Iceland Liechtenstein Norway grants | Norway grants

An overview of procurement integrity and introduction to opentender.eu

Procurement data: sources, possible errors, and examples of data availability

Péter Horn

First specialized regional training for R2G4P member, 7 July 2021

Implemented by:

CSD CENTER FOR THE STUDY OF DEMOCRACY

SELDI.net Southeast European Leadership for Development and Integrity

GTI

The R2G4P project, coordinated by the Center for the Study of Democracy, Bulgaria benefits from a € 1.5 million grant from Iceland, Liechtenstein and Norway through the EEA and Norway Grants Fund for Regional Cooperation. The aim of the project is to implement shared anti-corruption and good governance solutions in Southeast Europe through innovative practices and public-private partnerships.

Presentation overview

- The goals of a procurement database
- Relevant datatypes
- Key aspects of procurement data (scope, depth, quality, access)
- Examples from partner countries
- Errors in the data

Goals & Objectives of procurement databases

Main objective:

- **Goal**: To create a comprehensive public procurement dataset, which makes the evaluation of the integrity of countries' procurement systems possible.
- Create objective indicators to measure procurement integrity/corruption risk
- This requires high quality administrative data on:
  1. **public procurement tenders and contracts,**
  2. **bidding companies,**
  3. **awarding public organizations and**
  4. **political office holders.**

DIGIWHIST:

- Large-scale EU-funded research project which simultaneously aims to increase trust in governments and improve the efficiency of public spending across Europe.
- Supports corruption measurement by organizing and linking the four complex datasets.
- Its data template also serves the basis for collecting and republishing publicly available and sufficiently well-structured databases pertaining to corruption measurement in Europe.

Data types I.

1. **Public procurement data - (contract or item level) - *Mostly available***
    1. **Call for tender related information:** procedure type, product code, bidding period length, bidder limitation, estimated value, type of the contract, documentation fee, buyer, award criteria.
    2. **Contract award related information:** number of bids received, bidder and winner company related information (bid prices, location), final contract value, award signature date.
2. **Company data - *Partially available***
    1. **Registry information:** company name, location, legal form, date of incorporation, number of employees etc.
    2. **Financial information:** annual turnover, profit rate, return on assets, material costs, personnel costs, taxes, EBITDA.
    3. **Ownership information:** number of recorded shareholders, shareholder's name, shareholder's type (legal entity, individual etc.), shareholder's location, shareholder's direct and total shares.
    4. **Manager information:** number of directors, name of company directors, position of company directors, appointment and resignation date of directors, gender, date of birth, shareholder status.

Data types II.

1. **Public organization data - *Partially available***
    1. **Registry data**: name, ID, location, activity type, contact information.
    2. **Budget data:** annual budget figures, currency, classification of the budget item (IFRS).
3. **Public officials' data - *Mostly unavailable***
    1. Name, contracting authority, position, start and end date, political affiliation.

Key aspects of procurement data

1. **Scope**: The range of transactions the publicly available procurement data covers
   - E.g., publishing threshold
2. **Depth**: Amount of information disclosed for each contracts/tenders
   - Depth of information within each data types (e.g., does budget data available for public organizations or only registry data?)
3. **Quality**: Reliability of the data, share of missing information
4. **Access**: How easy is it to obtain the procurement data?
   - Is there an API or the website has to be scraped?

