Iceland Liechtenstein Norway grants

Norway grants

Péter Horn, GTI

Hands-on training: How to use public procurement data to identify PC companies?

Second specialised three-day regional training for R2G4P members

18-20 May 2022

Implemented by:

CSD CENTER FOR THE STUDY OF DEMOCRACY

Regional Cooperation Council

SELDI.net Southeast European Leadership for Development and Integrity

Government Transparency Institute

Iceland
Liechtenstein Norway
Norway grants grants

Outline

- **Procurement datasets**
    - Data availability, main variable groups, important filters & data cleaning
- **Integrity indicators**
    - Integrity indicator validation, main indicators
- **Q&A + short break**
- **Opentender and code-along session for the WP3 report**
- **Q&A session**

Procurement datasets

Procurement data availability

- Procurement data is available on the opentender.eu website for more than 30 countries.
- Data availability for Partner countries:

Fully available	Some data available	Not available
• Romania • Bulgaria • Hungary • North Macedonia • Croatia	• Serbia (currently scraping) • Albania (very limited data)	• Bosnia and Herzegovina • Montenegro

- Data is available in CSV and JSON formats → JSON is a NoSQL (non-tabular/non-relational database, therefore it can hold more information (discussed in end of seminar)
- Opentender not always has the most recent datasets, to get the most recent & cleaned datasets please contact us

Main variable groups (in CSV)

Tender specific variables	e.g. title, procedure type, supply type, CPV codes, final price, framework agreement
Buyer specific variables	e.g. name, location, type
Bidder specific variables	e.g. name, location, type
Lot specific variables	e.g. title, price, bid count, status
Integrity indicators	Calculated by GTI, e.g., single bidding, decision period, call for tender publication

- Datasets are on the lot level, therefore tender and buyer specific information are duplicated as many times as many lots a tender has.
  - Data should be deduplicated before using tender level variables (e.g., tender prices)!

Filters & important data transformations before analysis I.

- **Datasets are never perfect, therefore it is important to clean before analysis**

General cleaning/filtering steps:

- *Largely missing columns (e.g., bidder id/buyer id)*
  - Largely missing columns cannot be used, because it can restrict the scope of the analysis, and missing values might not be random.
  - Use alternative column instead (e.g., bidder name instead of bidder id; buyer NUTS code instead of buyer city)
- *Outliers*
  - Outliers should be identified and dropped or censored (capped at a maximum value)
- *Currency matching/PPP adjustments*
  - Price type variables should be adjusted to be denominated in the same currency
- *Duplicated values*
  - Duplicated values should be removed (use the lot_id/bid_id variable)
- *Data type transformation*
  - Check if numeric variables are strings, optionally transform categorical/ordinal variables from strings

Filters & important data transformations before analysis II.

Procurement data specific cleaning/filtering steps

- *Missing bidders*
  - As tender/lot status variables are often lacking rows with missing bidder names are often dropped, assuming these contract have no valid bids or tender/lot has been cancelled.
- *Opentender variable*
  - Special variable: Meant to deduplicate tenders that are both available in TED and national datasets. (For EU countries it is required to upload contracts to TED above threshold)
  - Keep only if opentender = „t" (or true)
- *Losing bids*
  - Some countries also report losing bids, for the current analysis these can be removed
  - Keep if bid_isWinning variable is true, yes
- *Framework filter*
  - New feature (not on opentender yet): if there is contracted value for framework procurement it keeps only that value, otherwise keeps all framework value as the contract value
  - Keep framework_filter = 1

Iceland Liechtenstein Norway grants Norway grants

Integrity indicators

Iceland Liechtenstein Norway grants   Norway grants

Qualities of integrity indicators

Main qualities:

1. They are based on the initial assumption that certain quantifiable features of public procurements are able to predict their corruption risk level.
2. Integrity indicators are selected based on thorough qualitative and quantitative research. Several working papers and published papers show the effects of these indicators on corruption risk.
3. Some features are equipped to measure corruption risks in
    1. the procurement planning & advertisement phase (e.g., procedure type, call for tender publications, submission period)
    2. Some in the selection phase (e.g., decision period, single bidding).

