Interpreting Open Data

Tomáš Kramár, @tkramar, http://minio.sk

Open Data

- **Open data** is the idea that certain data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control.

- Public money was used to fund the work and so it should be universally available.

-- Wikipedia

Web is full of Open Data

... it's just not the easiest kind of data to work with.

Open Data in Slovakia

- register of companies
- register of freelancers
- information from Statistical office
- procurements
- governmental contracts
- tax debts
- financial reports
- ...

Ideas?

Not so fast!

Barrier #1: Accessibility

```
<?xml version="1.0" encoding="utf-8"?>
<OznameniaSpravcov xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xsd="http
  <Upadca>
    <ObchodneMenoNazov>MRC s.r.o.</ObchodneMenoNazov>
    <Sidlo>
      <Ulica xmlns="">Železničná</Ulica>
      <Cislo xmlns="">324</Cislo>
      <Obec xmlns="">Kamenica nad Cirochou</Obec>
      <Psc xmlns="">06783</Psc>
      <Stat xmlns="" />
    </Sidlo>
    <Ico>36512966</Ico>
  </Upadca>
  <Spravca>
    <Typ Kod="PO">Právnická osoba</Typ>
    <ObchodneMenoNazov>Prvá arbitrážna k.s. </ObchodneMenoNazov>
    <Sidlo>
      <Ulica xmlns="">Prof. Sáru </Ulica>
      <Cislo xmlns="">5</Cislo>
      <Obec xmlns="">Banská Bystrica </Obec>
      <Psc xmlns="">97401</Psc>
      <Stat xmlns="" />
    </Sidlo>
    <Kontakt>prvaarbitrazna@zoznam.sk</Kontakt>
  </Spravca>
  <SpisovaZnackaSpravcovskehoSpisu>3K 41/2011 S 1429</SpisovaZnackaSpravcovskehoSpisu>
  <SpisovaZnackaSudnehoSpisu>3K 41/2011</SpisovaZnackaSudnehoSpisu>
  <DruhPodania Kod="15">Vylúčenie súpisovej zložky majetku zo súpisu</DruhPodania>
  <TextPodania>&lt;p&gt;Spr&amp;aacute;vca konkurznej podstaty &amp;uacute;padcu &lt;str
Cirochou, IČO: 36 512 966, na z&amp;aacute;klade predch&amp;aacute;dzaj&amp;uacute;ceho
```


