A note on Data Preparation for DAKSH High Courts Open Data Project V1.0

Background

From 2016 onwards, DAKSH India has been involved in collating and visualising data from Subordinate & District Courts in India as part of an effort to understand case volumes, pendency and quality of judicial data.[1] This initiative, which grew into the Rule of Law project, developed a protocol for addressing data quality issues while maintaining the DAKSH database.

The existing procedure had to be adapted when applied to data gathered from various High Court across India; certain fields are populated differently at this level and the task of mapping and categorising case types and stages within cases takes on greater value. The key challenge is that High Courts follow different conventions when it comes to naming & classifying cases – which makes it difficult to harmonise judicial data gathered from across the country.

This document outlines some of the issues observed and processes followed while preparing High Court data for analysis, representation and further research.

Data Collection

Data collection was done in January and February 2021 from 23 High Courts (out of 25) across India. In our scraping process, certain benches yielded troubling deviations in the unique cases[2] captured in both case and hearing datasets. For V1.0, we have selected only states for which all benches pass a threshold of 85% cases appearing in both tables. These High Courts and benches are as follows:

High Court	Bench	Overlap Percentage
High Court of Gujarat	Gujarat High Court	98.88 %
High Court of Sikkim	Principal Bench Sikkim	98.48 %
High Court of Jammu and Kashmir	Jammu Wing	98.28 %
High Court of Jammu and Kashmir	Srinagar Wing	97.29 %
High Court of Uttarakhand	Principal Bench Uttarakhand	96.07 %
Orissa High Court	Principal Bench Orissa	95.45 %
High Court of Kerala	Principal Bench Kerala	95.32 %

[1] DAKSH has already written about challenges of working with Judicial Data here

[2] Identified via unique CNR numbers (a 16-digit alphanumeric code assigned to each case)

Data Structure

To begin, the raw scraped High Court files for Case and Hearing records contain 87 and 41 fields respectively. However, many of these are either not consistently filled during data entry (see the **Appendix** to this document for an overview of the fields that appear in raw data tables).

Out of these, the following 30 fields are selected for our case and hearing tables:

DAKSH High Court Case table			
Sl No	Field name (in table)	Attribute/Class type	Notes
1	CNR_NUMBER	Character	16-digit alphanumeric unique identifier
2	CASE_NUMBER	Number	Generated by High Court during filing
3	CASE_TYPE	Character	Abbreviation as per each High Court's classification system
4	CASETYPE_FULLFORM	Character	Full form as per each High Court's classification system
5	CIVIL_CRIMINAL	Character	Mapped by DAKSH – CIVIL, CRIMINAL, WRIT or OTHER
6	SUB_CLASSIFICATION	Character	Mapped by DAKSH – if applicable: APPEAL, FIRST APPEAL, SECOND APPEAL
7	COMBINED_CASE_NUMBER	Character	Generated by High Court; consists of Case type abbreviation, Case number and Year of filing
8	COURT_NAME	Character	Bench within High Court
9	COURT_NUMBER	Number	
10	NAME_OF_HIGH_COURT	Character	High Court
11	CURRENT_STAGE	Character	As entered in eCourts
12	CURRENT_STATUS	Character	Either DISPOSED or PENDING
13	DATE_FILED	Date	Date on which case is accepted into system
14	DECISION_DATE	Date	Date on which case is disposed
15	DISPOSAL_DAYS	Number	
16	FILING_NUMBER	Number	
17	HEARING_COUNT	Number	
18	LAST_SYNC_TIME	Date	Last date of sync before scraping
19	NATURE_OF_DISPOSAL	Character	As entered in eCourts
20	NATURE_OF_DISPOSAL_BINARY	Character - mapped from previous column	CONTESTED, UNCONTESTED or null if case is pending in data
21	NJDG_JUDGE_NAME	Character	As entered in eCourts
22	PENDING_DAYS	Number	
23	PETITIONER*	Character	
24	POLICE_STATION	Character	

