Isabelle Adam, Government Transparency Institute (GTI)

Elizabeth Dávid-Barrett, University of Sussex

Bence Tóth, University College London, GTI

Mihály Fazekas, Central European University, GTI

INDIA'S FEDERAL PROCUREMENT DATA INFRASTRUCTURE: OBSERVATIONS AND RECOMMENDATIONS

FEBRUARY 2020

INTRODUCTION

Improving transparency in public procurement—that is publishing more and better-quality data—supports accountability by enabling greater scrutiny over processes and outcomes, and helping to achieve greater competition and better value for money. In India, according to the Ministry of Finance General Financial Rules (2017), all procuring authorities are responsible and accountable for ensuring transparency, fairness, equality, competition, and appeal rights in contracting. The transparency principle is about making information easily accessible to the public: it prescribes that all procuring entities should ensure the publication of all relevant information on the Central Public Procurement Portal (CPPP).

As part of the 'Curbing corruption in procurement' research project, our team has collected, cleaned, standardised, and analysed national public procurement data from a diverse set of countries in Latin America, Africa, and Asia, including India (federal level). This research involves mining large amounts of government contracting data from government portals and repositories in order to analyse how procurement can be manipulated for corrupt ends. We analyse the data to identify suspicious patterns widely associated with corruption, such as tailored bidding conditions and only a single bid being submitted on a market with multiple potential bidders.

Despite the General Financial Rules' formal requirement for transparency, we found that the Indian federal public procurement data that we could collect from public sources was insufficient for robust analysis. Besides a number of technical difficulties, the key problem is that many contract awards are not published; their publication seems not to be monitored or enforced and most contract awards are missing. This makes rigorous analysis impossible, since it is likely that our sample is biased and, moreover, it is impossible to determine the nature of any bias.

Given the Indian government's commitment to the transparency principle, this report seeks to inform future reforms by providing: (1) a description of our data collection efforts and our (incomplete) dataset; (2) our observations on the current data infrastructure; and (3) a set of recommendations for how to make the data more accessible and usable for analysis in the future.

DESCRIPTION OF THE DATASET

We collected data in three steps: (1) looked up the most comprehensive source of contract-level procurement publications and annotated the notices; (2) downloaded all available publications (calls for tender and contract awards); and (3) extracted information from the downloaded HTML documents and stored it in a standardized database.

In 2018–19, we annotated the public procurement web portal (https://eprocure.gov.in/eprocure/app) published by the Central Government at the time, hence the Government eMarketplace[1] is not considered in this analysis.[2] Our annotations covered call-for-tender notices, contract award notices, and corrigenda, linking different fields on the website to variables in our dataset. Based on these, the website was scraped and the available information put into JSON and CSV structures.

The resulting dataset contained 824,764 observations, representing information from around 190,000 contract award notices and around 690,000 call-for-tender notices. This big discrepancy between the data available for the two stages of the tendering process indicates that a lot of crucial information on awarded contracts is missing (Table 1).

Given data weaknesses, it is not possible to develop and test a corruption risk assessment framework the project has developed in other countries. Nevertheless, it is possible to show the range of potentially valid corruption risk indicators that could be developed and tested if the data, in particular data completeness, improves (Table 2).

Variable	Non-missing rate (%)
Tender title	83
Supply type (G/W/S)	5
Number of bids	100
Contract signature date	21
Product sub-category	83
Tender cancellation date	0
Award decision date	0
Estimated contract value	0
Total contract value	17
Buyer ID	0
Buyer name	100
Buyer contact details	83
Buyer type	0
Winner ID	21
Winner name	83
Tender publication date	83
Bid submission deadline	93
Procedure type	93

Table 1. *Key variables in the Indian federal public procurement dataset ($N_{CAs\&CfTs}$=824,764), 2013-2016.*

Corruption Risk Indicator	Non-missing rate (%)
Single bidding	86
Supplier dependence on buyer	84
Submission period length	21
Lack of call for tenders publication	100
Procurement method	69
Tender description length	100

Table 2. *Indicators calculable for the Indian federal public procurement dataset ($N_{CAs\&CfTs}$=824,764), 2013-2016.*

OBSERVATIONS ON THE EXISTING DATA INFRASTRUCTURE

In the following, we set out our detailed observations about India's federal procurement data infrastructure, based on our efforts to collect data from the e-procurement portal, CPPP.