You will rarely find structured data

Výpis z Obchodného registra Okresného sú

Tento výpis má len informatívny charakter a nie je pou

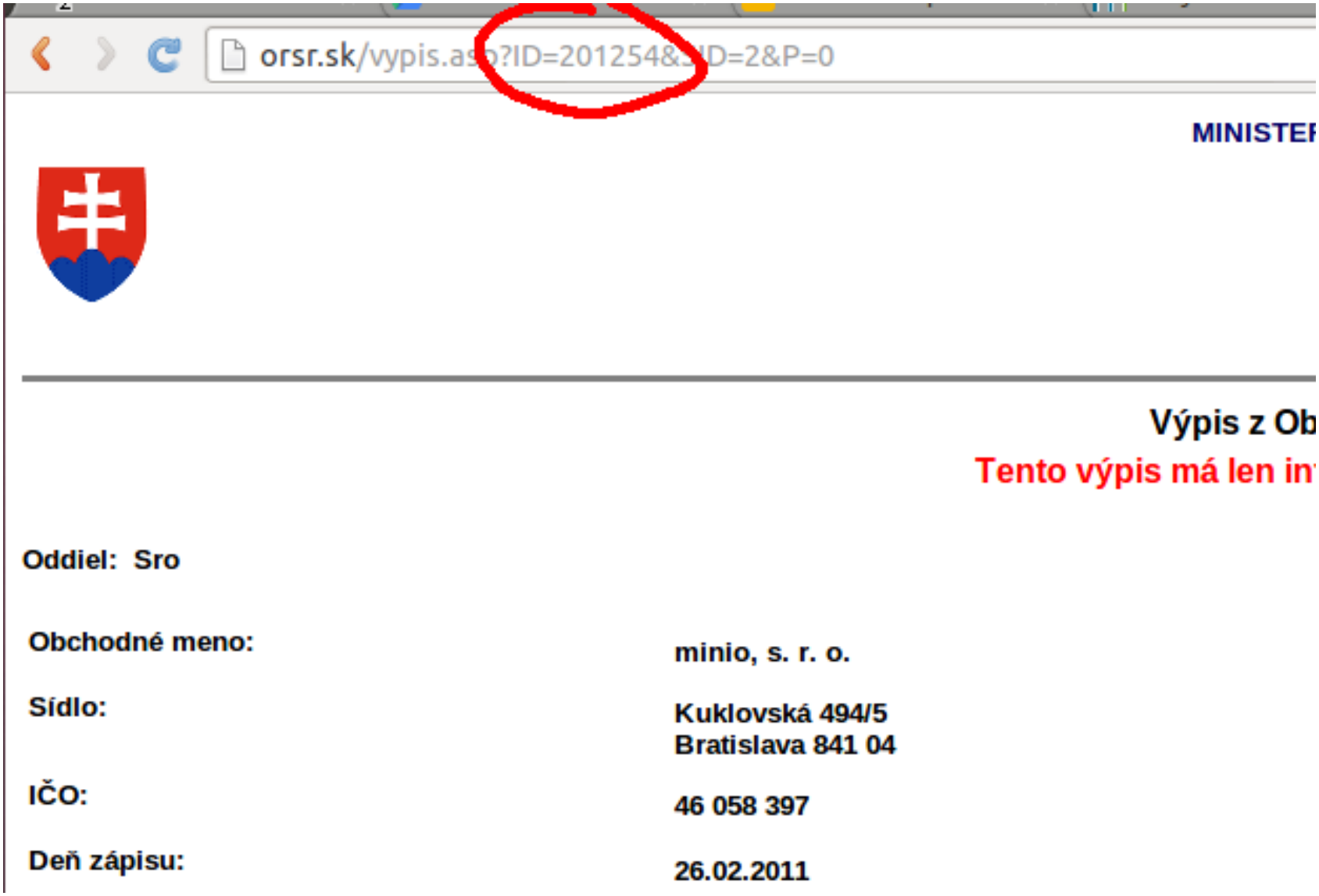
Oddiel: Sro

Obchodné meno:	**minio, s. r. o.**
Sídlo:	**Kuklovská 494/5** **Bratislava 841 04**
IČO:	**46 058 397**
Deň zápisu:	**26.02.2011**
Právna forma:	**Spoločnosť s ručením obmedzeným**
Predmet činnosti:	**počítačové služby** **služby súvisiace s počítačovým spracovaním údajov** **kúpa tovaru na účely jeho predaja konečnému spotrebiteľovi (maloobchod) alebo iným pr (veľkoobchod)** **reklamné a marketingové služby** **vydavateľská činnosť** **služby súvisiace s produkciou filmov alebo videozáznamov** **vykonávanie mimoškolskej vzdelávacej činnosti** **činnosť podnikateľských, organizačných a ekonomických poradcov** **sprostredkovateľská činnosť v oblasti služieb** **sprostredkovateľská činnosť v oblasti obchodu** **výroba komunikačných zariadení, spotrebnej elektroniky, počítačov a kancelárskych stroj**
Spoločníci:	**Michal Barla** **Kuklovská 494/5** **Bratislava 841 04** **Tomáš Kramár** **Voderadv 228**