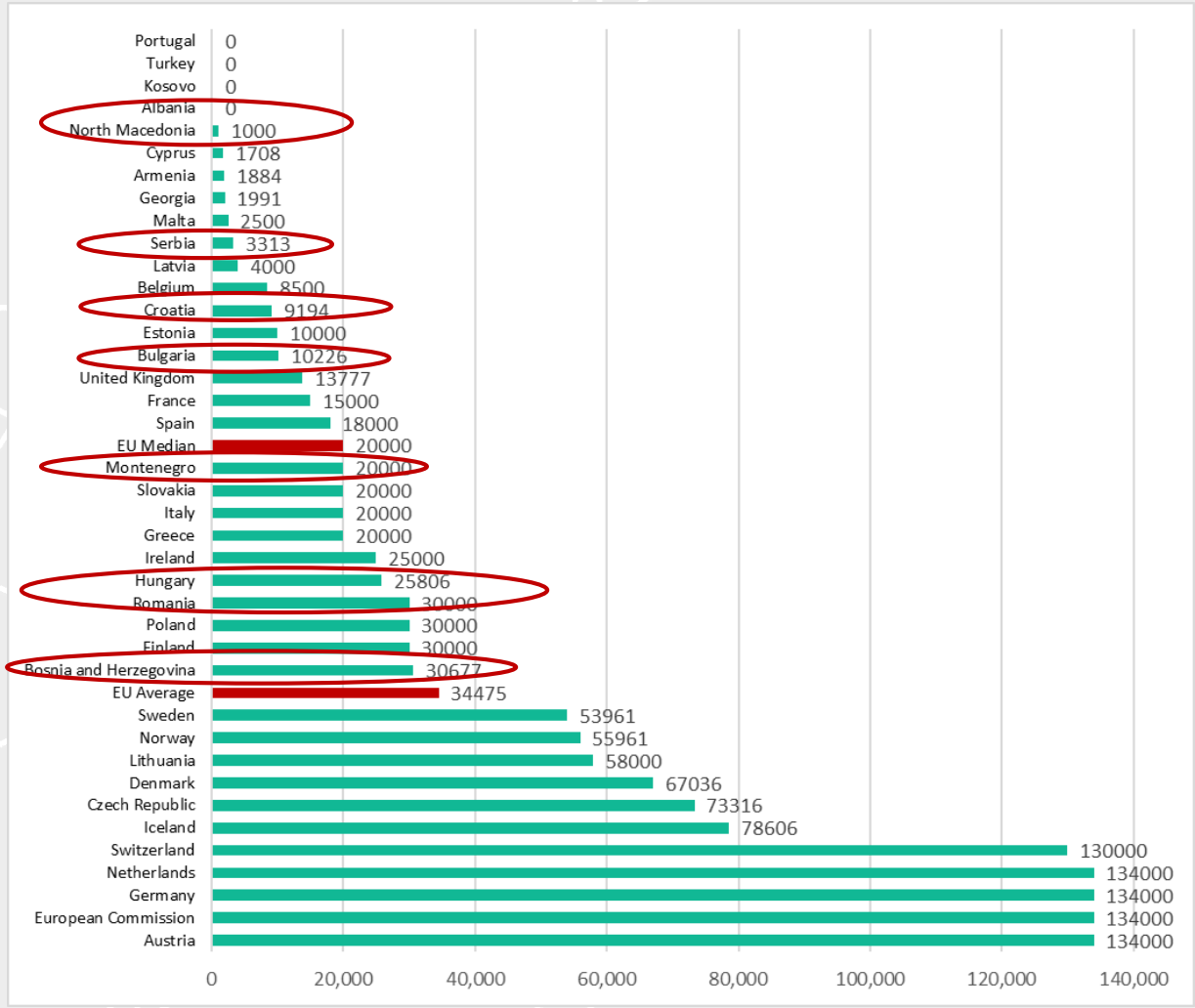
1. Data Scope I. – Reporting thresholds

Reporting thresholds: National contract value thresholds for mandatory publication of tenders on national or EU-wide portals

- Procurements over the threshold also have to comply with stricter rules
  - such as minimum length of advertisement period, or publication of scoring criteria.
  - Hence, lower threshold leads to more transparency.
  - Tenders under the threshold are significantly more likely to have restricted types (e.g., direct awards, negotiated tenders)

Reporting thresholds can have different meaning across countries and across time (e.g., in Turkey several public bodies are exempt from the threshold)

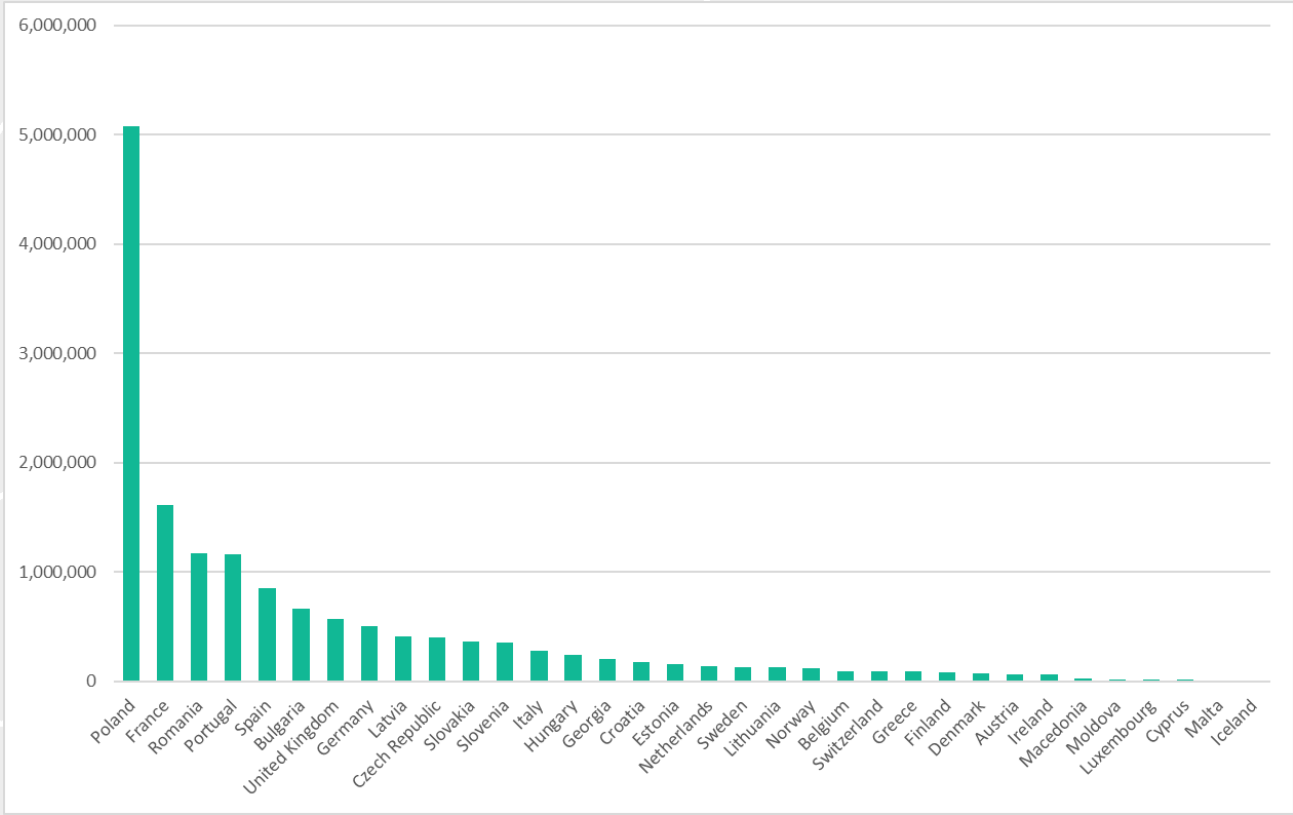
Scope of public procurement databases
Minimum contract value for publishing supplies and services contracts (EUR, 2015)



1. Data Scope II. – Number of tenders processed by DIGIWHIST

- Result of the variation in thresholds is that publicly available data quantity largely differs
- More data leads to better/less biased analysis

Number of contracts collected by DIGIWHIST per country
TED + National data, 2006(2007) - 2020