The central publication website

Generally, the rules governing publication on the website are laid down in the 2017 General Financial Rules and are further specified in manuals. For example, the Manual for Procurement of Goods (2017) states: "It is mandatory for all Ministries / Departments of the Central Government, Central Public Sector Enterprises (CPSEs) and Autonomous and Statutory Bodies to publish all their tender enquiries, corrigenda thereon and details of bid awards on the CPPP."[3] **The rules, however, are not fully followed and there is no indication that compliance is monitored or enforced.**

Information is published on two portals, both of which are updated, but the distinction between the two is not clear. A new portal (https://eprocure.gov.in/cppp/) was launched in 2019 in parallel to the old one (https://eprocure.gov.in/eprocure/app). The information published on the portals is not identical (e.g., some tenders are available on one but not the other, and the lists of latest Active Tenders are not the same) and searching for tenders with the same IDs on both portals does not yield consistent results. Contract awards are available on the new portal via a search form, but the same form on the old portal does not give any results in a manual search for specific Tender IDs. Additionally, the new portal contains dashboards covering data from various state and CPSE tender portals and offers links to these portals, including the old CPPP one.

Automated data collection is impossible because most information is only accessible via search forms, which are protected by a captcha. In addition, users need to enter a specific tender ID to find any information. Therefore, historical data with Tender ID cannot be downloaded unless the user knows the Tender IDs and can bypass the captchas.

Unique identifiers: Announcements and tenders

It is impossible to connect information from calls for tender and contract award notices that relate to the same tendering process because the IDs are inconsistent and not traceable across or within websites.

- Although most tenders have a Tender Reference Number and a Tender ID, this does not prove useful to find the corresponding contract award on either of the portals.
- In our dataset, we could only connect calls for tender and contract awards for 4% of all tenders. Manual cross-checks confirmed that, without consistent IDs, it is not possible to link these two types of publications.
- If information on calls for tenders and contract awards cannot be connected in our dataset, it is much less useful for analysis. This is because, for most procurement performance analysis, it is necessary to trace a tendering process from publication to award.

Unique identifiers: Organisations

It is not possible to consistently identify buyers or companies across contracts because the IDs of procuring entities (buyers) and companies (suppliers) are currently entered as free text rather than standardised unique IDs. This means that multiple versions of the same organisation name may appear owing to different characterizations of names (or even linguistic variation in spelling conventions or simply typing errors). This makes it difficult to analyse the practices and outcomes associated with any particular organisation.

Missing data

There is a great deal of missing data. Many data fields are not filled in on both the call for tender and contract award publications. This results in an incomplete picture and may be a source of systemic bias in any analysis. As Table 1 shows, crucial variables, such as supply type, buyer ID, contract value, tender award or decision date, or winner ID, are missing entirely or to a high degree.

Inconsistencies in terminology

Some key terms are used inconsistently. Some fields in the publications appear to have been inputted incorrectly. For example, “Tender type” in the call-for-tender document often refers to the procedure type used (e.g., open call or limited competition). However, “Tender type” in the contract-award document refers to the type of the purchase (e.g., whether the procurement is about works, goods, or services).

RECOMMENDATIONS

1. Make the publication of contract awards mandatory throughout the federal public procurement system and communicate the requirement to all stakeholders.
2. Monitor and enforce clear rules for procuring entities to collect and publish relevant public procurement data in a consistent and timely manner, including publication of contract awards.
3. Publish all data in one place (ideally the CPPP website) in machine-readable format (e.g., CSV, JSON, XML) to improve usability. Users should also be able to download data in bulk either as CSV or through an API.
4. Use unique standardised IDs for all tender announcements and contract-award notices to ensure that they can be linked.
5. Use unique standardised IDs for organisations—both buyers and suppliers—in addition to their names.
6. Collect information on more details of the tender process and in standardised formats (e.g., detailed product codes and structured addresses).
7. Publish information on amendments, modifications, and failed tenders in a structured and reliable format so that up-to-date information is available on all tenders.
8. Facilitate matching with other public datasets (e.g., it should be possible to match procurement data with budgets or other public financial management data, company registry data, court rulings).

ENDNOTES

[1] https://gem.gov.in

[2] In 2019, a new portal was launched in parallel: https://eprocure.gov.in/cppp/. Although we continued to collect data from the previous website, our findings in this paper address issues encountered with the new portal (e.g., regarding its search functionality, captchas). Moreover, since similar information is published on both portals, our observations about basic data quality and recommendations apply equally to the new portal.

[3] See https://www.finmin.nic.in/sites/default/files/Pub_tender_Enq_CPPPortal_1.pdf?download=1 and p. 9 of https://doe.gov.in/sites/default/files/Manual%20for%20Procurement%20of%20Goods%202017_0_0.pdf

UK aid
from the British people

ACE
GLOBAL INTEGRITY
Anti-Corruption Evidence
Research Programme

This project is part of the Global Integrity Anti-Corruption Evidence (GI-ACE) research programme, which supports 14 projects around the world generating actionable evidence that policymakers, practitioners, and advocates can use to design and implement more effective anti-corruption initiatives.
http://ace.globalintegrity.org

GI-ACE is part of the Anti-Corruption Evidence (ACE) research programme funded with UK aid from the British people.
http://anticorruptionevidence.org