Technical Debt in India's Open Government Datasets

Shashank Yadav
CPS / 214468003
2144680003@iitb.ac.in

Abstract

India's Open Government Data portal offers close to 500,000 open datasets. These datasets have been downloaded over 9 million times. We select a small sample and carry out a data quality assessment to ascertain the nature of the technical debt that any quality issues in these datasets might be incurring. Our results suggest that not only there are significant quality issues with Open Government Data, some government sectors bring in a disproportionately high share of data quality problems. Further study with a much larger sample and a wider range of checks is required to better understand the extent of the problem and validate these results.

1 Introduction

Technical Debt (TD) is defined as design or implementation constructs that are beneficial in the short-term, but incur a "debt" which leads to significant problems in the long-term. The origin of the concept can be traced back to Ward Cunningham, who in 1992 used the metaphor of economic debt while describing software management problems, particularly those relating to the source code *(Cunningham, 1992)*. Today, in the age of AI, as software systems have evolved to assimilate large components around data engineering, giving rise to data-intensive software systems such as those extensively dependent upon machine learning and data science workloads, the metaphor of technical debt too has been expanded to cover the problems originating from predominantly data quality issues. In context of data, put simply, Technical Debt becomes the implied cost of future rework required to improve the quality of data that makes it easy to understand and work with.

Some of the most widely used and freely-available datasets, which fuel the analytics processes of myriad social sector and local governance initiatives, are the Open Government Datasets [OGDs]. The Indian government too, through its National Informatics Center (NIC), makes available a plethora of datasets from a diverse range of governance sectors. These datasets are available at https://data.gov.in/ and have been a key enabler for not only building socially impactful software services but also informing the analyses of industry, research, and development professionals about the ground realities in India.

India's OGDs have been downloaded over 9 millions times, consequently the debt incurred from any data quality issues therein is also not isolated but propagated throughout the application development and data science processes based upon those OGDs across the country. This make an assessment of the nature of technical debt being incurred from OGDs absolutely critical in times when anything and everything is morphing into a "data-driven" or "AI-enabled" structure.

2 Related Work

Most studies of OGDs have focused mostly on the web experience, accuracy, navigation and search/discoverability of information on the OGD portal (Saxena, 2019), asymmetry between different states and government bodies in releasing data (Saxena, 2018), as well as stakeholder and policy issues(Verma and Gupta, 2013). A systematic assessment of actual data quality issues in India's OGDs from an analytics point-of-view has been found lacking. Notwithstanding, as a general area of research, the quality of OGDs has received some attention internationally. For example, Purwanto et al. (2020) suggest that data quality issues undermine the citizenry's trust in the OGDs. Bicevskis et al. (2018) take a more quantitative approach in measuring the actual quality issues and look at aspects like missing values in the OGDs of Norway, Britain, Estonia, and Latvia – proposing a platform specific and a platform independent data quality management framework.

Since governments across the world have been aggressively pushing for AI in everything, including the Indian government with its ambitious AIFORALL approach, it is worth noting that the second largest source of anti-patterns in AI come from deficiencies or suboptimal practices around data, of which most are associated with data quality (Bogner et al., 2021). It is well established that in machine learning systems, data dependencies cost more than code dependencies and lead to higher technical debt (Sculley et al., 2015).

In a leading work on Technical Debt, Avgeriou et al. (2016) expound upon the Cost of Change (CoC) in legacy systems, and suggest that expected CoC is much lower than actual CoC, which increases at a much greater rate, leading to most people more making the change when required. This is a classic debt creating situation where the short-term incentives continue to reinforce the existing debt. When the matter reaches data-intensive systems, the quality of data has also been termed as the single greatest debt incurring aspect of data (Foidl et al., 2019).

That said, an assessment of Technical Debt in OGDs is also lacking. Now, we'll undertake an assessment of data quality in India's OGDs and try to characterise the debt therein.

3 Methodology and Process

3.1 Data Collection

The API documentation to access the OGDs tells us that full access to the requested datasets will be given to the user once he has provided his own API-key. If the user does not have an API-key, upon receiving the API request, the platform would return only the initial 10 records of the requested dataset. However, the information in the documentation proved incorrect as it turned out upon experimentation that the platform only returns the initial 10 records for all requests unless a limit parameter is also specified along with the user's API-key.

Therefore an estimate was made about the size of datasets based upon exploration and then a large limit was specified which would cover all the records contained in the OGD. This is poor API design, however upon having ascertained that complete datasets were being downloaded after providing a large enough limit parameter, the top 10 datasets from each main governance sector available at `https://data.gov.in/` were downloaded after being sorted for Relevance. There are 33 main sectors under which datasets are available. Some sectors like Biotechnology, Judiciary, and Foreign Affairs etc have only 2-3 datasets available. There were also a few instances (12) in the selected sample of datasets from different sectors having same identifying Resource-Ids, and in some instances the API access was not available. At the end, 288 datasets were scraped.