2. Data depth I. - Tender cycle

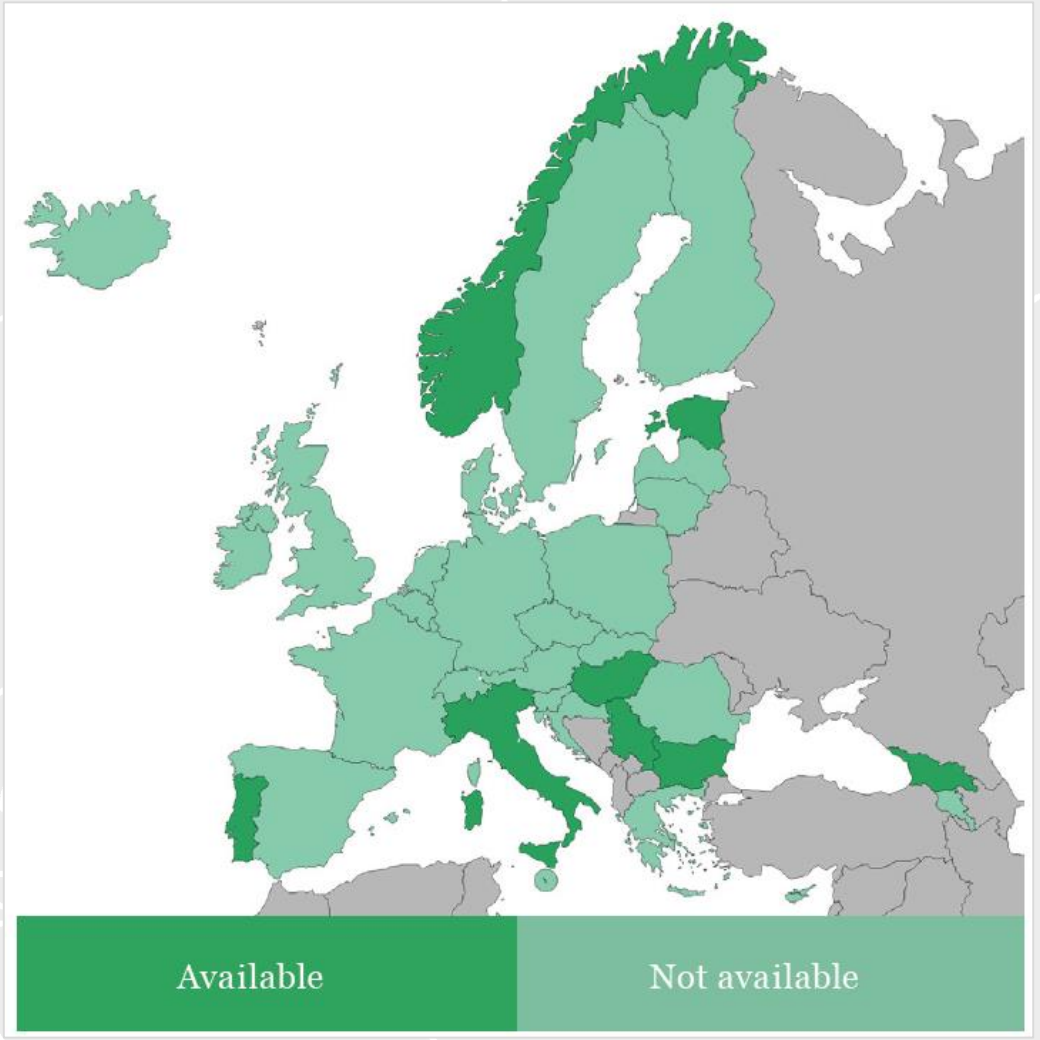
The tender cycle consists of:

2. Data depth II. – Tender cycle coverage

Problems with tender cycle coverage in Europe:

1. Most of the procurement systems in Europe only cover the tendering phase and the awarding phase.
   - Only a handful of EU countries' procurement system disclose information on implementation
   - No information on implementation can give a false picture of the procurement
     - The project can go over budget, or it can be poorly implemented
2. The depth of information within a cycle can vary greatly across countries, due to different (and often changing) legislature
   1. E.g., the UK does not collect bidder number information significantly reducing data usability
3. In many public procurement data systems, modifications and failed tenders are not adequately logged
   - There is no data point indicating tender failure, making failed tenders look like tenders with incomplete information

