields for date of filing, first and last listing, decision, and whether the final order/judgment is available (for disposed cases only).

Lastly, we also examine the trends in these variables over time and across states to check if these have affected the data quality. For brevity, we will refer to this data-set as the Contractual Disputes Dataset.

2.2.2 Commercial suits data-set

As mentioned, the platform allows for querying data by case-types and Act names.[3] We built the first data-set using Act name queries for the specified Acts. This gives insights into the problems with identifying the subject matter of a case based on the Act name. However, there may be issues with the Act name field itself. This second data-set was thus intended to assess the issues with the entries in this field. We picked cases filed under the "Civil Suit: Commercial" case type across court complexes in the same districts as the earlier data-set. We picked 10 random cases from each year since 2010 in each court complex where this case type is available. This gave us a sample of 1368 cases.

In this set, we examine the Act name and the section number field. Ideally, these fields should have the name and applicable section number(s) of both the procedural law – in this scenario, the CPC – and the substantive law (e.g. SR-Act). We use this to understand how many cases are tagged with the applicable substantive law. We also check for obvious errors, such as missing values, irrelevant Act names, or entries that are not names of any statute. In the section number field, we check for missing values, entries of something other than a section number, or malformed entries. For brevity, we will refer to this data-set as the Commercial Suits Dataset.

3 Findings

In this section, we describe the results of our error-checking. We have divided the analysis thematically based on the issue or the field of information under question.

3.1 Data availability

The e-Courts project was implemented in phases from 2008 to 2014. Districts and court complexes were on-boarded in phases, which means that the data does not accurately reflect the total stock of cases consistently across years and geographies. Also, the full stock of cases has not yet been digitised. Case details are only recorded and reported if the case had a hearing after a given court complex was added onto the platform. If a court complex was brought onto the platform in 2014, we would have information on a case filed in 2010 only if it had a hearing in or after 2014. Any cases concluded before 2014 at such a complex are not reflected in the data. Since the implementation across court complexes was completed in 2014, we see data more consistently from 2014. This can be seen in table 2.[4]

Year of filing	Pending cases	Disposed cases
<2009	116647	43560
2010	20764	1530
2011	19397	1876
2012	22915	3311
2013	33737	6892
2014	51447	12113
2015	65506	23448
2016	76716	40358
2017	67605	65068
2018	43515	104522

[3] Some examples of case types are Execution Petition, Interlocutory Application, Arbitration Petition, etc.
[4] The data for 2019 is only till October.

Year of filing	Pending cases	Disposed cases
2019	27525	172413
Total	545774	475091

Table 2: Number of cases filed across years

3.2 Issues with case identification by statute

There is considerable variance in how data is recorded across court complexes and states on the platform. For example, cases under the AC-Act are filed under 58 unique names, while those under NI-Act are filed under 21 unique names. The raw data records a total of 844 unique case types. This is the case even though the NJDG reports only 31 case types. Cases could be tagged under a procedural law, substantive law, or both. Non-uniformity in tagging cases means that there is a high likelihood of a search for cases under an Act would contain an incomplete set of cases for a given court complex.

This inconsistency in tagging of cases results in incomplete information about the actual litigation under a given statute. Table 3 shows the distribution of cases under each statute in our data-set. Certain states show inconsistencies in the data. For example, Tamil Nadu does not report any cases under any Act apart from NI-Act. It is highly unlikely that litigation concerning SR-Act is not filed in Tamil Nadu. Instead, it appears that these cases are filed under the procedural law i.e. the CPC and not under the substantive law, and hence cannot be traced back to the Act names.

State	AC-Act	IC-Act	NI-Act	SR-Act	TP-Act
Andhra Pradesh	2212	2	46845	44	77
Assam	—	—	19180	—	—
Bihar	4152	70	23658	107	11
Chandigarh	2994	—	16410	480	1
Chhattisgarh	6273	—	36987	2589	—
DNH at Silvasa	56	—	2072	9	—
Delhi	30122	—	100172	1757	2
Goa	1240	27	41734	2686	5
Gujarat	508	327	67809	10755	1
Haryana	4624	1	23631	312	1
Himachal Pradesh	1205	1	20360	1886	8
Jammu and Kashmir	72	—	10435	6	3
Jharkhand	661	—	31377	34	1
Karnataka	6575	—	46597	—	—
Kerala	15596	3	22636	424	21
Madhya Pradesh	9596	167	22989	8516	61
Maharashtra	2793	—	66210	903	1443
Manipur	18	209	556	1647	29
Meghalaya	251	—	398	46	—
Mizoram	1	—	21	—	—
Orissa	1932	1	2229	3213	12
Punjab	8495	3	43270	5812	4
Rajasthan	22938	12	12705	60	3
Sikkim	26	—	335	48	1
Tamil Nadu	—	—	17635	—	—
Telangana	5472	20	30988	1525	86
Tripura	253	3	2017	2080	2
Uttar Pradesh	3307	3	29721	7	—
Uttarakhand	1412	1	37819	5	—
West Bengal	17245	10	19907	26552	12