Validating integrity indicators

- Before being used in empirical research, each potential indicator must be tested using **national** procurement datasets
- Single bidding is the most robust (and most thoroughly tested indicator)
  - Therefore it is used to validate other indicators
  - Each indicator is validated on its country-level dataset based on its association with single bidding
- Two step validation process
  - Regressing „raw” indicator on single bidding with additional controls (such as buyer type, log contract value, market fe, year fe., etc.)

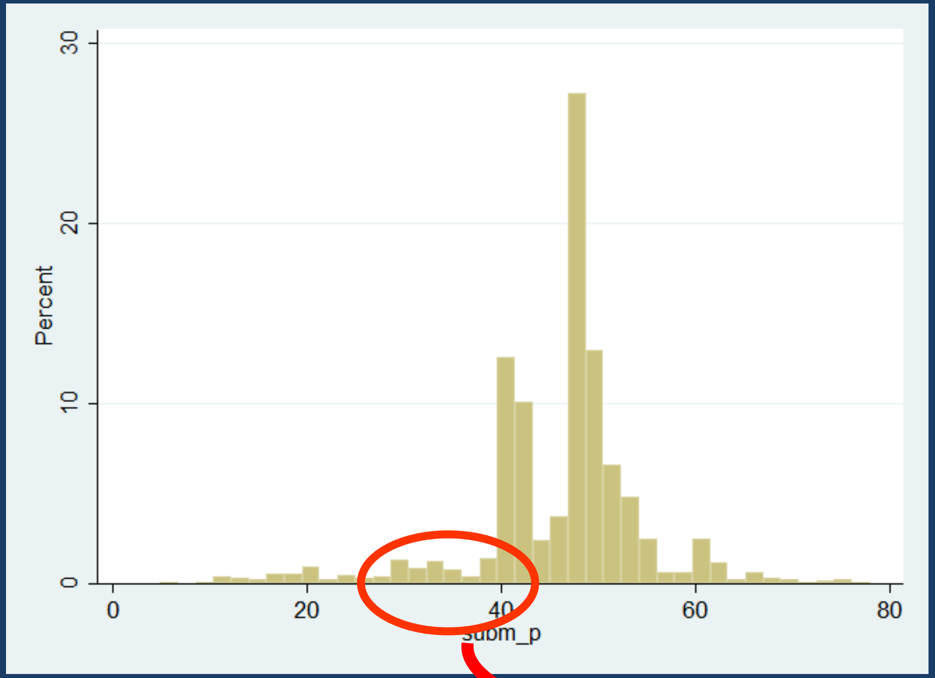
(1) $\boldsymbol{\mathit{Single\ bidding\ integrity} = B0 + B1 * \mathit{indicator} + B2 * \mathit{institutional\ and\ market\ controls} + \varepsilon}$

  - If coefficients are significant and intuitive, they are sorted into into high-, medium and low-integrity categories based on their association with single bidding. Then regressed on single bidding again.