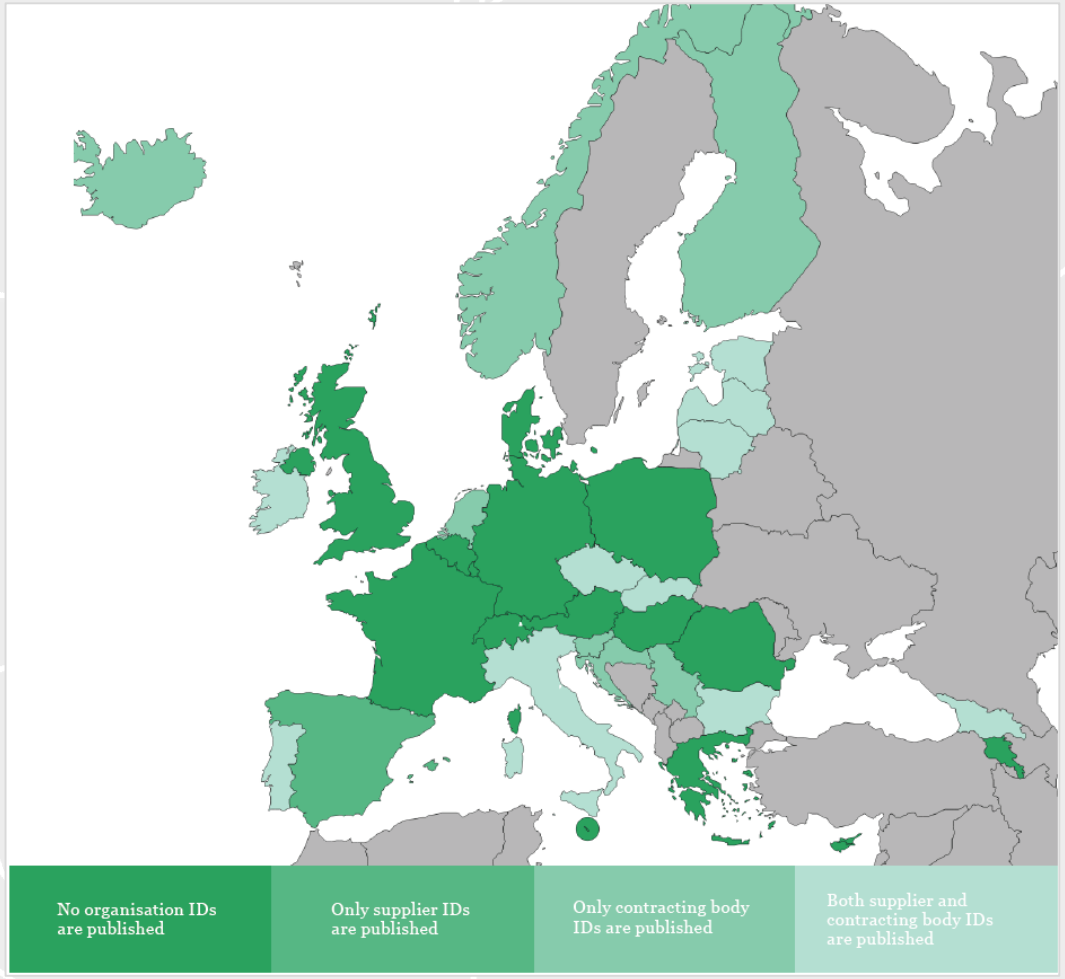
Coverage of the full tender cycle



2. Data depth III. – Organizational IDs

- Many countries only publish the name and location of organization without any unique identifiers
- Organizational IDs for buyers and suppliers are important, to track organizations over time
  - E.g., how different firms perform across different governments
- Without IDs only name-location pairs can be used to identify different buyers and suppliers
  - These can change over time and prone to grammatical errors (typos)

Coverage of organizational IDs



2. Data depth IV. – Minimum data scope

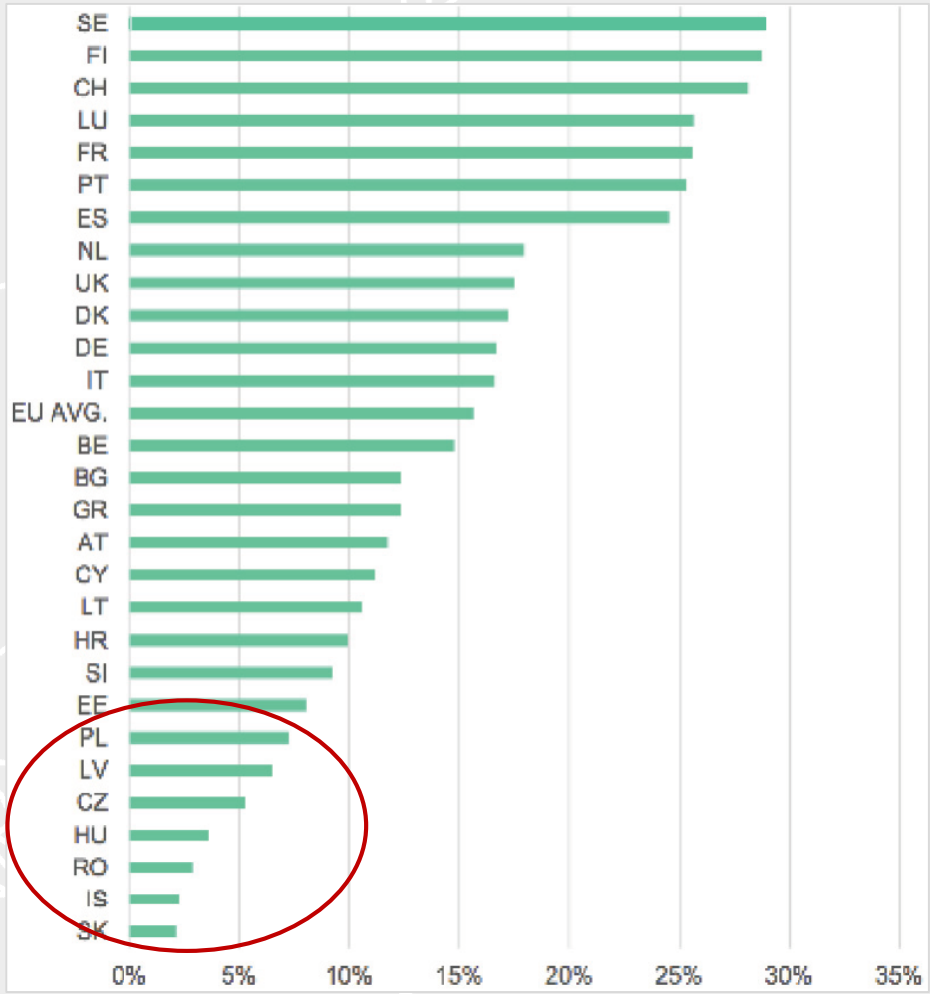
Minimum required information for comprehensive corruption risk assessment

Variable group	Variable
Buyer	Buyer’s name, Buyer’s unique ID, Buyer’s address
Bidder/bids	Bidder’s name, Bidder’s unique ID/tax ID, Bidder’s address, Number of bids submitted, Number of bids excluded, Bid price, Exact time of bid submission, Bid type (winner/loser bid), Beneficial owners
Tender/contract	Procedure type, Framework agreement, Estimated price, Procurement type (service, supply, work), CPV codes, NUTS codes, Status (cancelled, pending etc.)
Dates	Call for tender publication date, Bid submission deadline, Contract start and end dates, Publication date of contract award, Date of contract completion
Subcontracting	Subcontractor’s name and unique ID, Subcontractor’s share
Consortium	Consortium members’ name and unique ID, Consortium member’s unique ID
Contract performance	Contract performance end date, Was performed according to contract, Explanation in case of deferring from contract, Information on contract modification, Information on performance quality

3. Data quality – Share of missing information

- In some countries even legally required administrative information is missing from tender announcements
  - Such as buyer name, tender price, bidder information etc.
- Data quality is low throughout Europe with 15% of mandatory fields are missing on average
- Eastern-European countries pp system fare much better than more developed nations'

Extent of missing information
EU-wide TED data between 2009-2015



4. Data accessibility I. – Extraction method

Goal:

To create structured database from non-structured/semi-structured (text, html, pdf) data

Method:

- Prerequisite is machine readability. (HTML, readable PDF)

1. Web crawling/scraping → collecting the data from the webpage (Python, R)
2. Database creation (JSON, NOSQL, MongoDB)
3. Parsing → automatic text extraction to create data from text (Human assisted) data correction / cleaning, imputation
4. Testing data quality (manual/automatic)
5. Data analysis and indicator creation

