



národní
úložiště
šedé
literatury

How much do we need to clean?

Chvalková, Jana Gutierrez
2012

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

Dílo je chráněno podle autorského zákona č. 121/2000 Sb.

Licence Creative Commons Uveďte autora 3.0 Česko

Tento dokument byl stažen z Národního úložiště šedé literatury (NUŠL).

Datum stažení: 28.04.2024

Další dokumenty můžete najít prostřednictvím vyhledávacího rozhraní nusl.cz .

zIndex.cz

How much do we need to clean?

Jana Gutierrez Chvalková, 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

- ✘ 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

- 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 concerning highway tolls)
2. Data available online, but only as html
3. Screen scraping of the data

- 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 concerning 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



Historie Záložky Nástroje Nápoředa

http://www.isvsuz.cz/usisvz/usisvz01005Prepare.do?znackaForm=6000056502001

Ekonomické su... PMA loc-phpMyAdmin loc-Adminer datafix

Profile Extra Stuff

Search for PHP information:

Search 15°C Login E-mail Notifier YouTube [550] Music Games

http://localhos...lier_rating.php http://localho..._deadlines.php http://local

INFORMAČNÍ SYSTÉM O VEŘEJNÝCH ZAKÁZKÁCH - Uveřejněno

Nabídka veřejných zakázek

- Úvodní stránka
- Dodavatel
- Zadavatel
- Vyhledávání veřejné zakázky/koncesního řízení
- Přímé zadání veřejné zakázky / koncesního řízení
- Sledování veřejných zakázek
- Veřejné zakázky archiv
- Odkazy
- Legislativa
- ATEST IS VZ US

[verze pro tisk](#)

OZNÁMENÍ O ZAKÁZCE

Značka formuláře: 6000056502001
Evid. číslo v IS VZ US: 60000565
Limity: Nadlimitní
Datum zveřejnění: 15.08.2006
Zadavatel dle zákona č. 137/2006 Sb. nebo č. 139/2006 Sb.

ODDÍL I: VEŘEJNÝ ZADAVATEL

I.1) NÁZEV, ADRESA A KONTAKTNÍ ÚDAJE

Úřední název:	Obec Dynín
Poštovní adresa:	Dynín 48
Obec:	Dynín



MINISTERSTVO
PRO MÍSTNÍ
ROZVOJ ČR

VĚSTNÍK VEŘEJNÝCH ZAKÁZEK

Kontakty

[Přihlásit](#) [Zaregistrovat](#)

Úvodní stránka

Aktuality

Vyhledávání

Podle více parametrů

Podle data uveřejnění

Podle evidenčních čísel

Podle názvu zakázky

Podle zadavatele

Vítězové veřejných zakázek

Seznam profilů zadavatelů

Seznam zrušených profilů zadavatelů

Provést uveřejnění

Podání formuláře k uveřejnění

Číselníky a klasifikace používané při uveřejňování

XML rozhraní pro VVZ

Video postupy

Vyhledávání

Podání formuláře online

Podání formuláře offline

Odeslání XML

Správa formulářů

Správa uživatelů

Informace o systému

Nastavení elektronického podpisu

Nejčastější dotazy a odpovědi (FAQ)

Evidenční číslo zakázky: 222557

Evidenční číslo formuláře: [7202031022557](#)