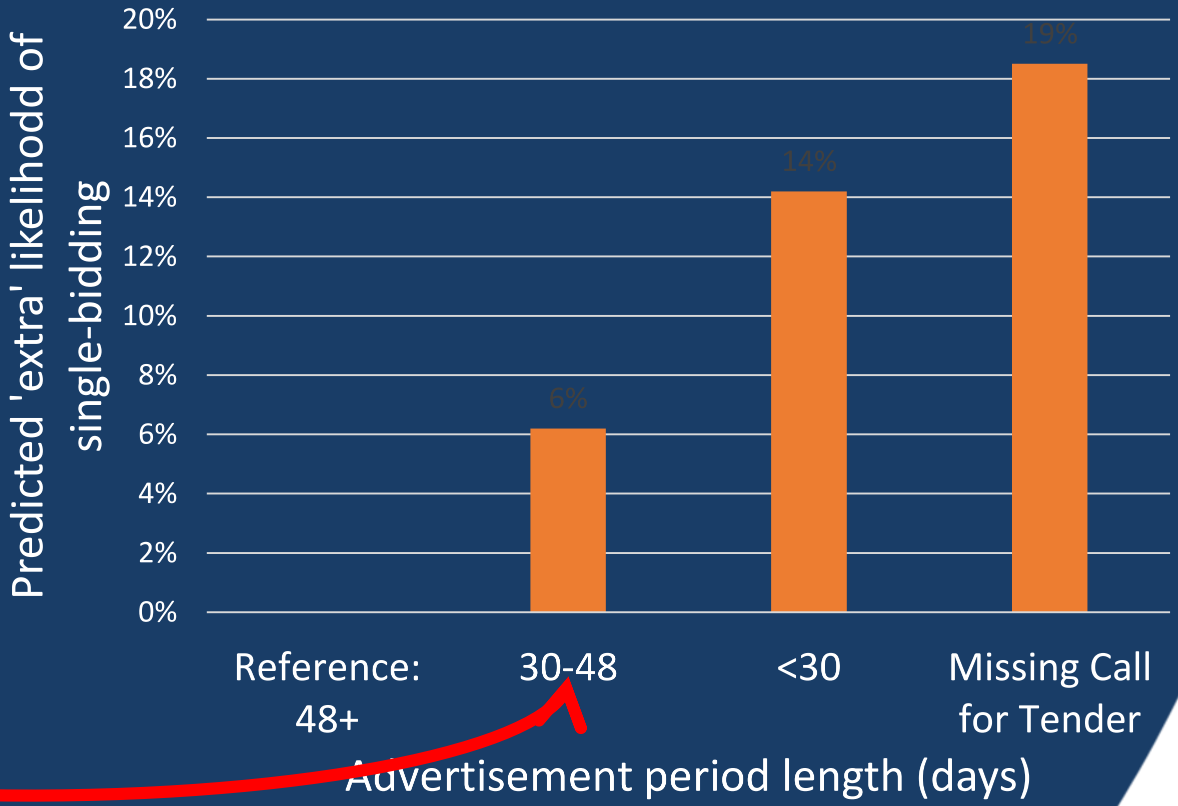
(2) $\boldsymbol{\mathit{Single\ bidding\ integrity} = B0 + B1 * \mathit{integrity\ indicator} + B2 * \mathit{controls} + \varepsilon}$