	A	B	C	D	E
1		persistent_id	tender_id	tender_title	tender_proceduretype
2	0	EU_50dbf565ff8131df1927afe8ee(	4bceb2d8-53ed-45ac-8a	NAIK - villamos energia	OPEN
3	2	EU_f2ea868c3e38ea9c0a08fcc195(	4c74a037-0eb4-4835-af	MTVA villamos energia beszerzés 2020/2021	OPEN
4	3	EU_12511f9b9a13aac2ff3f0d12cd9	54f5aea9-1352-40dd-85	Földgáz energia kereskedelmi szerződés.	RESTRICTED
5	4	EU_12511f9b9a13aac2ff3f0d12cd9	54f5aea9-1352-40dd-85	Földgáz energia kereskedelmi szerződés.	RESTRICTED
6	5	EU_49d47d2d4659e62074f4d3781a	4be99821-23cd-4fa2-8a	Az Erkel Színház felújítási programjának keret	OPEN
7	6	EU_9176184452d8d26f0e7144eac(	4ba315a4-cc03-4682-8df	Villamos energia beszerzés 2020–2021	OPEN
8	8	EU_f87d6df44288f2fa85f8a113ba9	4bfa97a8-bcf9-414f-91fc	A MÁV Zrt., a MÁV - START Zrt., a MÁV - TRAKCIÓ	NEGOTIATED_WITHOUT_PUBLICATION
9	9	EU_e1b67cc0de74fad6ce6e58bbb(	5593c49b-9a81-4d40-b0	A társult három egészségügyi intézmény részé	NEGOTIATED_WITH_PUBLICATION
10	10	EU_2de6d5260b406248ec63b729d(	61490b25-0eae-4c64-9d	Nyírbátor korszerűsített biomassza alapú fűtő	OPEN
11	11	EU_7778df6293615a6c60dea44555	589e47a2-e531-4a77-84	Földgáz beszerzés.	OPEN
12	12	EU_8d1caa92f3aa241a65110aeed	55efe194-8ad7-4f0c-915	Villamos energia beszerzés.	OPEN
13	13	EU_97322663d29f0e70ee6fe9b954	4d171a55-3667-4468-a0	Földgáz energia beszerzése.	OPEN
14	14	EU_7e7e72c55a32e61957bb925f0(	4bfdfcf4-865d-42ae-9b3	Szállítási szerződés.	OPEN
15	15	EU_1e9c0a01f397a92582b61d5248	00013814-62b3-45f9-9c7	132 kV-os földkábel és 145 kV-os kábelvégelzá	OPEN
16	16	EU_499f62c2afd44825757ca89dbb	00143b95-78ca-4aca-b2	Elektromos anyagok beszerzése	OPEN
17	17	EU_499f62c2afd44825757ca89dbb	00143b95-78ca-4aca-b2	Elektromos anyagok beszerzése	OPEN
18	18	EU_499f62c2afd44825757ca89dbb	00143b95-78ca-4aca-b2	Elektromos anyagok beszerzése	OPEN
19	19	EU_499f62c2afd44825757ca89dbb	00143b95-78ca-4aca-b2	Elektromos anyagok beszerzése	OPEN
20	20	EU_499f62c2afd44825757ca89dbb	00143b95-78ca-4aca-b2	Elektromos anyagok beszerzése	OPEN
21	21	EU_499f62c2afd44825757ca89dbb	00143b95-78ca-4aca-b2	Elektromos anyagok beszerzése	OPEN

4. Data accessibility II. - Machine readability I.

Data can be obtained in a:

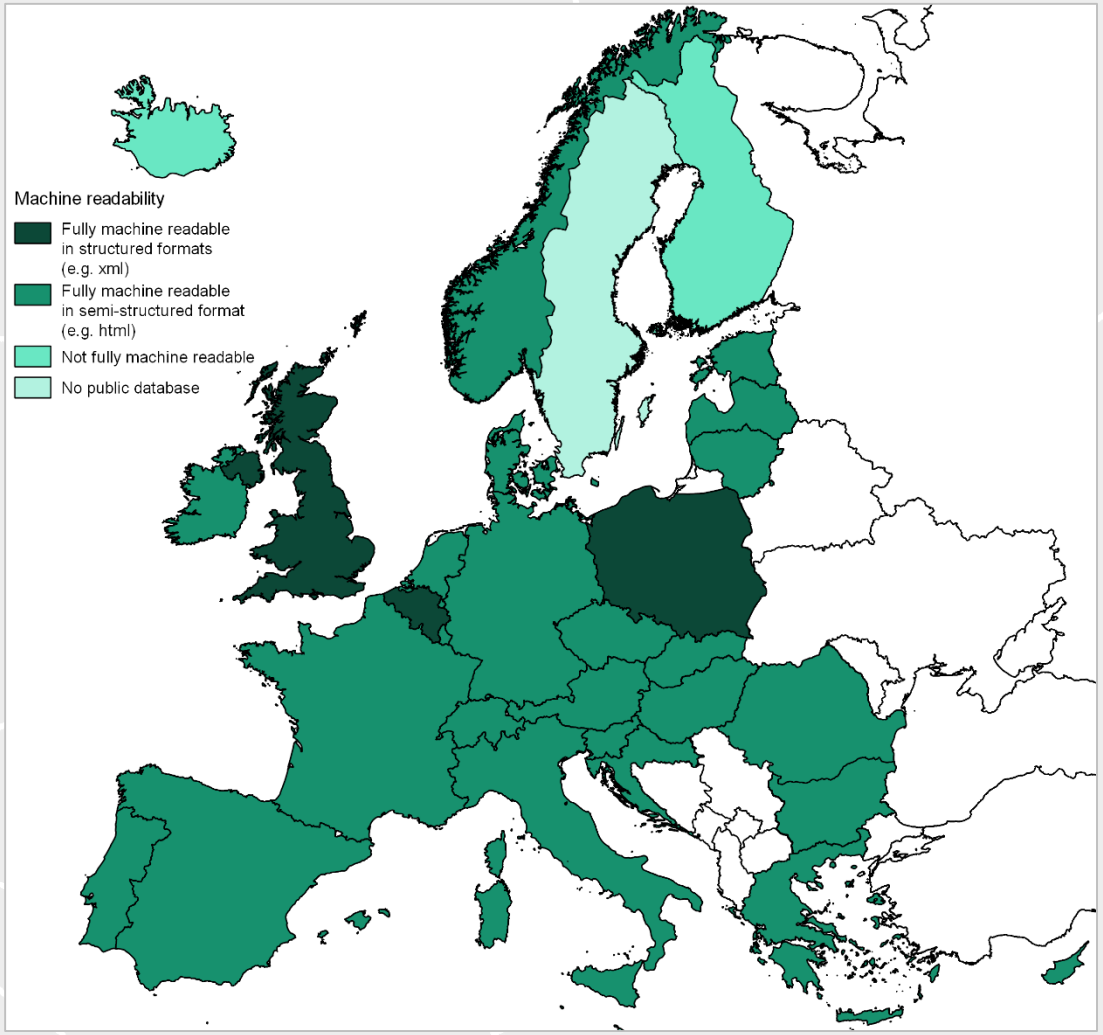
1. Structured format
   - Whole dataset can be downloaded into an excel/json file
2. Semi-structured format (Semi machine-readable)
   - Information is available in a html format, can be scraped and parsed
3. Not fully machine readable
   - Part of the data only accessible by manual cleaning (e.g., scanned pdfs)
4. No public database