Datum uveřejnění ve VVZ: 02.11.2012

Datum odeslání do TED: 02.11.2012

Typ: Opravný



Evropská unie

Vydání dodatku k Úřednímu věstníku Evropské unie

2, rue Mersier, 2985 Luxembourg, Luxembourg

E-mail: oj@publications.europa.eu

Fax: +352 29 29 42 670

Informace & online formuláře: <http://simso.europa.eu>

Oznámení o zakázce

směrnici 2004/18/ES

Oddíl I: Veřejný zadavatel

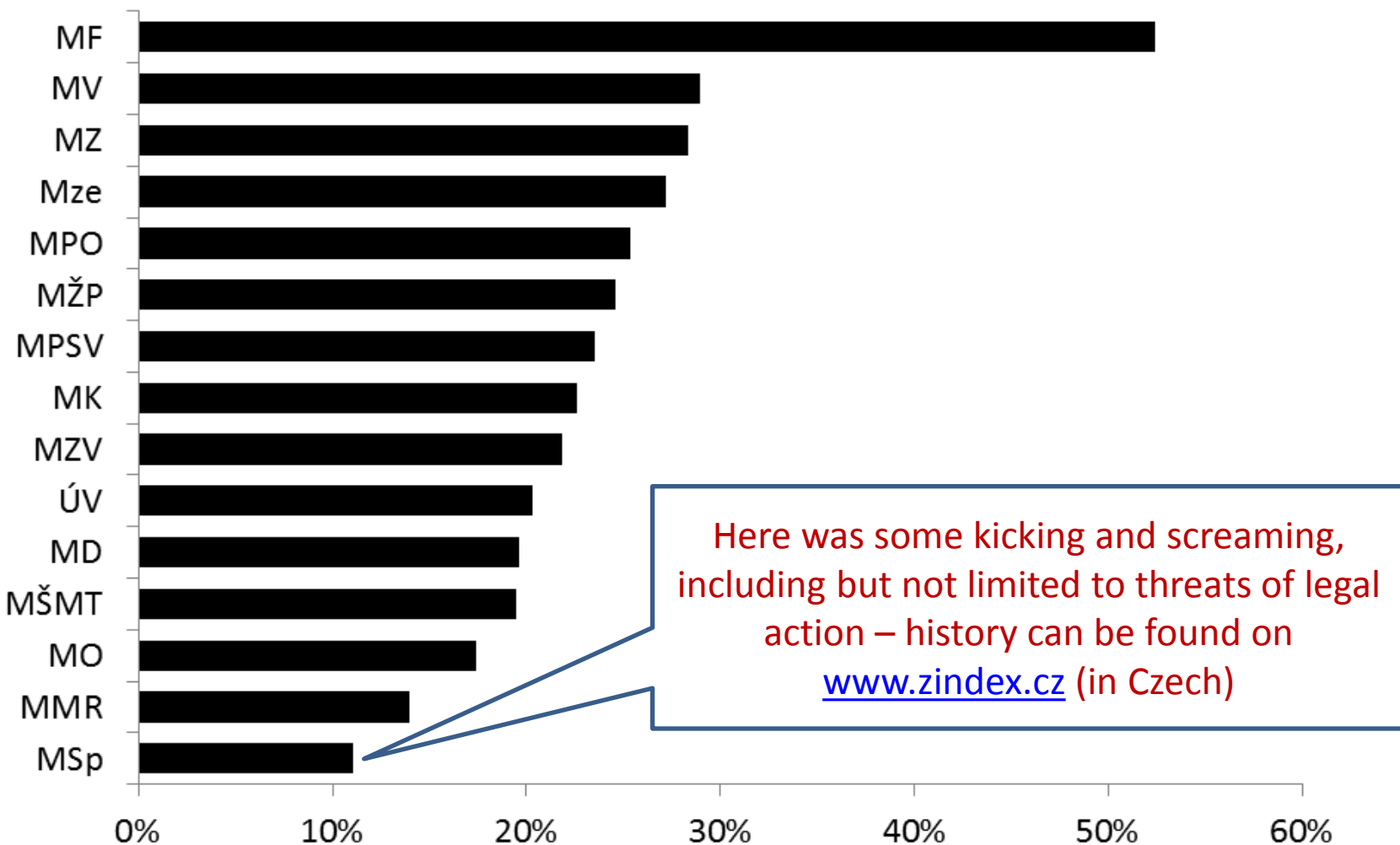
I.1) Název, adresa a kontaktní místo/místa

Úřední název	Identifikační číslo (je-li známo)
Česká republika – Ministerstvo zemědělství	00020478
Poštovní adresa	
Těšnov 17	
Obec	Praha 1
PSČ	117 05
Stát	CZ
Kontaktní místo	Tel.
K rukám	
E-mail	Fax
Internetové adresy (jsou-li k dispozici)	