Setting thresholds for continous indicators

Distribution of contracts by advertisement period length

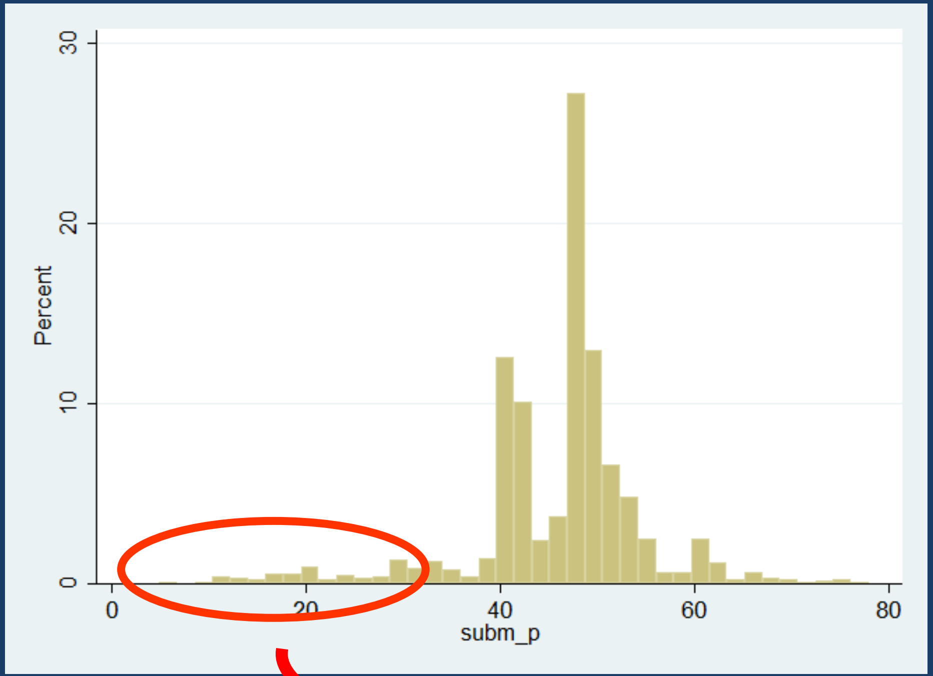


Likelihood of single-bidding

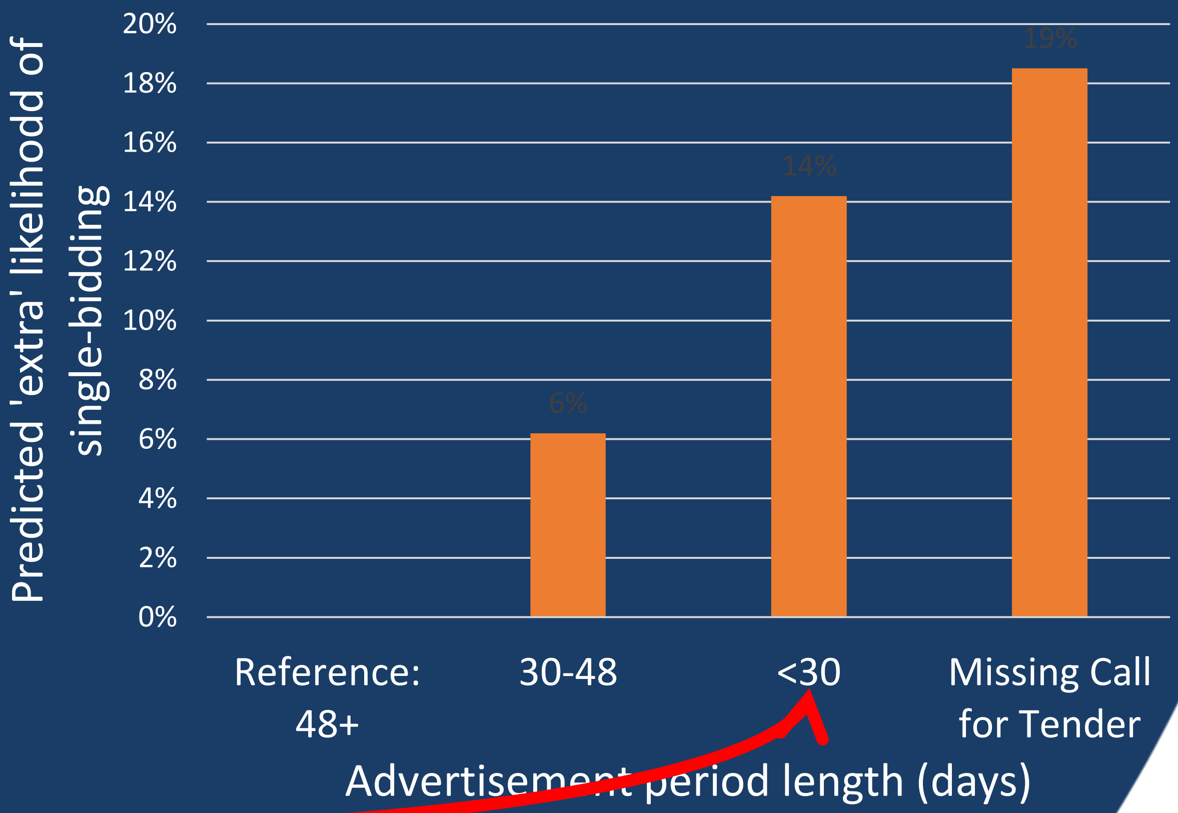


Setting thresholds for continous indicators

Distribution of contracts by advertisement period length



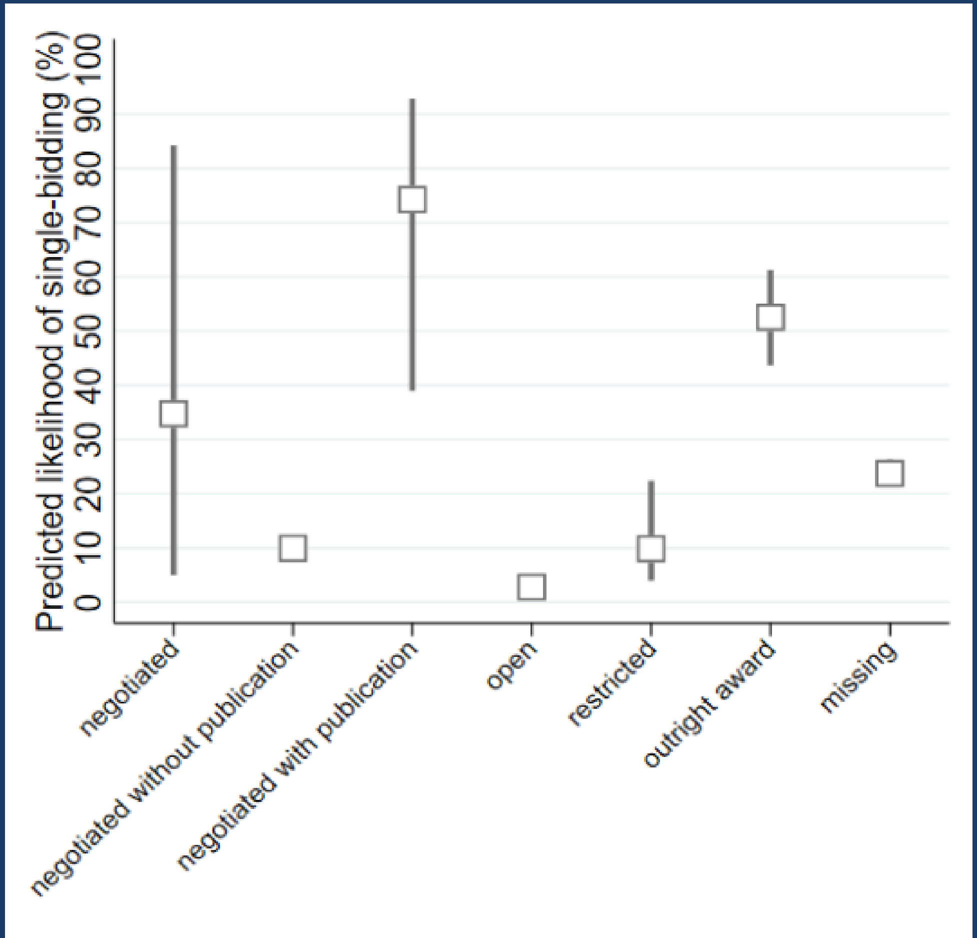
Likelihood of single-bidding



Iceland Liechtenstein Norway grants

Norway grants

Setting thresholds for categorical indicators



Main integrity indicators

- Only those indicators are kept for each country that correlate with single bidding and robust enough to stay statistically significant in the full model (containing all the integrity indicators)

Some of the main indicators:

Advertisement period length (country specific)	100 = length of advertisement period is unrelated to corruption risks 50 = length of advertisement period has intermediate relationship with corruption risks 0 = length of advertisement period or missing advertisement period has a strong relationship with corruption risks
Decision period length (country specific)	100 = length of decision period is unrelated to corruption risks 50 = length of decision period is somewhat related to corruption risks 0 = length of decision period OR missing decision period is related to corruption risks
Single bid	100 = more than 1 bid received 0 = 1 bid received
Call for tender	100 = call for tender/prior information notice published in official journal 0 = NO call for tender/prior information notice published in official journal
Procedure type (country specific)	100 = open, or does not have significant effect on single bidding 50 = negotiated 0 = non-open + has significant effect on single bidding
Tax haven	100 = winning bidder is not registered in a tax haven country, and is a foreign bidder 0 = company is registered in a tax haven country
(New company) – many missing	100 = if company is older than 1 year when winning a public contract 0 = if company is younger than 1 year when winning a public contract