4. Data accessibility II. - Machine readability II.

<u>Machine readability & automatization can be problematic:</u>

- Only in three countries and the EU-wide TED's public procurement data is machine readable in a structured format
- In 26 countries, data is only semi machine-readable,
- In 5 countries it is not machine-readable or has no public data av.
- These barriers prevent researchers and NGO's to efficiently analyze the region's public procurement systems

Machine readability of pp databases in European



4. Data accessibility III. – Usual data sources

Sources:

- Open accessibility requires data sources to be publicly available such as:
  - National procurement websites (e-tendering)
  - EU's Tender Electronic Daily (TED)
  - Public organizations' registry and budget data

DIGIWHIST data on opentender.eu:

- Over 40 million public contracts from 32 countries (more to come) in a standardized format
  - JSON, CSV, NDJSON
- Over 5 million government suppliers and 1 million public organizations

More on this in the last section…

Q: Do you know any organization from your country that aggregates procurement data in a similar manner?

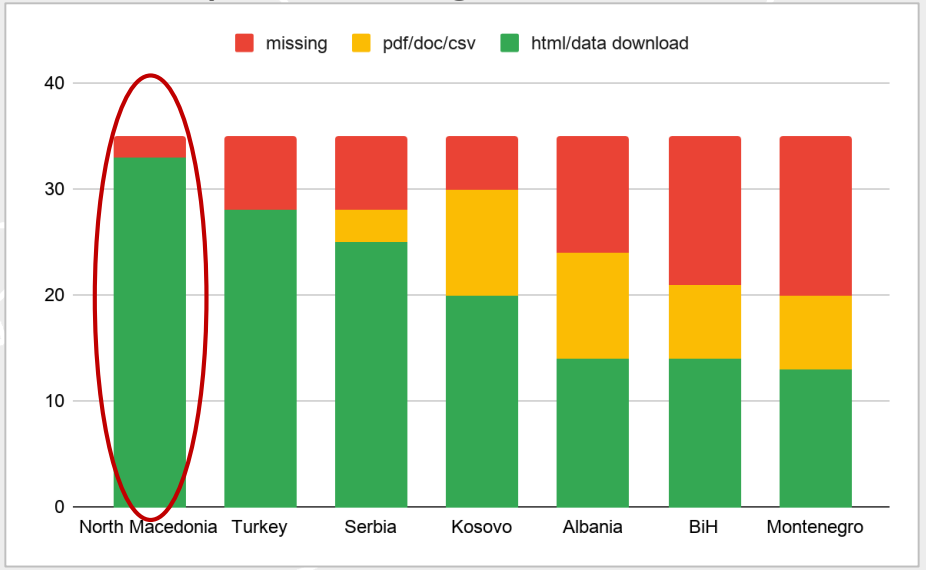
Examples from partner countries I. – North Macedonia

The good:

- The Electronic System for PP (ESPP) was set up and running in 2006 (Longest running system in the region)
- Complete tender documentation is required to publish new tender notice
- The publication includes information on all phases of procurement from planning to contract implementation (this is hard to get).
- Compared to the other countries in the Western Balkans, North Macedonia performs highest on accessibility and usability of standard data fields

Points for improvement:

- Most of the organization ID-s are missing
  - 71% of buyers and 99% for suppliers
- Adding full data download (or API) could further improve accessibility



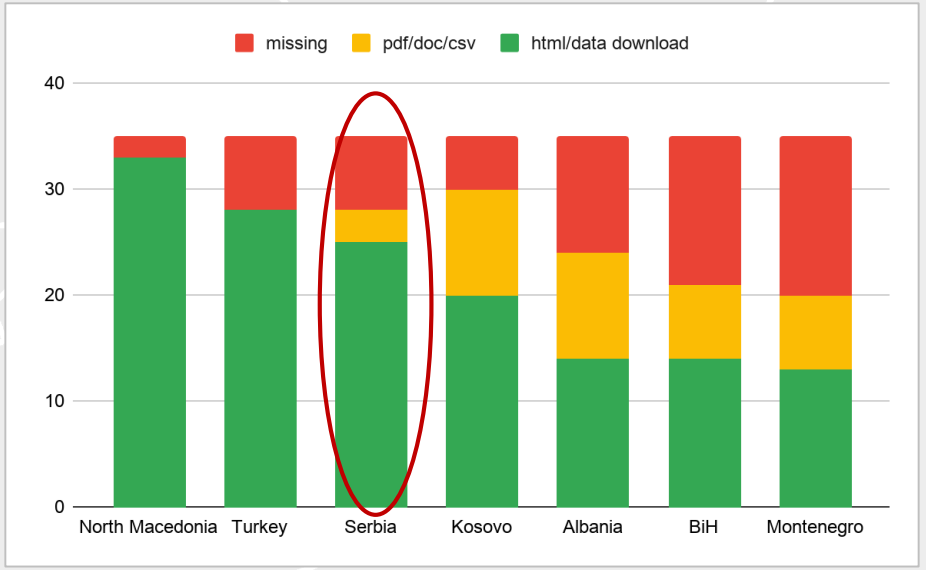
Examples from partner countries II. – Serbia