25	REGISTRATION_DATE	Date	
26	REGISTRATION_NUMBER	Number	
27	RESPONDENT	Character	
28	UNDER_ACTS	Character	
29	UNDER_SECTIONS	Number	
30	YEAR	Date	Year of filing or institution

DAKSH High Court Hearing table			
Sl No	Field name (in table)	Attribute/Class type	Notes
1	CNR_NUMBER	Character	16-digit alphanumeric unique identifier
2	NATURE_OF_DISPOSAL	Character	As entered in eCourts
3	NATURE_OF_DISPOSAL_BINARY	Character - mapped from previous column	CONTESTED, DISPOSED or if case is pending in data, NA
4	CURRENT_STATUS	Character	Either DISPOSED or PENDING
5	REGISTRATION_DATE	Date	
6	DATE_FILED	Date	Date on which case is accepted into system
7	DECISION_DATE	Date	Date on which case is disposed
8	POLICE_STATION	Character	As entered in eCourts
9	HEARINGID	Character – generated by DAKSH	Consists of CNR no. and Business-on date
10	PETITIONER*	Character	
11	PETITIONERADVOCATE*	Character	
12	RESPONDENT	Character	
13	RESPONDENTADVOCATE*	Character	
14	CURRENT_STAGE	Character	As entered in eCourts
15	COMBINED_CASE_NUMBER	Character	Generated by High Court; consists of Case type abbreviation, Case no and Year of filing
16	BUSINESS_ON_DATE	Date	Date on which each hearing occurred
17	HEARING_DATE	Date	
18	PURPOSE_OF_HEARING	Character	As entered in eCourts
19	COURT_NAME	Character	Bench within High Court
20	YEAR	Number	Year of filing
21	NAME OF HIGH COURT	Character	High Court
22	COURT_HALL_NUMBER	Number	
23	NJDG__JUDGE_NAME	Character	As entered in eCourts
24	CASE_TYPE	Character	Abbreviation as per each High Court's classification system

25	CASETYPE_FULLFORM	Character	Full form as per each High Court's classification system
26	SUB_CLASSIFICATION	Character	Mapped by DAKSH – if applicable: APPEAL, FIRST APPEAL, SECOND APPEAL
27	CIVIL_CRIMINAL	Character	Mapped by DAKSH – CIVIL, CRIMINAL, WRIT or OTHER
28	MAPPED_STAGE_NAME	Character	Mapped by DAKSH
29	DAYS_BETWEEN_HEARING	Number	
30	PREV_HEARING	Date	As entered in eCourts

Data Preparation

As part of the cleaning process:

Cases were de-duplicated & filtered; V1.0 is predominantly interested in cases filed in the decade from 2010 to 2019 as digitisation of court records increased significantly during this period under the eCourts Mission.

Discrepancies between fields that should concur (i.e. Date of filing & Year) were edited in accordance with the field that provided more detail

Researchers at DAKSH then applied their knowledge of the court systems to map relevant fields. For High Court data, this included the nature of case (Civil, Criminal or Writ), level of appeal, the expansion of the abbreviated Case Type, nature of disposal and the purpose of particular hearings.

Calculations were generated for 'Disposal Days', 'Pending Days' and 'Days between Hearings' and computation method checked

Final checks were done to ensure no essential fields were empty and that no case records appeared in only one out of the case and hearing tables.

In the interest of protecting personal data of petitioners are advocates, the 'Petitioner', 'Petitioner Advocate' and 'Respondent Advocate' fields have been removed from the dataset being released. We have maintained a version with these fields included; please reach out to DAKSH at info@dakshindia.org if you are interested in accessing this data for research.

In V1.0 of this project, the goal is to put out a usable dataset of High Court judicial data and collaborate with like-minded researchers and organisations to determine what insights could be reached at a system-level. We also hope to identify easy-to-implement modifications in data structuring that could better enable rigorous pan-India analysis of judicial workload.

Further changes to this codebook and the data preparation process are anticipated as this project progresses.