Example of integrity indicators

- Due to country specific validation of integrity indicators, their definition can differ.
- However, the logic behind each of these indicators remains the same; for example, in each country relatively short decision and submission/advertisement periods are considered risky, and non-open procedure types are also associated with low integrity

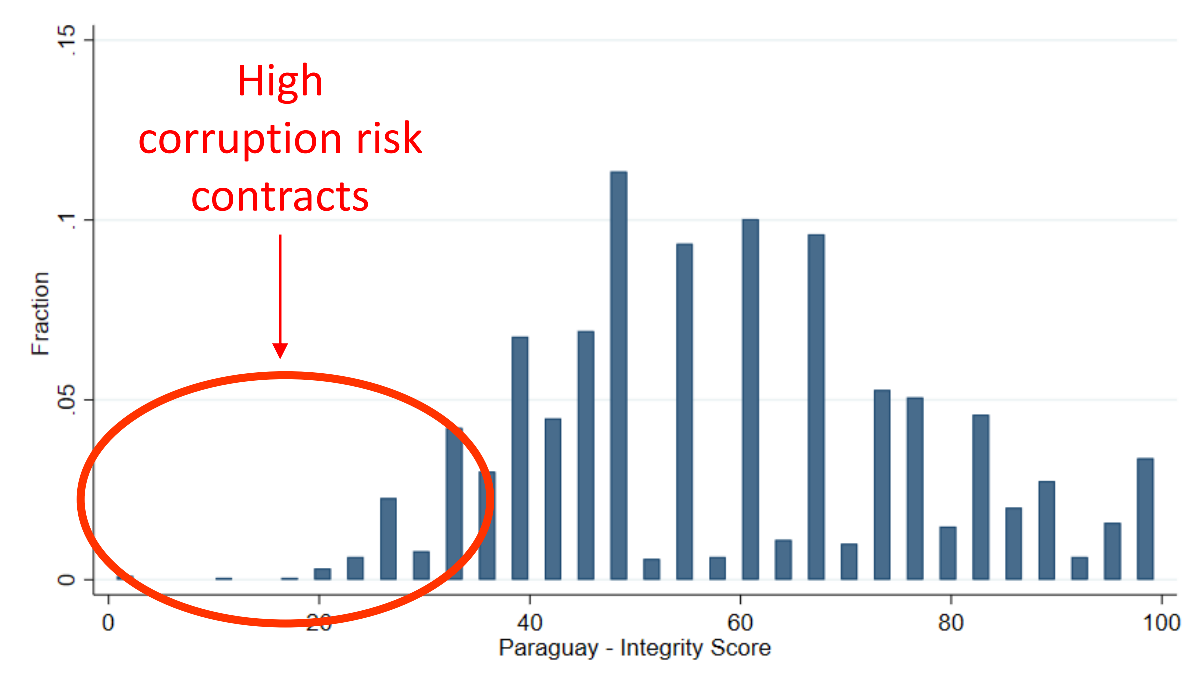
Example: Croatian integrity definitions:

Indicator	High integrity (100)	Medium integrity (50)	Low integrity (0)
Single bidding	1+ bidders	-	1 bidder
Cft publication	Published Cft	-	Not published Cft
Procedure type	Open, Negotiated, Competitive dialog, Restricted, Negotiated with publication	-	Negotiated without publication, Missing
Advertisement period	34+ days		0-33 days
Decision period	93+ days	53-92 days	0-52 days
Tax haven	Supplier not in tax haven country		Supplier in tax haven country

Integrity Score

- The score is the simple arithmetic average of the individual integrity indicators, falling between 0 and 100, with 100 representing the highest possible integrity and 0 the lowest