The good:

- New procurement website since 2020.
- Certain information on tenders and contracts is available in a semi-machine readable format (html) in both the old and new websites
- The new portal gives an opportunity to download data in XLSX, XML or Json formats
- Organization IDs and tender number is available, which allows matching databases

Points for improvement:

- Only new tenders are recorded in the new website
- Some of the attachments are non-machine readable (e.g., scanned pdf-s)
- Downloadable datasets only include buyer name, tender ID and title, date of publication, type of announcement
- Every other information must be traced back from the website



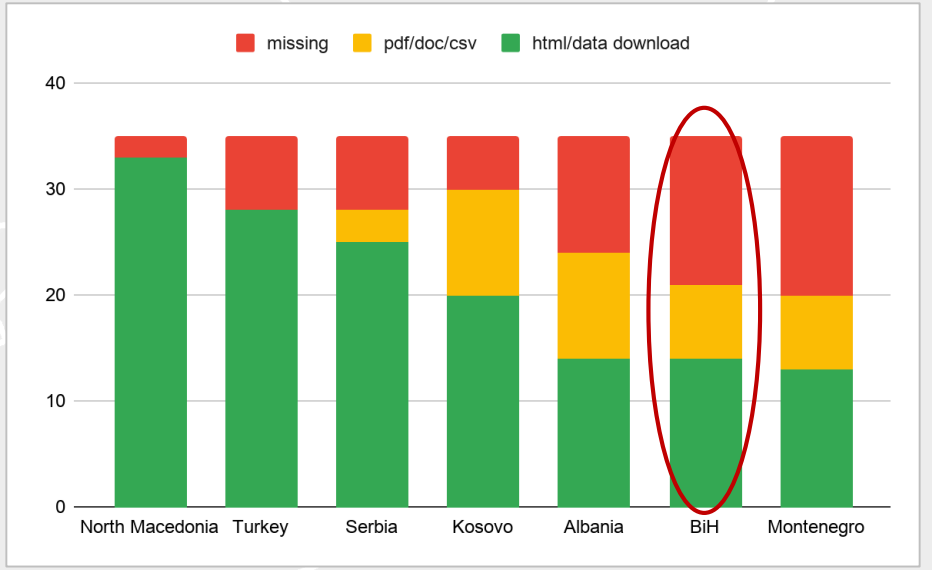
Examples from partner countries III. – Bosnia and Herzegovina

The good:

- Certain information on tenders and contracts is available in a semi-machine readable format (html)
- All the information on buyers and bidders (name, ID, address, agency type) is published in standardized (and readable) pdf format, the same is true for contract details, number of bids, eligibility details and deadlines.

Points for improvement:

- More than ⅓ of the pp data is missing
- Most of the details only available in PDF (not html)
- The source lacks pre-tender information such as procurement plans, as well as details on supplier's performance or contract completion
- The details provided in PDFs differs by tender



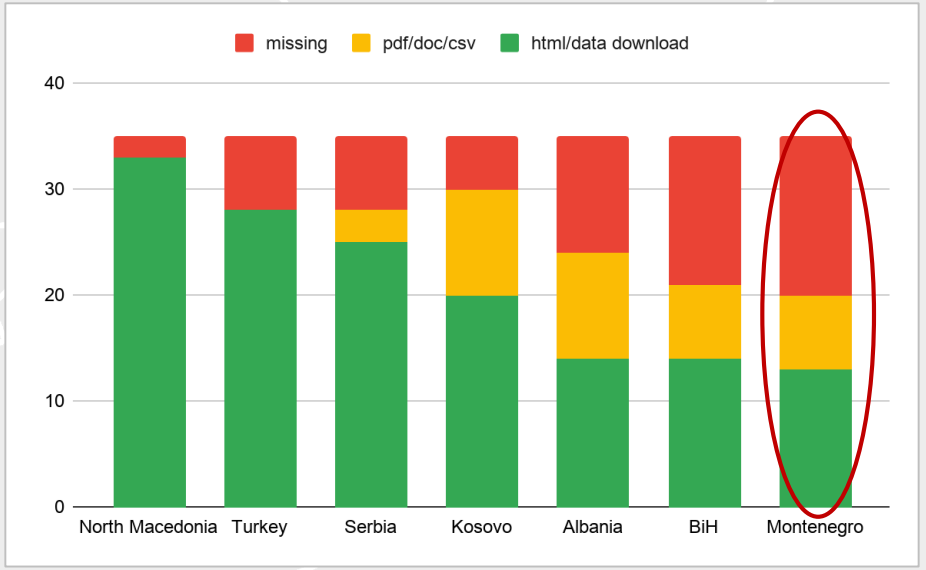
Examples from partner countries III. – Montenegro

The good:

- Certain information on tenders and contracts is available in a semi-machine readable format (html)
  - Type of procurement, price details such as estimated value and currency
- Limited information can be exported in CSV, XLSX, XLM and PDF format

Points for improvement:

- More than ⅓ of the pp data is missing
- Most of the information is provided in many separate world/pdf files.
  - Many files are scanned, and badly structured lengthy documents
  - Even within one tender the types of the documents may differ
- Bidder and buyer IDs are always absent
- Exportable information is limited and only one page can be downloaded at once



Q: What are your experiences in your own countries?

