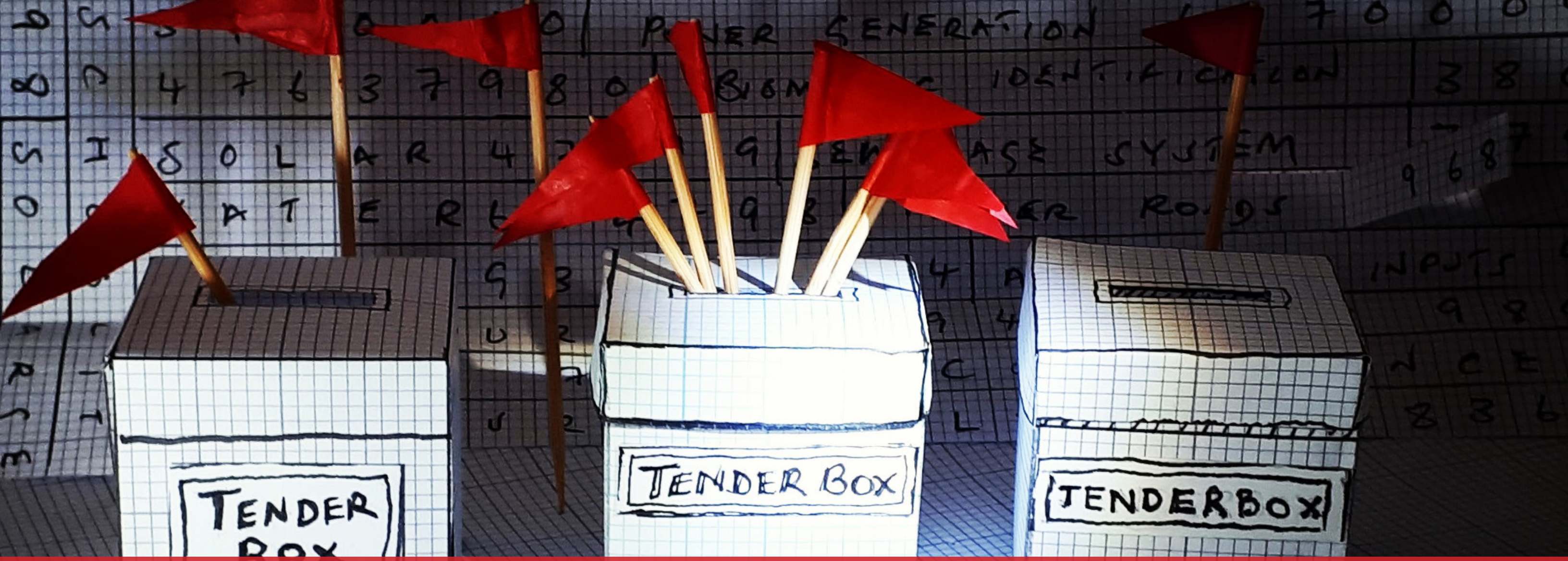

RED FLAGS
FREQUENTLY ASKED QUESTIONS

Where do you get public procurement data from?

Most countries have centralized websites, often maintained by procurement authorities, where they put out standardized publications, such as ‘calls for tender’ or ‘contract award’ documents (stating who won a given contract).

How do you collect the data from these websites? What are the key stages of the process?

The first stage is **mapping what kind of publications are published on these central procurement sites** and **which bodies are required to publish there**. We need to know, for example, whether they publish contracts for goods, works, and services—the three main categories of items that are procured. If they do, we also need to know what the threshold value for publication is (usually only contracts above a certain value have to be published). We also need to know which public sector entities are obliged to publish—is it only the central government or local / regional bodies also? What about state-owned enterprises, such as the national oil company?

Next, we have to **scrape these documents from the website** unless they are available in bulk, which typically is not the case. Mainly they are in xml or html formats. We need to understand what everything means in these publications and connect each item to a database standard (e.g., OCDS or DIGIWHIST data standard, i.e., a relational dataset). Based on these correspondences, our

The research of the ***Curbing Corruption in Procurement*** *project analyses how procurement can be manipulated for corrupt ends using a prize-winning ‘red flags’ methodology developed by Mihály Fazekas. We collect datasets of procurement tenders and contracts, with a range of variables that indicate corruption risk, and analyse the data to identify suspicious patterns and trends, by procuring entity, supplier, and over time.*

Team Members

The project is led by ***Liz Dávid-Barrett****, Senior Lecturer in Politics at the University of Sussex and Director of the Centre for the Study of Corruption; and* ***Mihály Fazekas****, Assistant Professor at Central European University. The research team also includes Bence Tóth, Ágnes Czibik, and Isabelle Adam.*

programmers can then build an algorithm to automatically collect or ‘scrape’ all of the data from the websites. At this stage we need to do **data mastering,** or collating the same information coming from different sources (e.g., the buyer’s name might be published both in the call for tender and in the contract award notice, but we need to have a rule about which version to keep).

The next step is **cleaning the data.** Usually the scraped data contains a lot of values that do not make any sense and, hence, are probably errors. For example, if there is a contract for highway construction with a price of only 5 EUR, that is probably a mistake! We have to identify all of these kinds of errors, as they could distort our analysis.

The final step is to **validate the data.** Once we have a dataset, we have to validate whether our scraping has collected every document published, check that our mastering did not combine different data points under the same headline, and ensure that we collected the data correctly from the publications (a process often called parsing).