Example: Paraguayan Integrity Score distribution



Iceland
Liechtenstein Norway
Norway grants grants

Q&A session

Opentender guide for company filtering

15 mins. break

Code-along exercise for WP3

Iceland Liechtenstein Norway grants Norway grants

First steps (if you would like to code-along with me)

1. Create a similar folder structure
   - Folder name: Code
     - *File: EEA_asset_declarations.R*
   - Folder name: Data
     - *File 1: HU_mod_ind202203.csv.gz*
     - *File 2: pc_firms.csv*
   - Folder name: Output
2. Open the EEA_asset_declarations.R using Rstudio
   - It will work only if you have downloaded both R and Rstudio
   - If R-studio is not set as default you might need to: right-click->Open with ->Rstudio

Iceland
Liechtenstein Norway
Norway grants grants

Q&A session

Iceland
Liechtenstein
Norway grants

Norway
grants

Thank you for your attention!

Iceland
Liechtenstein
Norway grants

Norway
grants

EXTRA: Gentle introduction to the JSON file format

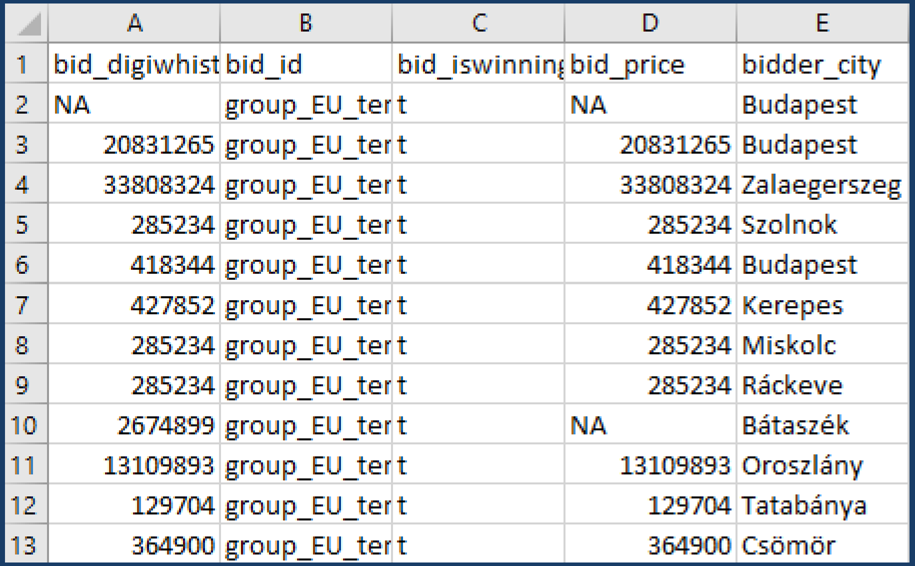
What is JSON?

- „JSON (JavaScript Object Notation) is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate.” (json.org)
- It is a NoSQL data format, meaning that it is non-tabular/non-relational database
- Instead it holds „key-value” pairs or properties:
    - The object is surrounded by curly braces {}
    - Every key-value pair is separated by a comma
    - A key-value pair consists of a key and a value, separated by a colon (:)
    - The key is a string, which identifies the key-value pair, therefore at each „level” keys are unique
    - The value can have several data types such as string, number, float, array, boolean, etc.