- *Is the depth of published contract details are good enough?*
- *Are company names traceable (across time & different contracts)?*
- *Is the quality of the procurement website satisfactory?*

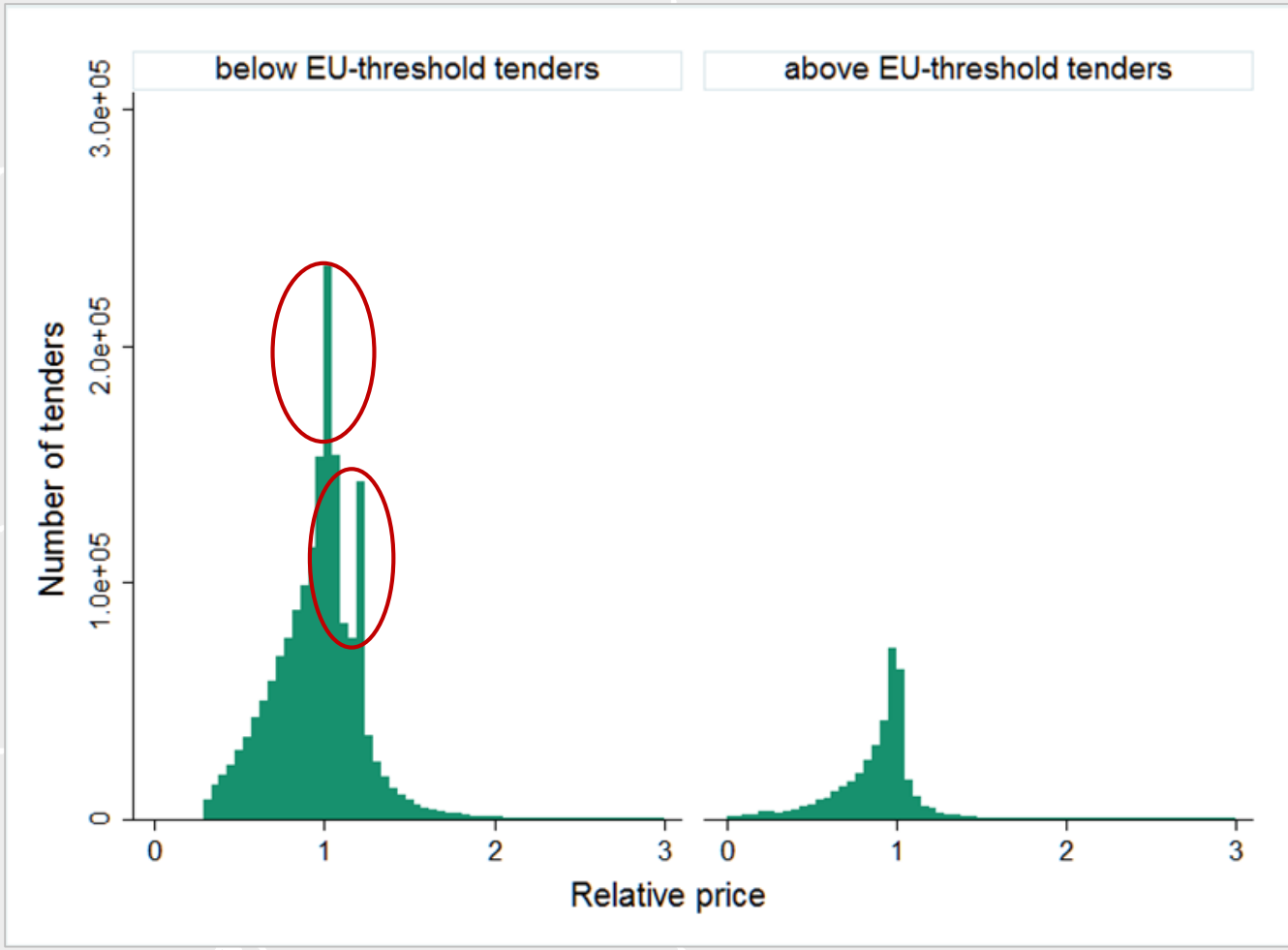
Possible errors in the data I. – Common errors

Common errors:

- *Missing attributes* → No information provided (not necessary an error).
- *Lexical error* → The value is provided is not consistent with the column name (e.g., country id column shows currency id)
- *Irregularity error* → E.g., the unit of measurement differs from the other observations'
- *Formatting error* → E.g., date is in different time format leading to errors when data is loaded
- *Duplication error* → There are duplicate observations in the data; each variable is the same
- *Contradiction error* → Two columns measuring (almost) the same thing show different values for the same observation.
- *Outlier* → Given variable for a given observation is significantly different from the others (not necessary an error, but usually should be delt with)

Possible errors in the data II. – Example of lexical error

- Figure shows relative price distribution of tenders below and above the EU-threshold
- The distribution has two „peaks” because in some cases prices were recorded with VAT even though a net value should have been recorded



Data wrangling good practice

No dataset is unique to a different set of errors; hence it is always important to:

- Explore the data before deeper analysis (check column values, distribution for numeric columns, averages, etc.).
- Only use columns that are sufficiently “clean” and not much is missing (~15%)
- Cross-check/sanity-check every result
- Use alternative – qualitative – methods such as expert interviews
- **Procurement data is always just an approximation of reality**
  - With many information missing or contradicting we cannot see the whole picture, hence all results should be treated in their proper place

Thank you!

Questions?

Sources

- Mara Mendes, Mihály Fazekas (2015): DIGIWHIST Recommendations for the Implementation of Open Public Procurement Data - An Implementer‘s Guide
- Mihály Fazekas, Luciana Cingolani, Bence Tóth (2016): A comprehensive review of objective corruption proxies in public procurement: risky actors, transactions, and vehicles of rent extraction, Working Paper series: GTI-WP/2016:03, link
- Mihály Fazekas (2021): Using Macro Analytics (Big Data Analytics) to Monitor Fraud and Corruption in Public Procurement In: Using Data Analytics To Combat Fraud And Corruption In Public Procurement In Healthcare In Eastern Europe, Ceeli Institute, link
- Mihály Fazekas, Ágnes Czibik (2021): Measuring regional quality of government: the public spending quality index based on government contracting data, Regional Studies, DOI: 10.1080/00343404.2021.1902975
- Mihály Fazekas (2021): Corruption Risks in Public Procurement in the Western Balkans and Turkey (Draft)
- Nicolas Rangeon, Anderson Prewitt (s.a.): Identify Different Types of Errors, link