State	AC-Act	IC-Act	NI-Act	SR-Act	TP-Act

Table 3: Distribution of cases across Acts in each state

Theoretically, the IC-Act would govern all contract disputes. However, as table 3 shows, most states do not tag cases under the Act. Furthermore, any contract dispute in which a particular relief is sought both the IC-Act and SR-Act would be necessary. But as seen, the SR-Act too is often not tagged. Similarly, we find that there are very few cases filed under TP-Act, despite it being known that a majority of litigation in the courts is related to land and property.

This issue results from a variation in what each state defines as a case. During the course of litigation, parties file Interlocutory and Interim Applications, Injunction Petitions, Execution Petitions, and other motions and applications under the CPC. These can be considered the components of the process of litigating a contract dispute. Some states treat these as cases in themselves. Some record what is known as the *main matter* and file it under the substantive law, and the procedural matters as linked matters, tagged under the CPC. Some states tag the procedural matters under both the substantive and procedural law. Moreover, some, like Tamil Nadu and Karnataka, do not tag certain cases with the substantive law at all.

States also follow varying protocols for different kinds of disputes. Karnataka, for instance, files contract disputes based on the applicable Order number of the CPC but does not do so for all civil litigation. To wit, it records arbitration-related matters under the AC-Act. However, it files many of the matters falling under the SR-Act under Order XXXVII of the CPC, which is the provision for summary hearings of disputes. This implies that disputes cannot be identified reliably by Act name in many states, especially in contract disputes.

3.3 Issues in reporting statute name

In our conversations with practising advocates, we were informed that case-type is a fairly reliable identifier of cases within any given state. While cases may not be reliably queried under a particular statute, they may be identified under given case types. Therefore, to confirm our hypothesis of cases being filed under procedural rather than substantive laws, we look at the Act names reported for cases in which the case type is reported as "Commercial Suit" or "Civil suit: Commercial", i.e. our Commercial Suits Dataset.

As table 4 shows, we observe a considerable variance in reporting statute names when cases are queried by case type. As expected, we find that a majority of the cases are tagged only under the CPC. However, out of the total sample of 1368 cases, 126 cases have missing values in the Act Name field. We also find cases under the Protection of Women From Domestic Violence Act, 2005, Juvenile Justice Act, 2000, and the CrPC despite the case type being "commercial suits". This is either an error in the case type or statute name. However, given our experience with the inconsistency in recording statute names described in the previous sub-section, we surmise that the errors here are in the latter.

Act	Count	Percent
Code of Civil Procedure	425	34.21
Recovery of Money	202	16.26
Arbitration and Conciliation Act	145	11.67
Specific Relief Act	94	7.56
Code of Criminal Procedure	82	6.60
Commercial Courts Act	79	6.36

Act	Count	Percent
Labour laws	41	3.30
Domestic Violence Act	35	2.81
HP State Labour Laws	35	2.81
Trade Marks Act	27	2.17
Minimum Wages Act	18	1.44
Permanent Injunction	16	1.28
Indian Contract Act	11	0.88
Limitation Act	9	0.72
Copyright Act	8	0.64
All India Services Act	7	0.56
Hindu Marriage Act	7	0.56
Juvenile Justice Act	1	0.08
Others	43	3.46

Table 4: Act-wise distribution of cases filed under the Civil Suit- Commercial case type

Out of a total of 1,242 cases that reported the Act names in this sample, 1,209 were tagged under a single law. Out of these, more than one-third were tagged under a procedural law. Table 5 reports the values for our sample.

Description	Count	Percent
Cases with Act names	1,242	90.78
Cases tagged only under CPC	402	29.38
Cases tagged only under CrPC	82	5.99
Cases with missing Act name	126	9.21
Total number of cases	1368	–

Table 5: Errors in reporting Act names

3.4 Issues with recording the section number

In addition to issues in reporting the statute name, we also find errors in reporting section numbers in our Commercial Suits Dataset. The section numbers are supposed to be strings of either only digits or digits followed by an alphabet (for example 25, 25(e), or 25E are all acceptable values). For cases filed under CPC, Orders and Rules are reported as "Order" or "Rule", followed by the number either in Roman numerals (e.g. Order XXXVII) or Arabic numerals (e.g. Rule 10). Both these are also acceptable. Table 6 shows the count of different kinds of errors in recording section numbers. Broadly, these errors can be classified as:

Case type mentioned in the section field: Certain cases do not have a section value, but the case type value, such as "civil suit" or "commercial suit". This is an error in tagging the section names.