What is the OCDS? If a country is part of that, does it mean that its data is top-notch quality?

OCDS stands for Open Contracting Data Standard and it is driven by the Open Contracting Partnership. OCDS is about the **format** of the data and how it is published—in json structure, either as an API or in large data dumps—**not about data quality.** So OCDS data typically does not have any different quality than the source data.

In general, improving data quality requires thousands of public procurement officials to understand the reporting requirements correctly and take their time to accurately report what has happened in the procurement procedure. Unsurprisingly, improving data quality takes a major culture shift or a whole new transactional information technology (IT) system (i.e., the system is used to conduct the transactions themselves, rather than merely used for reporting to the public separately from the procedure itself). There is a lot of interest in e-procurement because this means that parts of the transaction take place through the system, both removing the scope for discretion and having the consequence that data are automatically collected.

Can you collect and analyse public procurement data for all countries? If I want to study and compare six countries, can I choose any six?

No, probably not. First, **countries differ in terms of what kinds of contracts they publish.** Sometimes they publish only works, not goods and services. Sometimes they publish only federal or central contracts, not state- or local-level contracts. Usually, countries only publish contracts above a certain value threshold, but those thresholds vary among countries and change over time. So, depending on what kinds of procurement you are interested in, the data may or may not be available for a given country.

There also are **important differences in the depth of the information** they publish. If they only publish the calls for tender, but not the final price or the name of the winning bidder, then it is difficult for us to do any meaningful analysis.

Another problem is that we **need to be able to trust the accuracy of the data that governments publish.** We need to know how much data is missing from the official documents and, ideally, we would need some insight into whether the gaps are random. Otherwise, we might find that the data we are analysing contains some systemic biases.

How do you establish whether a country has data of good enough quality?

A good start is to u**nderstand the procurement law** and, in particular, what the **mandatory requirements are about publishing data.** If the law only requires a small subset of the data to be published, then that is not good for us.

Some countries have wide requirements to publish data, and yet do not really enforce it. Ultimately, the centralised datasets that we rely on in turn rely on a wide range of procuring entities uploading their data. If they do not do that, and nobody chases them up to ensure that they comply with the rules, then we are likely to have very incomplete datasets. Moreover, we do not know if the gaps reflect deliberate efforts to keep deals secret, or rather just a lack of efficiency in uploading the data.

Ultimately, there is no single answer to whether a dataset is good enough quality; **different sets of analyses require different qualities.** Also, there is a lot of trial and error in looking at general patterns in the data (e.g., breaks in the time series of contract values). Where possible, we also look at related datasets to sanity-check our procurement data so, for example, if we see a country where 50% of the GDP is produced by the capital city, but we barely have any contracts from that city, it is an indication that there is a fundamental problem with the data.

Is it easy to tell whether the data are going to be good quality before investing in collecting it? What can go wrong?

Tanzania is a good example. Before we started to collect data, we looked at the overall size of the data—there were thousands of records—and that looked okay. We also saw that there were similar numbers of calls for tender and contract awards, and IDs were used in various publications, suggesting that we would be able to link them. But then we started collecting the data and found many problems. In the end, it turned out that the IDs were not linkable, and that most calls for tender could not be matched to contract awards or vice versa. *(More details in our report.)*

One key red flag for corruption risks in procurement is that there is only one bidder for a contract. Why might that indicate corruption, and what do you need to construct that indicator?

This is an **'outcome' indicator**—an outcome that might reflect corruption risk. For example, if most suppliers think that a deal is already pre-agreed, it deters them from bidding. If there is only one bidder, maybe a lot of companies judged that they would have no chance, and the only one who did bid was one who had some special information or access.

Sometimes there are other signals of this. For example, if the tender was specified really narrowly, that also might put off some companies because it indicates that the tender has been tailored to suit a particular company. But you will not necessarily be able to judge whether the specification was narrow. It's a good proxy to look at the number of bidders and, if it is low, then this may be indicative of favouritism.

In terms of constructing the red flag from data, if you have the names of the bidders, then you can count how many there were and find out that there was only one. However, buyers often only record the name of the winning bidder and forget to record the others, so relying on bidder names often leads to under-estimation. An alternative way to publish and record bidder numbers is simply to require procuring entities to report a numeric value. That is easy for buyers to collect and does not require the laborious typing in of many company names.

Another indicator relates to the period for which the tender is advertised. What do you need for that one?

The logic here is that one easy way of favouring a particular bidder is to give it advance warning about the tender, but then make the official deadline for submitting bids short. That means that few companies will have time to prepare a bid, a process which often requires assembling a lot of documents as well as careful planning of how to deliver the contract.

You can construct this indicator from the date of the call for tenders and the deadline for submitting bids. This is a **'process' indicator**—by looking at the process, you can see that something might be amiss. Process indicators are best used together with outcome indicators, such as single bidding.

What about matching up procurement process and outcome indicators with other data, that is not about procurement, such as company records?

In principle, this is a great idea and leads to a lot of useful government tools—such as conflict-of-interest checkers—and great research—such as that on the impact of party donations on procurement success. However, the main challenge here is to find a unique key—a variable which allows the different datasets to be linked. Even though it sounds simple and easy, matching up company names that may have been entered in hundreds of ways with a corporate registry is a laborious and error-prone task. It is far better if you have widely used IDs—like company tax registry IDs—in the procurement data to start with.

UKaid
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