How much do we need to clean?

Chvalkovská, Jana Gutierrez
2012

Dostupný z http://www.nusl.cz/ntk/nusl-126844

zIndex.cz
How much do we need to clean?

Jana Gutierrez Chvalkovská, Jiří Skuhrovec

Center of applied economics, Datlab, Institute of Economic Studies of Charles University

Topic of the presentation

- About me - very briefly
- The story of zIndex – would it work without data cleaning?
- Other projects
- Data to go – sources of data that are not yet clean, but they should…

About me

- Studied law and economics at the Charles University in Prague
- For last nearly 6 years works as:
  - Market analyst
  - Consultant - mainly for public sector
- During last 2 years started to:
  - Together with Jiří Skuhrovec run a anti-corruption NGO
  - Perform manual data cleaning, software testing, raise public funds, selling databases
- I am practical user of both cleaned and raw data, with no IT knowledge

The story of zIndex.cz

- zIndex is a research project which rates public institutions according to quality and transparency of their tender competitions
- Started as research project, but quickly evolved due to dual use of IT and know-how of public procurement
- Included screen scraping, data cleaning and data mining and analytical work

The story of zIndex.cz

1a. Necessity to find some data for a dying PhD research in economics

1b. Depressed by elaborating numerous fruitless bids for public tenders and by large-size tenders won by companies like Deloitte (CZK 280 mil. for advisory concerining highway tolls)

2. Data available online, but only as html

3. Screen scraping of the data

The story of zIndex.cz

1a. Necessity to find some data for a dying PhD research in economics

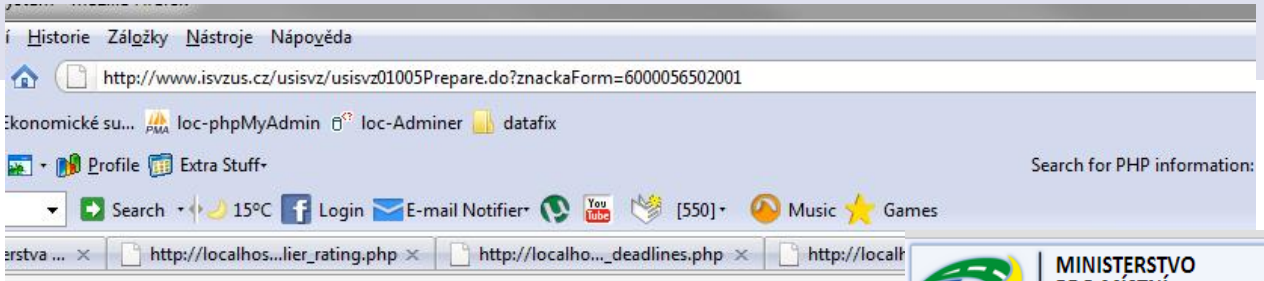
1b. Depressed by elaborating numerous fruitless bids for public tenders and by large-size tenders won by companies like Deloitte (CZK 280 mil. for advisory concerinin highway tolls)