Section names contain another Act name: Certain section names contain another Act name in their section column. This is especially seen when the case is filed under a state law, such as Himachal Pradesh State Labour Law or a state amendment to the Shops and Establishments Act.

Section names are not clearly delimited: When cases are filed under separate sections of

the same Act, they are usually delimited by a comma. However, in certain cases, this delimitation is not present, which provides an absurd value for the section names. For example, if the actual section values are "17,18,19" for a certain Act, they are reported as "171819" under the section column.

Missing values in section names: Certain cases report null values for section columns.

Description	Value	Percent
Missing values for Act field	126	9.21
Sections properly defined	980	71.63
Case type in the section field	124	9.06
Act name and section name given	91	6.65
Improper delimitation	43	3.14
Missing section names	37	2.70

Table 6: Types of errors in reporting the section names

3.5 Missing data for final orders

Even when the Act names and section numbers are available, the links for final orders are missing. In more than 70% of cases in our Contractual Disputes Dataset, final orders are not available. Given their importance for litigants and other users of the data, this is a serious issue. Table 7 shows the percentage of cases with final orders and how their availability has changed.

Year of filing	Cases with final orders	Total disposed cases	Percentage
<2009	20021	116647	17.16
2010	5133	20764	24.72
2011	4574	19397	23.58
2012	5491	22915	23.96
2013	9142	33737	27.10
2014	14559	51447	28.30
2015	17514	65506	26.74
2016	25109	76716	32.73
2017	27409	67605	40.54
2018	17830	43515	40.97
2019	10744	27525	39.03
Total	157526	545774	28.86

Table 7: Percentage of cases with links to final orders over time

The performance of courts on this front has improved slightly over time. However, the number is still low. This is the case even when the total disposed cases do not represent the entire universe of cases decided by district courts. If a case was decided before a court was on-boarded, that case would not be counted in disposed cases. Hence, the 70% absence of orders is in addition to those cases where no information (including orders) is available.

However, across geographies, we find that certain states are more consistent in uploading final orders. Table 8 shows the percentage of cases with links to final orders across different states in the sample.

State	Cases with final orders	Total disposed cases	Percentage
Andhra Pradesh	5783	42307	13.67
Assam	173	18852	0.92
Bihar	981	7032	13.95
Chandigarh	380	550	69.09
Chhattisgarh	1538	15801	9.73
DNH at Silvasa	849	1456	58.31
Delhi	34423	85694	40.17
Goa	4081	31628	12.90
Gujarat	13579	44382	30.60
Haryana	5851	6535	89.53
Himachal Pradesh	2225	9698	22.94
Jammu and Kashmir	813	3406	23.87
Jharkhand	1320	18620	7.09
Karnataka	33838	35356	95.71
Kerala	6265	32413	19.33
Madhya Pradesh	1001	16669	6.01
Maharashtra	4013	15126	26.53
Manipur	175	1402	12.48
Meghalaya	25	370	6.76
Orissa	552	2889	19.11
Punjab	26607	41993	63.36
Rajasthan	432	24027	1.80
Sikkim	326	360	90.56
Tamil Nadu	166	6376	2.60
Telangana	2530	23745	10.65
Tripura	2436	2473	98.50
Uttar Pradesh	3698	21389	17.29
Uttarakhand	515	11982	4.30
West Bengal	2989	23350	12.80

Table 8: State wise percentage of cases with links to final orders

States like Haryana, Karnataka, Sikkim, and Tripura are more consistent in uploading judgments of disposed cases. In contrast, Uttar Pradesh, Tamil Nadu, and Uttarakhand do not upload orders as consistently. Interestingly, even though Haryana and Punjab have a common High Court that monitors their performance on the platform, there is a marked difference in the availability of orders. This means that district courts in certain states have better case management systems to upload their judgments. Larger states such as Delhi and Maharashtra also fare poorly. One possible reason could be the caseload handled by particular court complexes. The data shows that states which have a relatively lower caseload upload final orders more frequently.

3.6 Missing date fields

The platform reports four dates of interest for each case: (i) the date of filing, (ii) the date of first hearing, (iii) date of last hearing, and (iv) the date of the decision. We found a significant number of cases with missing data in these fields. These dates are important for calculating indicators related duration of the case. Table 9 shows the number of cases in our Contractual Disputes Dataset which have missing data in important date fields.