3.2 Analysis

For assessing the quality of OGDs, the tool DeepChecks was found to be an excellent choice. It is quite well received in the machine learning community for testing and validating models, comparing and evaluating between them, as well as for assessing the integrity of single datasets. It is the latter use of it which made it useful for this project.

Consequently, the scraped OGDs were brought to be "deepchecked" for the following:

- Data duplicates (DD)
- Single values (SV)
- Mixed data types (MT)
- Mixed nulls (MN)
- Special characters (SC)
- Multiple variations of same strings (StrV)

It turned out at the beginning of the analysis that 18 datasets contained no data. As a result of which, the output of analysis was further narrowed down to the remaining 270 datasets.

4 Results and Discussion

Overall 316 columns with single values, 64 columns with mixed data types, 19 with mixed nulls, 26 having a variety of special characters, and 73 columns having multiple variations of the same strings were found, other than significant duplication of data in some of the key sectors, the maximum of which was in a data asset from water sector, containing 67.5 percent duplicate data. There were only 8 datasets with duplication. However, in the total of over 500,000 OGDs the Technical Debt from Data Quality issues can be expected to be proportionally, very high. One very curious phenomena which emerged through analysis was that of Sectoral Debt.

It turned out that data quality issues are not uniformly distributed throughout all 33 sectors – some

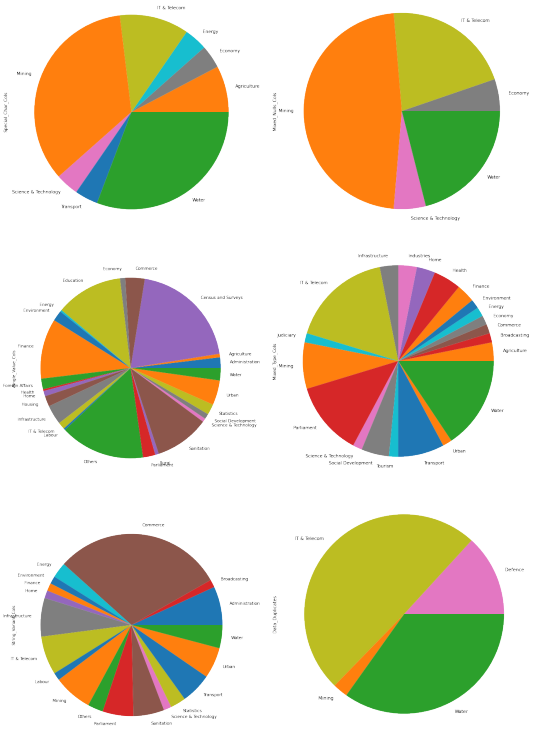

Figure 1: Data quality issues are not uniformly distributed across all sectors – some government sectors consistently contribute a disproportionately high share of TD.

government sectors bring in more technical data debt than others, for example, most instances of data duplicates occurred in IT & Telecom, Defence, Water, and Mining. This phenomena of Sectoral Debt in OGDs does not seem to be a mere coincidence as it spreads throughout all verticals of analysis. The prevalence of special characters, for example, is largely found in Mining, IT & Telecom, Water, and Economy. Similarly, the numbers of columns with single values were found to be relatively much higher in Census and Surveys, Finance, Education, Sanitation and 'Others'. Special characters and mixed nulls were found to be primarily distributed in Mining, Water, IT & Telecom, and Economy. Similarly, Commerce sector showed the most variations of same strings.

This phenomena of sectoral data debt should be concerning for policy-makers for it raises questions over data handling practices in some of the strategic sectors of governance. However, the study has limitations as the sample size is too small. Further enquiry is needed with a much larger sample size and a wider range of data quality checks, also pertinent is collaborating with respective data controllers and NIC to figure out the actual time and cost of fixing these issues.

For all of the 270 OGDs which were analysed, the following table summarises all the instances of data quality issues in them in a sector-wise manner for all the 33 sectors:

Sector	DD	SV	MT	MN	SC	StrV
1	0	6	0	0	0	5
2	0	2	2	0	2	0
3	0	0	0	0	0	0
4	0	0	1	0	0	1
5	0	63	0	0	0	0
6	0	11	1	0	0	22
7	0.39	0	0	0	0	0
8	0	3	1	1	1	0
9	0	38	0	0	0	0
10	0	1	1	0	1	2
11	0	6	1	0	0	1
12	0	34	2	0	0	1
13	0	6	0	0	0	0
14	0	1	3	0	0	0
15	0	3	2	0	0	1
16	0	6	0	0	0	0
17	0	0	2	0	0	0
18	0	11	2	0	0	5
19	1.46	4	11	4	3	5
20	0	0	1	0	0	0
21	0	1	0	0	0	1
22	0.072	0	5	9	9	5
23	0	48	0	0	0	2
24	0	7	8	0	0	4
25	0	2	0	0	0	0
26	0	31	0	0	0	4
27	0	2	1	1	1	1
28	0	3	3	0	0	0
29	0	6	0	0	0	2
30	0	0	1	0	0	0
31	0	0	5	0	1	4
32	0	14	1	0	0	4
33	1.025	7	10	4	8	3
Total	2.947	316	64	19	26	73