Example data structure

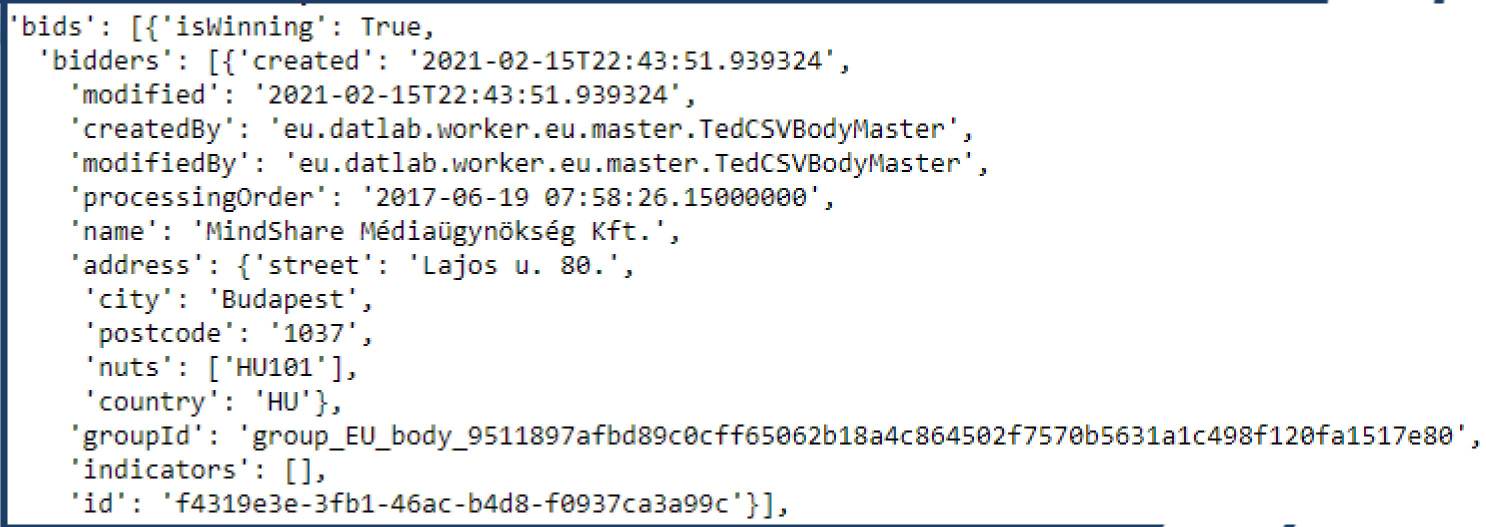
- To access a specific value, have to navigate to the appropriate data „layer/level”
- For example in Python programing language
    - json_data[1030]["lots"][0]["bids"][0]["bidders"][0]["name"]
- Or you can use database programes such as MongoDB to view and filter JSON data

Tabular/relational database (CSV)



	A	B	C	D	E
1	bid_digiwhist	bid_id	bid_iswinning	bid_price	bidder_city
2	NA	group_EU_ter	t	NA	Budapest
3	20831265	group_EU_ter	t	20831265	Budapest
4	33808324	group_EU_ter	t	33808324	Zalaegerszeg
5	285234	group_EU_ter	t	285234	Szolnok
6	418344	group_EU_ter	t	418344	Budapest
7	427852	group_EU_ter	t	427852	Kerepes
8	285234	group_EU_ter	t	285234	Miskolc
9	285234	group_EU_ter	t	285234	Ráckeve
10	2674899	group_EU_ter	t	NA	Bátaszék
11	13109893	group_EU_ter	t	13109893	Oroszlány
12	129704	group_EU_ter	t	129704	Tatabánya
13	364900	group_EU_ter	t	364900	Csömör

Snippet of JSON „dictionary”



```
'bids': [{'isWinning': True,
  'bidders': [{'created': '2021-02-15T22:43:51.939324',
    'modified': '2021-02-15T22:43:51.939324',
    'createdBy': 'eu.datlab.worker.eu.master.TedCSVBodyMaster',
    'modifiedBy': 'eu.datlab.worker.eu.master.TedCSVBodyMaster',
    'processingOrder': '2017-06-19 07:58:26.15000000',
    'name': 'MindShare Médiaügynökség Kft.',
    'address': {'street': 'Lajos u. 80.',
     'city': 'Budapest',
     'postcode': '1037',
     'nuts': ['HU101'],
     'country': 'HU'},
    'groupId': 'group_EU_body_9511897afbd89c0cff65062b18a4c864502f7570b5631a1c498f120fa1517e80',
    'indicators': [],
    'id': 'f4319e3e-3fb1-46ac-b4d8-f0937ca3a99c'}],
```