Variable	Missing values	Percentage
Date of filing	39216	3.84
Date of first hearing	48476	4.75
Date of last hearing	44979	4.40
Date of decision	2797	0.51
Total	135468	13.26

Table 9: Missing values in the data-set

In addition to these missing date fields, we also found that the entire history of hearings was missing in 69305 (6.78%) of cases in our sample. There is considerable variation between states concerning the proportion of cases with missing data in these fields. We delineate these differences in the following subsections.

3.6.1 Missing values for date of filing

Table 10 shows the number of cases with missing values for the date of filing across states. The error rates are concentrated in states such as Assam and Gujarat, while other states perform better. Again, it is interesting that though Goa and Maharashtra report to the same High Court (High Court at Bombay), there is a marked difference in their performance.

State	Cases with missing date of filing	Number of cases	Percentage
Andhra Pradesh	2504	49180	5.09
Assam	10792	19180	56.26
Bihar	4	27998	0.01
Chhattisgarh	50	45849	0.11
Goa	5969	45692	13.06
Gujarat	18352	79400	23.11
Himachal Pradesh	577	23460	2.46
Karnataka	374	53172	0.70
Kerala	579	38680	1.50
Maharashtra	2	71349	0.00
Orissa	1	7387	0.01
Tamil Nadu	5	17635	0.03
Telangana	4	38091	0.01
Tripura	1	4355	0.02
Uttar Pradesh	2	33038	0.00

Table 10: Missing values for date of filing across states

In the remaining states, we found no cases with missing values in the date of filing field.

3.6.2 Missing values for date of first hearing

Table 11 reports the missing values for the date of first hearing across states. The error rates are again concentrated in similar states i.e. Gujarat, Assam, and Goa.

State	Number of cases	Cases with missing date of first hearing	Percentage
Andhra Pradesh	49180	2637	5.36
Assam	19180	10802	56.32
Bihar	27998	136	0.49
Chandigarh	19885	24	0.12
Chhattisgarh	45849	143	0.31
DNH at Silvasa	2137	1	0.047
Delhi	132053	104	0.079
Goa	45692	6271	13.72
Gujarat	79400	19522	24.59
Haryana	28569	18	0.06
Himachal Pradesh	23460	659	2.81
Jammu and Kashmir	10516	71	0.68
Jharkhand	32073	63	0.20
Karnataka	53172	4227	7.95
Kerala	38680	883	2.28
Madhya Pradesh	41329	645	1.56
Maharashtra	71349	170	0.24
Manipur	2459	21	0.85
Meghalaya	695	1	0.14
Orissa	7387	336	4.55
Punjab	57584	72	0.13
Rajasthan	35718	699	1.95
Sikkim	410	6	1.46
Tamil Nadu	17635	76	0.43
Telangana	38091	339	0.89
Tripura	4355	96	2.20
Uttar Pradesh	33038	176	0.53
Uttarakhand	39237	264	0.67
West Bengal	63726	286	0.45

Table 11: Missing values for date of first hearing across states

3.6.3 Missing values for date of last hearing

Table 12 reports the missing values for the date of last hearing across states. The error rates are again concentrated in Gujarat, Assam, and Goa.

State	Cases with missing date of last hearing	Number of cases	Percentage
Andhra Pradesh	2504	49180	5.10
Assam	10825	19180	56.44
Bihar	151	27998	0.54
Chhattisgarh	72	45849	0.16
Delhi	103	132053	0.078
Goa	6280	45692	13.74
Gujarat	18700	79400	23.55
Haryana	1	28569	0.00
Himachal Pradesh	581	23460	2.47

State	Cases with missing date of last hearing	Number of cases	Percentage
Jammu and Kashmir	25	10516	0.23
Jharkhand	15	32073	0.046
Karnataka	2752	53172	5.17
Kerala	797	38680	2.06
Madhya Pradesh	640	41329	1.55
Maharashtra	63	71349	0.09
Manipur	3	2459	0.12
Orissa	299	7387	4.04
Punjab	7	57584	0.01
Rajasthan	685	35718	1.91
Telangana	19	38091	0.05
Uttar Pradesh	112	33038	0.34
Uttarakhand	180	39237	0.46
West Bengal	167	63726	0.26

Table 12: Missing values for date of last hearing across states

3.7 Other missing data

Some fields can be of great use but are rarely populated. One example is that of related matters. The data for each case contains a field for case-identification numbers of the subsidiary and main matters. But these fields are often not populated. Hence, the data does not allow us to explore the entire web of court-procedures that ensue after a suit is filed. To illustrate the point further, in a case filed under the SR-Act, parties may file numerous appeals, applications, and petitions for injunctions, all of which have to be heard and disposed of before the main matter. On the platform, irrespective of whether the individual state treats these as separate cases, these are treated as separate cases. The fact of them being related is not reported. So, while we may get somewhat reliable data about the duration of a matter, we do not truly understand what kind of procedural complexity each case goes through or the duration of a case in its entirety. This can also affect skewing estimates of case-duration downwards since these supplementary matters usually take less time than the main matter.