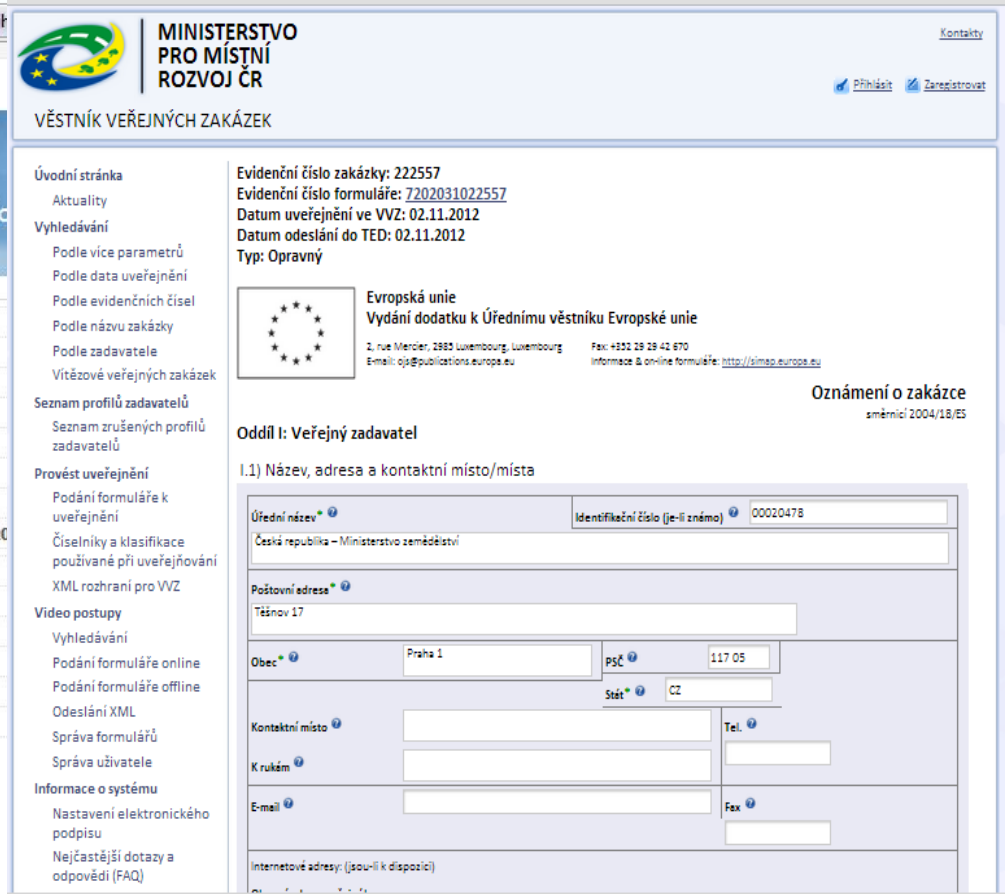
2. Data available online, but only as html

3. Screen scraping of the data

4. Database full of mistakes – version 1

The story of zIndex.cz





The story of zIndex.cz

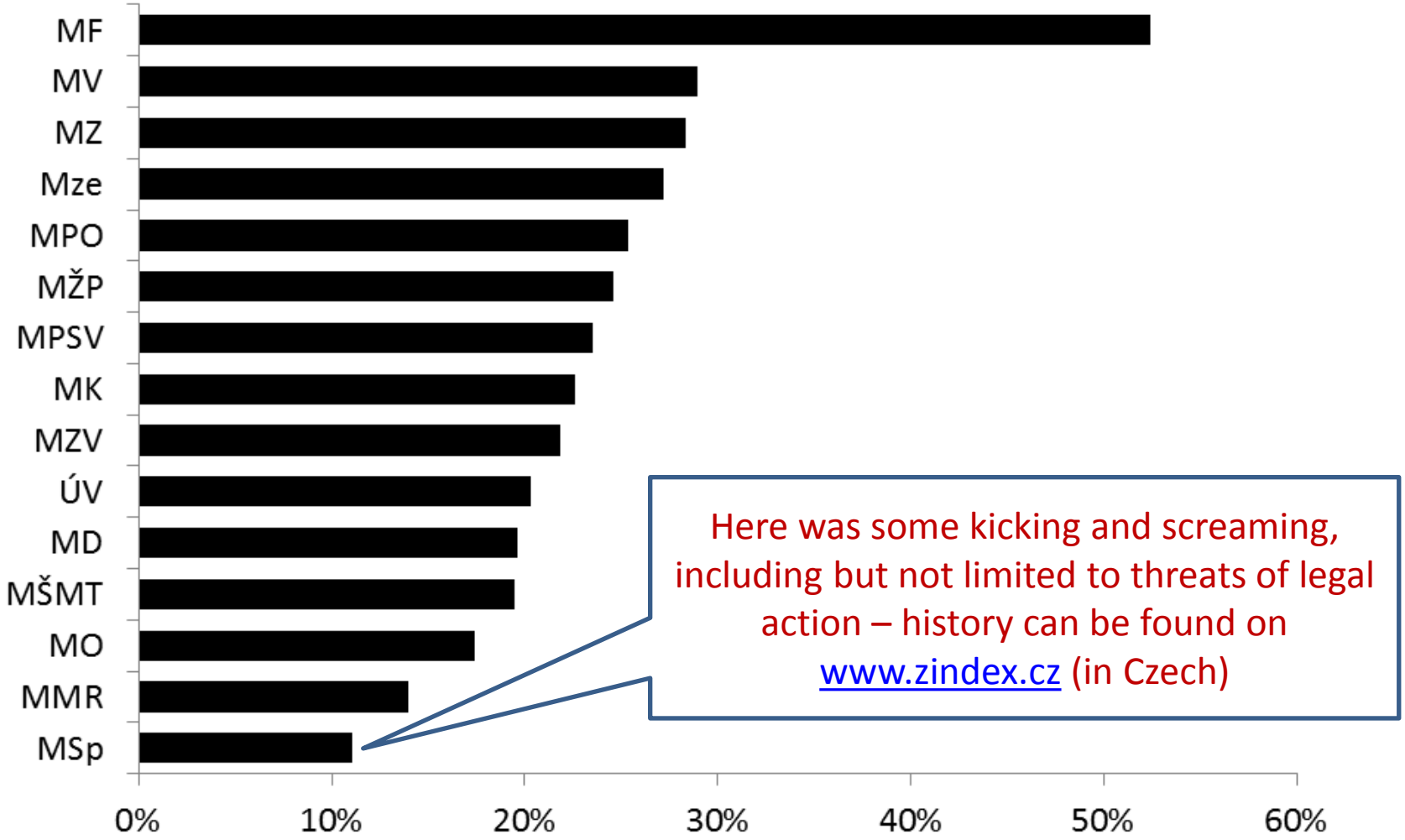
5. – 10. Data cleaning. Why?