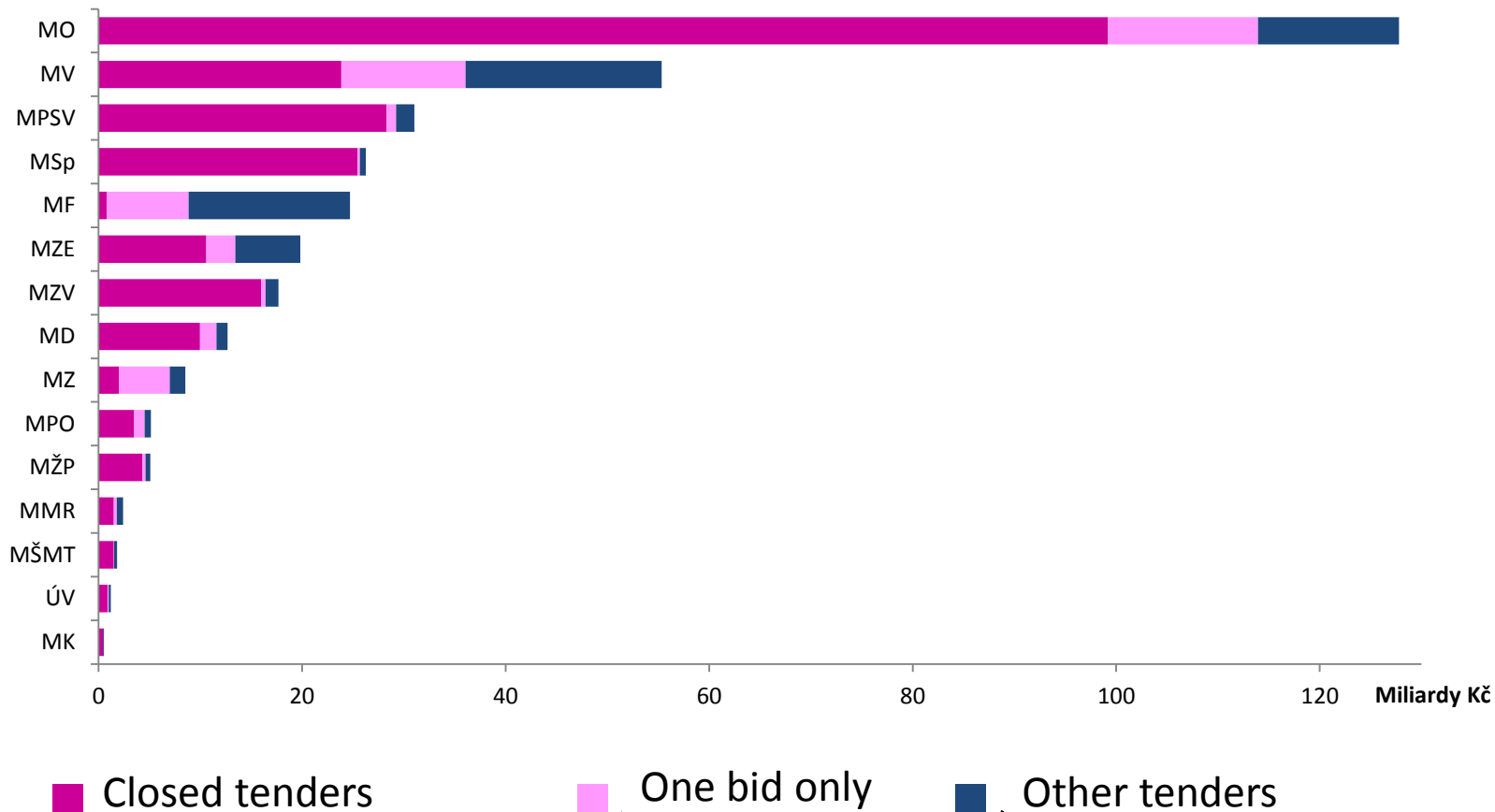
5. – 10. Data cleaning. Why?

- 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



The story of zIndex.cz - results



✘ Testing of outliers

✘ 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-zjednodusenym-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 Chvalková (jana.chvalkovska@zindex.cz)
- ✖ Jiří Skuhrovec (jiri.skuhrovec@zindex.cz) – IT questions 😊
- ✖ [zIndex.cz](http://zindex.cz)