4 Interventions by the eCommittee

Many of these issues are well known (although the scale of the issues was previously unknown). The Supreme Court's eCommittee has also taken cognisance of many of these issues, and some others which we have not addressed in this paper.[5] Specifically (among other things), the eCommittee plans to introduce standardisation across the Case Information System 3.0 (CIS 3.0) (the new case-management software). This will include case-types, statute names, and notations for stages of a case and section numbers.

The *eCourts Phase II Accomplishments Report* report notes that some standardisation is already underway, although the exact extent of completion is not reported. The report even mentions the possibility of a public API for data access in the future. While these are all good measures, some issues have still not been addressed.

[5] eCommittee, Supreme Court of India 2018, See :

4.1 Gaps in eCommittee's interventions

Largely, there are five areas that the eCommittee's plans and interventions ignore:

1. Management of archival data;
2. Definition of what constitutes a case;
3. Process manuals for data quality control;
4. Data quality reviews; and
5. Institutional arrangements.

While the CIS 3.0 documentation talks about how case data will be better organised in the future, it does not mention what will be done with the millions of cases already in the system. The report on *Case Management through CIS 3.0*, which documents the features of the CIS 3.0 mentions in passing that archival data will be used for Big Data Analytics. However, it does not mention what will be done with the existing case data, which has not yet been entered into the system. This is compounded by the fact that the *eCourts Phase II Accomplishments Report* states that digitising existing case records is no longer within the scope of the e-Courts project. It also states that under the 14th Finance Commission, the budget for this activity is being given directly to the State Governments. Hence, it is unclear how the integration will take place.

The eCommittee's documentation also does not provide any solutions for the disparity between states on what constitutes a case. Standardising statute names will resolve the issue of alternate spellings, but it will not resolve the problem of cases not being tagged with substantive law. To resolve this issue, states have to reach a consensus on how to define a case, or the Supreme Court has to issue guidelines for the same. The committee has planned to introduce a new field in CIS 3.0 for the "subject matter of dispute", which could partially mitigate this problem. Finalising what constitutes a case also requires that fields of related matters are populated. The *eCourts Phase II Accomplishments Report* does not mention the progress on this parameter, nor does it mention how if there is any plan to resolve this in the future.

The design for CIS 3.0 will likely reduce instances of malformed entries in the statute name and section number fields. The *eCourts Phase II Accomplishments Report* report mentions that the erstwhile system of allowing free text entries for the statute name and section number field will be replaced by a drop-down menu. However, as we have seen, there are instances of the statute names and section number fields being left blank. This is likely because statute name and section number are "non-mandatory fields", according to the *Registration User Manual eCourt Information Systems*.[6] The change to a drop-down menu will not address this problem. It will also not address the incongruity between case-types and statutes (e.g. civil suits being tagged with statutes that deal with criminal offences). Further, the e-Courts manuals merely contain instructions on how to operate the software. To our knowledge, there are no process manuals for data management and quality control practices. There also does not seem to be any institutional arrangement for ensuring the system's quality of data. None of the courts seem to conduct data quality reviews either.

This feeds into a larger point of lack of systematic data quality control protocols. The eCommittee does not appear to have any systems for assessing and monitoring the quality of the data. The *eCourts Phase II Accomplishments Report* does not mention any data quality audits. The documentation also lacks any standard manuals for the data quality. The *National Policy and*

[6] Other non-mandatory fields are the names of petitioner and respondent, name of the subordinate court, and registration details of a case.

Action Plan for Implementation of Information and Communication Technology in the Indian Judiciary, 2005, which was the genesis of the e-Courts project, also does not mention data management or quality control.[7] This is supplemented by a dearth of technical staff capable of handling such a complex data management system. Documentation on the technical capacity (or the lack thereof) is sparse. Ideally, the state would build dedicated teams with specialised knowledge to collect data. It would improve upon processes that have been used previously and consequently create greater capability. However, the Bombay High Court, in its response to a questionnaire issued by the Chief Justices' Conference of 2016, reported that they have no technical staff at all in subordinate courts. Some existing staff who have demonstrated competency in the use of the system serve as master-trainers for staff members, courts, and even other states.[8]

The institutional arrangements for technical support for users of the e-Courts platform are also inefficient. All troubleshooting and technical support for users of the e-Courts system across the country is done by one National Informatics Centre team stationed in Pune.[9] This team is only responsible for providing support for the e-Courts platform. It is unclear what the arrangements are for technical support for the rest of the IT systems.