a) Because it increases the credibility of the project – especially if it is an anti-corruption (high probability that public bodies will kick and scream)

b) Because otherwise the data do not fulfill their purpose (as they do not in case of the official governmental webpage)

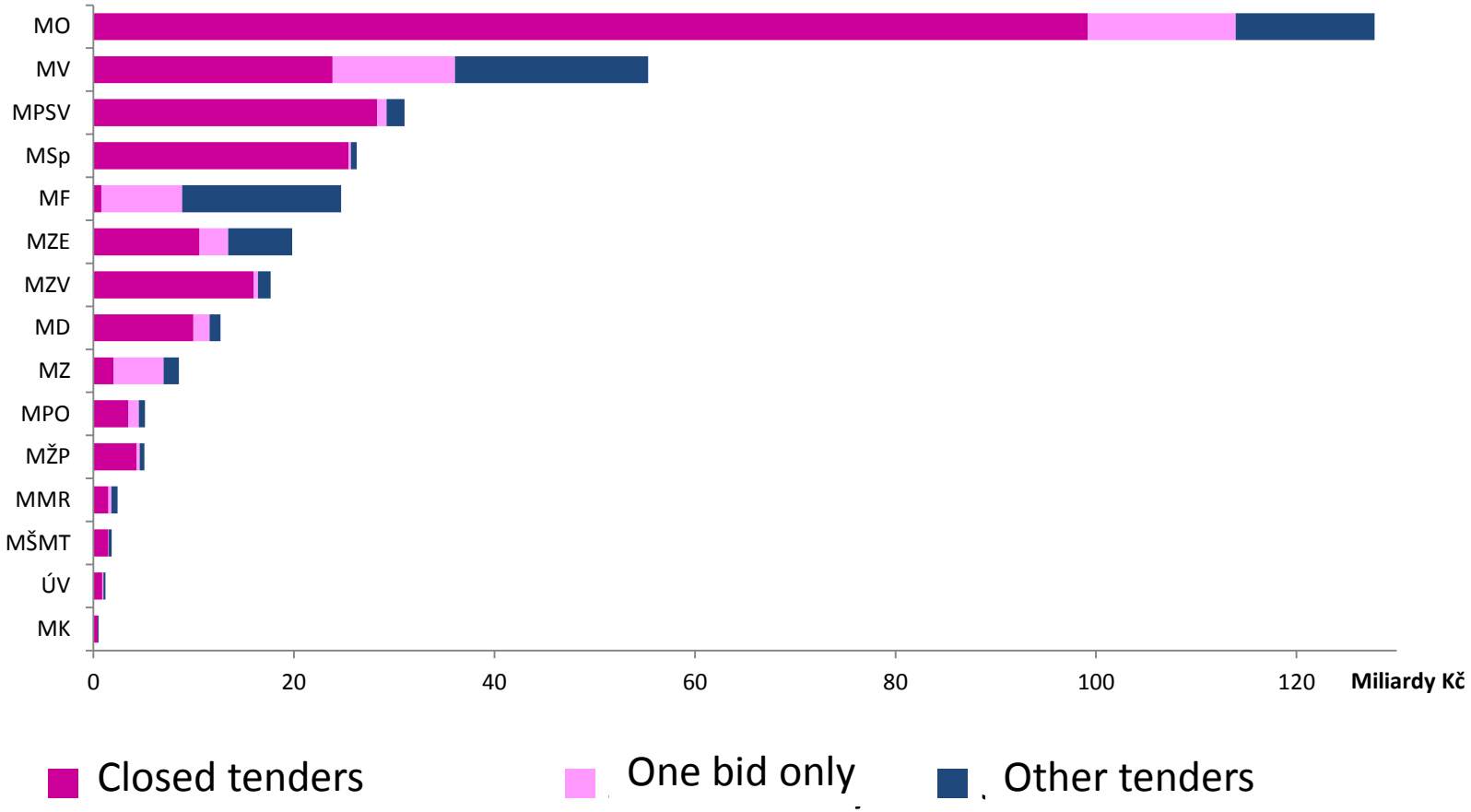
c) Because it increases the value of the database – NGO funding vs. Business