Table 1: [Administration (1), Agriculture (2), Biotechnology (3), Broadcasting (4), Census and Surveys (5), Commerce (6), Defence (7), Economy (8), Education (9), Energy (10), Environment (11), Finance (12), Foreign Affairs (13), Health (14), Home (15), Housing (16), Industries (17), Infrastructure (18), IT & Telecom (19), Judiciary (20), Labour (21), Mining (22), Others (23), Parliament (24), Rural (25), Sanitation (26), Science & Technology (27), Social Development (28), Statistics (29), Tourism (30), Transport (31), Urban (32), and Water (33)]

4.1 Technical Debt Estimation

Following the suggestion of Curtis et al. (2012), we can look at the principal technical debt as a function of the time and cost of resolving these problems, further classifying problems by their importance/severity and weighting them accordingly. With a very conservative estimate of just two hours (10 for duplicates) and INR 200 an hour to resolve these issues, the TD works out as follows:

TD-Type	Number	Weight	Time	Cost
DD	8	1	10	16000
SV	316	0.2	2	25280
MT	64	1	2	25600
MN	19	0.8	2	6080
SC	26	0.8	2	8320
StrV	73	0.8	2	23360
Total Cost				104640
Cost per dataset				387.56
Principal-TD for 500,000 OGDs				193777777.78

Table 2: Principal TD Estimation (in INR)

The principal technical debt itself, based on conservative estimates, works out to be in hundreds of millions of rupees for all the OGDs – how much it might incur across the pipeline of data usage in terms of rework is beyond the scope of this work. That said, it must also be taken with a grain of salt as our estimates of Time and Cost are fairly assumptive and need to be corroborated with NIC data controllers. The dependence of DeepChecks as a tool of analysis is also a limitation of this study, and a more thorough method would be to also validate these with other tools and techniques as well.

5 Concluding Remarks

OGDs can be used to power socio-economically high impact applications, but any technical debt emerging from data quality issues of OGDs spreads across the government and its services. Out litle study suggests that the overall technical debt originating from the quality of India's OGD's can be quite high. In the small sample chosen for the study itself, a range of issues were encountered and it was observed that data quality is not uniform acros sectors. Some government sectors bringing in more debt than others is an interesting phenomena which calls for further enquiry with larger sample and wider range of checks. To improve the quality of OGD's, it is important to set internal technical standards for government departments and use regular human audits, internal red-teaming exercises, and persistent automation based methods to ensure that the datasets being produced are adhering to those standards.

References

Paris Avgeriou, Philippe Kruchten, Ipek Ozkaya, and Carolyn Seaman. 2016. Managing technical debt in software engineering (dagstuhl seminar 16162). In *Dagstuhl Reports*, volume 6. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Janis Bicevskis, Zane Bicevska, Anastasija Nikiforova, and Ivo Oditis. 2018. An approach to data quality evaluation. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 196–201. IEEE.

Justus Bogner, Roberto Verdecchia, and Ilias Gerostathopoulos. 2021. Characterizing technical debt and antipatterns in ai-based systems: A systematic mapping study. In *2021 IEEE/ACM International Conference on Technical Debt (TechDebt)*, pages 64–73. IEEE.

Ward Cunningham. 1992. The WyCash portfolio management system. *ACM SIGPLAN OOPS Messenger*, 4(2):29–30. Publisher: ACM New York, NY, USA.

Bill Curtis, Jay Sappidi, and Alexandra Szynkarski. 2012. Estimating the size, cost, and types of technical debt. In *2012 Third International Workshop on Managing Technical Debt (MTD)*, pages 49–53. IEEE.

Harald Foidl, Michael Felderer, and Stefan Biffl. 2019. Technical debt in data-intensive software systems. In *2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pages 338–341. IEEE.

Arie Purwanto, Anneke Zuiderwijk, and Marijn Janssen. 2020. Citizens' trust in open government data: a quantitative study about the effects of data quality, system quality and service quality. In *The 21st Annual International Conference on Digital Government Research*, pages 310–318.

Stuti Saxena. 2018. Asymmetric Open Government Data (OGD) framework in India. *Digital Policy, Regulation and Governance*, 20(5):434–448.

Stuti Saxena. 2019. Proposing a total quality management (TQM) model for open government data (OGD) initiatives: implications for India. *foresight*, 21(3):321–331.

Stuti Saxena and Marijn Janssen. 2017. Examining open government data (OGD) usage in India through UTAUT framework. *foresight*, 19(4):421–436.

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. 2015. Hidden Technical Debt in Machine Learning Systems. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Neeta Verma and M. P. Gupta. 2013. Open government data: beyond policy & portal, a study in Indian context. In *Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance*, pages 338–341, Seoul Republic of Korea. ACM.