Together, these issues indicate a larger problem of the lack of technical capacity in designing, operating, and maintaining a complex data management system such as e-Courts. Data management on such a scale requires specialised knowledge and technical capacity. In addition to knowledge of how to operate the system, such systems need to establish protocols and manuals for data quality and integrity. These are aspects of management rather than the operation of the system, and they are currently missing from the eCommittee's thinking about e-Courts.

4.2 Lack of integration between internal systems and disseminated data

From a system design point of view, the issues of missing and wrong data we observe in our dataset arise out of the fact that there is a disconnect between the IT system the courts use internally for managing case-level data, and the system through which they disseminate the data. The data which is disseminated through the e-Courts platform for public consumption is entered into the system by data-entry-operators after the fact. As an example, if a case on a contractual dispute is heard today, the proceedings of the hearing are recorded in one system. The data that appears on the e-Courts platform regarding that hearing, is entered into it some time after the day's proceedings are done. The fact that the courts' system for recording information in the course of a working day and the system which reports these data (i.e. e-Courts) are two separate systems, is evident from the *Registration User Manual eCourt Information Systems* and the *Court User Manual eCourt Information Systems*. The *Registration User Manual eCourt Information Systems* states that *"Although all the fields are not mandatory, it is advisable to enter all the data at the time of Registration, so that correct and accurate management reports as required by the management can be generated in future."*. This indicates two things. First, it explains why we find so many instances with data missing. Second, it indicates that the data can be entered in the system well after the event where it is generated. This is further evidenced by the instructions for data entry by court clerks in the *Court User Manual eCourt Information Systems*.[10] It states that the evidence, filings, and proceedings of any hearing are first entered into a diary, and then subsequently entered into the e-Courts system. This means there are two parallel sets of information on any case, the paper-record and the digital one.

[7] The policy plan contains a lot of other details such as the design of a data-centre and choice of software for the digital library

[8] Bombay High Court 2017.

[9] See: Kaushik 2019.

[10] See Page 195: of NIC, Pune 2016a.

This lack of integration introduces the possibility for mismatches in the first place. Ideally, both these systems have to be part of one integrated data-management system. The data which is meant for dissemination should be created as a matter of course in the proceedings of the court, rather than requiring a separate data entry process. Each record must be entered in the system as and when it is generated. This record itself has to be digital, and should not require someone to enter it into the system separately, on the basis of a paper record, and after the fact.

While the lack of data quality reviews is an important issue, the lack of an integrated data-management system creates opportunities for these errors (of missing and erroneous data) to arise in the first place.

5 Conclusion

As shown, our experience of using the e-Courts data brought forth issues with:

Data availability: Data for cases that were filed before 2014 is not reliably available. Further, as we go back in time, the data for disposed cases gets sparse.

Identifying cases by statute: The field for statute name contains several errors and inconsistencies. Part of the problem is with the disparity between states on what is considered a case. Many cases are only tagged with procedural laws. This means we cannot, for example, reliably identify contract disputes purely based on the statute name. In our Commercial Suits Dataset, we found that most cases were filed solely under the CPC, i.e. procedural and not the substantive law.

Incongruity between statute names and case-type: We found instances of commercial suits, which are always purely civil, tagged with the Prevention of Domestic Violence Act, 2005 and the CrPC. These statutes deal with criminal suits, and should not appear as a commercial suit.

Names of procedures in statute name field: The field which is supposed to contain the name of the statute instead contained names of procedures, such as "Recovery of money" and "Permanent injunction".

Missing statute name: There were a significant number of cases where the statute name was blank. This is likely because the statute name field is non-mandatory, i.e. a case can be filed and processed without these data-points being entered.

Errors in the section number field: We also show that the section number field has a large number of missing or malformed values. In some cases, the field contains the name of the statute.

Missing final orders: Final orders/judgments are not available for a vast majority of the disposed cases.

Missing date fields: There are a significant number of cases where the fields for filing date, date of the first hearing, and last hearing are blank. Further, there also is a significant number of cases where the case hearing history is not available at all. This phenomenon, however, is state-specific. Some states such as Gujarat, Goa, Assam have more missing values, while states like Maharashtra, Delhi, Punjab, and Telangana perform better.

Related matters field not populated: Related matters are not populated in a vast majority of the cases.

As a result of these problems with the data, despite their availability, their utility for research is limited. If these errors in the data did not exist, it could open up several, interesting and important lines of inquiry into the functioning of the judiciary and its impact on the Indian economy and society.