The story of zIndex.cz - results



Here was some kicking and screaming, including but not limited to threats of legal action – history can be found on www.zindex.cz (in Czech)

The story of zIndex.cz - results



The story of zIndex.cz - results

- Testing of outlyers
  - **IKEM – Kardioport**
    - **Company with CZK 50 000 turnover in 2009**
    - **In 2008 won a tender of IKEM for CZK 2 bn.**
    - **Bearer shares in paper form**
    - **Investigated by HN – attracted attention to by that time not very interesting segment of public healthcare (media more interested in constructions)**
    - **Managers of IKEM – Malý and Netuka losing their functions**

The story of zIndex.cz - results

- Commercial results
  - SaaS application for public tendering zInfo (currently under certification)
  - Selling of cleaned database
- Support of not-for profit projects by provision of reliable data
- Own not-for profit projects using the data (e.g. Bearer-shares in paper form – supported by NFPK, detection of real owners of public tender winners supported by NFPK, Motejl Fund and HN)

Other projects

- www.vasmajetek.cz – centralization of auctions on one webpage; not vulnerable to data (limited cleaning); commercial
- Data-mining in databases of CENIA – no data cleaning – since the purpose is merely general market analysis; commercial
- Public budgets – visualizations and data analysis – e.g. http://data.blog.ihned.cz/c1-58197600-za-co-hlavni-mesto-utrati-45-miliard-proklikejte-si-zjednoduseny-rozpocet
- Published uncleaned data with better search tools on www.vsechnyzakazky.cz

Data to go – no cleaned, but should be…

- Tender Electronic Daily (http://ted.europa.eu/TED/main/HomePage.do) – the jungle of all large EU tenders
- The Research and Development and Innovation Information System of the Czech Republic (http://www.isvav.cz/prepareResultForm.do) – RIV
- IS CEDR – Central Register of Subsidies (http://cedr.mfcr.cz/cedr3internetv415/default.aspx) and register of agricultural subsidies (http://eagri.cz/public/web/mze/farmar/registr-prijemcu-dotaci/)
- Many more….

zIndex tým:
Dotazy? Komentáře?

- Kontakt:
  - Jana Chvalkovská (jana.chvalkovska@zindex.cz)
  - Jiří Skuhrovec (jiri.skuhrovec@zindex.cz) – IT questions ☺
  - zIndex.cz