5.1 Unrealised potential of e-Courts data

There are several potential research designs that one could imagine with the e-Courts data, if it did not have the flaws we have highlighted viz.:

1. **Case-loads under different statutes, and resources required to dispose of them**:
   Suits, under particular Acts cannot be studied, owing to the inconsistency in recording and errors in the statute name and section number fields. One cannot, for instance, study the volume of litigation that concerns contract disputes and the sections of laws that lead to those disputes. This also prevents researchers from examining problems where the solution may be legislative rather than related to court administration.

2. **Inter-state comparison of dispute resolution of various types**:
   Inter-state comparison is not possible with the current state of the data. The first issue is with differing definitions of what constitutes a suit, from state to state. The second is that there is great variation in the data quality and availability between states. Some states such as Maharashtra, Telangana, and Delhi seem to have better data than others like Goa, Gujarat, and Assam.

3. **Intra-state comparison of dispute resolution may be possible for only a few select states**:
   We limited our data collection to two most populous districts in each state, in hopes that the error rates would be within an acceptable range, and it would allow us to use the data. However, we find that that is not the case, with many states having high error rates despite us picking the more developed districts. We suspect that remote districts might have worse data, which means that an intra-state comparison is not likely not possible for most states.

4. **Pendency and disposal rates for years before 2014**:
   Since the entire stock of cases has not been digitised, and there are fewer older cases that have been disposed, the possibility for studies of durations of cases over time are also restricted.[11] Survival models for estimating likely durations of cases on the basis of the statute, nature of dispute, location etc. are not currently possible. This also limits studies on the causal linkages between various interventions on case-duration and on pendency rates.

5. **Studies on the substantive aspects of law, and the court's interpretation of statutes**:
   The data also cannot be used to study substantive aspects of law since the name of the substantive law governing a dispute is often not cited. The lack of availability of final orders/judgments for a majority of cases also means that one cannot study the finer aspects of the nature of litigation, such as the arguments made by either party, or the court's reasoning in arriving at its verdict.

6. **Study of the complexity of dispute-resolution**:
   The fields for linked matters is most often not populated for any case. This prevents

[11] In statistics parlance, this phenomenon is known as right and left censoring. In this particular scenario, it means that we do not have an accurate number of cases instituted for any given starting year. At the end of the observation period i.e. 2019, we still have some pending cases from the starting year.

anyone from mapping the true complexity of any litigation. As a result, we are limited to, at best, studying individual motions rather than suits in their entirety.

As it currently stands, the e-Courts data can primarily be used for state-level studies of certain case-types, for certain states and only in certain districts. Thus, researchers should be careful about selecting which state, case-type and district to study using these data. These issues affect not just academic research on judicial data but are a hurdle for the judiciary to create sophisticated data-driven decision support tools for judicial resource estimation.

6 Way forward

The e-Courts data, in its current form, has very limited utility for research. Our analysis raises questions on the conclusions that can be drawn from the data. However, the quality and utility of the data could be improved with certain targeted interventions.

The primary interventions needed here are:

- Create an integrated system for day-to-day court functioning and data dissemination;
- Conducting regular data quality reviews for existing and newly generated data;
- Creating process manuals for data-entry and validation; and
- Assigning roles and responsibilities to separate people for data-entry and data-validation.

The data dissemination system must not be seen as a separate system, but as an integral part of the court's data-management system. The data for dissemination should be generated as a matter of course in the court's functioning, rather than being entered separately. At the same time, even if this sort of integrated system were realised, data quality reviews are essential for making the data usable for robust research.

Some data-quality controls and validation issues can also be solved using technology. Missing fields, for instance, are easy to solve. The validation tool only has to ensure that the field is populated. Similarly, date fields where the date for a subsequent hearing is before a preceding one should get flagged. Another example would be to have a matching algorithm to flag an instance of an Act that does not match a case-type (such as an Act that governs criminal offences being assigned to civil cases). Another measure, as the eCommittee has recommended, could be to minimise the fields where free text can be entered, and providing multiple-choice options as far as possible. These technical solutions, however, are not a substitute for regular data quality reviews.

It is important to recognise that data management is just one aspect of court administration. The discussion about data management has to be situated in the context of modernising court-administration as a whole. It is a well-accepted idea that court administration in India needs an overhaul. Better data management has to be part of that larger reform. One idea for modernising court administration is to create a separate administration agency – an Indian Courts and Tribunals Service.[12]

One model that can be adopted is the one implemented by the Central Board of Direct Taxes (CBDT) to create the Tax Information Network. The CBDT did not have the experience or expertise required to design, build, and operate a complex tax information system. It entered into a service contract with National Securities Depository Limited (NSDL). CBDT specified only the functionality of the system in broad-terms and the functional requirements for its users.

[12] Dutta et al. 2019.

NSDL then designed the architecture of the whole system, prepared the tender documents to sub-contract other firms to build the system, and is responsible for all of the back-end functions such as data management and maintaining the IT infrastructure.

In the *National Policy and Action Plan for Implementation of Information and Communication Technology in the Indian Judiciary, 2005*, the Supreme Court eCommitee describes multiple coordination problems in working with NIC as a software provider. We see today's data quality issues arise from this old paradigm of treating software as a separate component rather than as one cog in the court administration. For lasting improvements to court data, courts need to start thinking about the entire court administration in a new way.

No development program has ever been effectively completed without a sound monitoring network. The failure of a program is often connected with poor feedback and monitoring systems. To co-ordinate different administration capacities and improve processes, data is required in a form and at a time that is consistent with the prerequisites of the users working at different geographical and administrative hierarchies. These attributes make it necessary to have robust linkages among the data components and progression, and also to how it is shared across various users. The pay-off from improved, reliable, and timely accessibility of data on key indices can be significant. It will not only give the correct type of signals to policymakers, but also enable precise allocation of scarce state-resources for interventions where these are most needed, and can have the greatest impact.

References

[1] Bombay High Court. *Information Required for Implementation of Information and CommunicationTechnology under Resolution 6 (iii) of the Chief Justice Conference 2016*. 2017. URL: http://www.nja.nic.in/CJ-CM_Resolution/Bombay%20HC%20as%20per%20Questionnaire.pdf.

[2] Justice Chandrachud. "32.45 Million Cases Pending In India; 10% Over 10 Years Old : Justice Chandrachud Shares NJDG Statistics". In: *LiveLaw* (2020).

[3] DAKSH. *State of the Indian Judiciary*. 2017.

[4] Pratik Dutta et al. *How to Modernise the Working of Courts and Tribunals in India*. 2019.

[5] eCommittee, Supreme Court of India. *National Policy and Action Plan for Implementation of Information and Communication Technology in the Indian Judiciary, 2005*. 2005.

[6] eCommittee, Supreme Court of India. *Case Management through CIS 3.0*. 2018.

[7] eCommittee, Supreme Court of India. *eCourts Project Phase II: Objectives Accomplishment Report As per Policy Action Plan Document*. 2019.

[8] eCommittee, Supreme Court of India. *eCourts Project Phase II: Objectives Accomplishment Report As per Policy Action Plan Document*. 2019.

[9] Atul Kaushik. *Implementing e-Courts Project at district level*. 2019.

[10] Sudhir Krishnaswamy, Sindhu K Sivakumar, and Shishir Bail. "Legal and judicial reform in India: a call for systemic and empirical approaches". In: *Journal of National Law University Delhi* 2.1 (2014), pp. 1–25.

[11] Law Commission of India. *Report No. 207: Arrears and Backlog: Creating Additional Judicial (wo)manpower*. 2014.

[12] Justice MB Lokur. "NJDG for High Courts is going to be a game changer", Justice Madan Lokur". In: *Bar and Bench* (2017).

[13] Ministry of Finance, Government of India. *Economic Survey 2017–18*. 2017. Chap. 9.

[14] Ministry of Law and Justice, Government of India. *Annual Report of the Department of Justice*. 2018.

[15] National Informatics Centre, Pune. *Court User Manual eCourt Information Systems*. 2016.

[16] National Informatics Centre, Pune. *Registration User Manual eCourt Information Systems*. 2016.

[17] Manaswini Rao. "Whither Justice?: Judicial Capacity Constraints Worsens Trial and Litigants' Outcomes". In: (2020).

[18] Kshitiz Verma. "Analyzing HC-NJDG Data to Understand the Pendency in High Courts in India". In: (2018).

MORE IN THE SERIES

- Chakraborty, L. S., Thomas, E., and Gandhi, P., (2020). Natural Resources Revenue Buoyancy in India: Empirical Evidence from State-specific Mining Regime, W.P. No. 313 (July).
- Karun, H., Vinod, H., Chakraborty, L.S., (2020). Did public investment crowd out private investment in India? W.P. No. 312 (July).
- Pandey, R., Gupta, M., Sachdeva, P., Singh, A., and Sugand, S., (2020). Biodiversity Conservation in India: Mapping Key Sources and Quantum of Funds, W.P. No. 311 (July).

Devendra Damle, is Research Fellow, NIPFP
Email: Devendra.damle@nipfp.org.in

Tushar Anand, is Research Fellow, NIPFP
Email: tushar.anand@nipfp.org.in

National Institute of Public Finance and Policy,
18/2, Satsang Vihar Marg, Special Institutional Area (Near JNU),
New Delhi 110067
Tel. No. 26569303, 26569780, 26569784
Fax: 91-11-26852548
www.nipfp.org.in