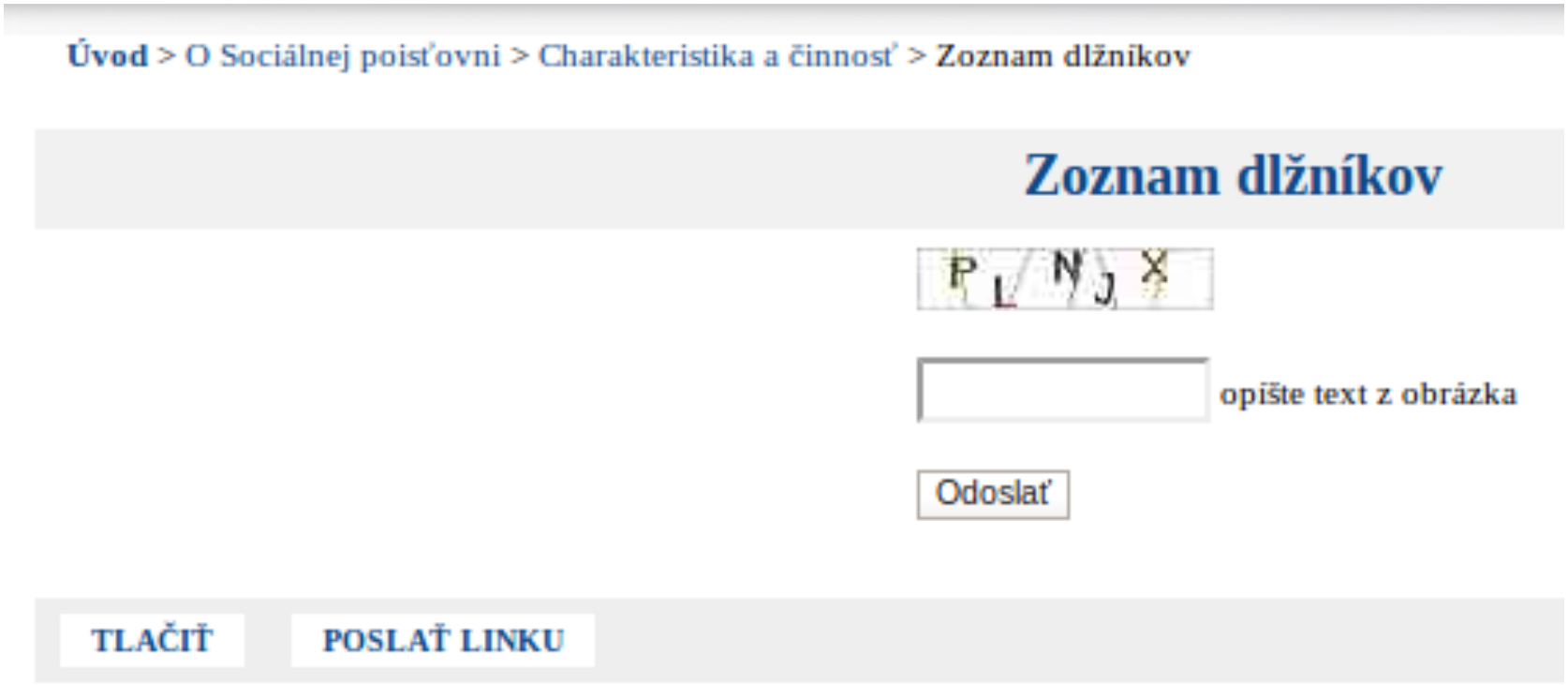
Parsing HTML is your best bet

Perils of Web scraping

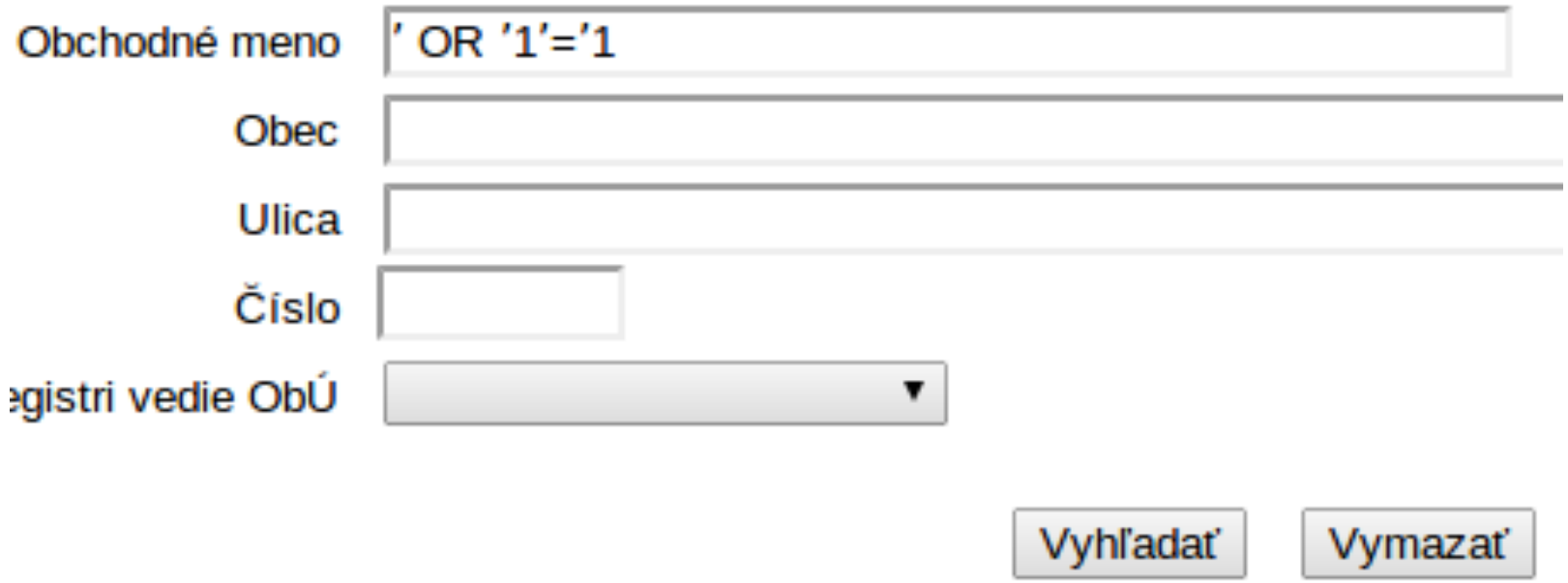
- Parsing HTML
  - Plenty of edge cases
  - Malformed HTML
- Crawling whole site
  - ideal case: list of all records with links
  - usually: deep Web, hidden behind a search form



Sometimes you can try different IDs



Sometimes you need to break a CAPTCHA



And sometimes you need to be more creative

Barrier #2: Data quality

- Noisy data
- Deliberately missing data
- **Linking data within and between datasets**

Notable OpenData projects

foaf.sk - Social network of slovak companies

- companies register, public procurements, debts, internet domains and other public data.
- 300K companies and 500K+ people
- current and historic data about companies in aggregated and usable form
- visualizes connections of people and companies as network (graphs) for deeper insights.
- 500K+ pageviews per month.

otvorenezmluvy.sk - analyzing contracts

- 300K+ scanned documents (over 1TB of raw data)
  - various formats
  - multiple sources
- fulltext and advanced faceted search
- automatic analysis of contract problematicity
- in-browser document viewer, visual annotations of any part of the contract
- embedded by largest slovak online newspaper (sme.sk)

govdata.sk - API for structured public data

- Scraping, cleaning and deduplication of many public and private data sources scattered through the web.
- Identifying and linking subjects from different datasets.
- REST-based API for searching 1.2M+ subjects in realtime.

otvorenesudy.sk - transparent court decisions

- data from Ministry of Justice
  - completely unusable site, search takes minutes to complete
- in a usable form
- fulltext search, statistics, visualisations

tender.sme.sk - public procurements

- OLAP tool built on top of Slovak procurements
- Slice the data

Pattern?

Interpret meaningless data.
Link it.
Find the (hidden) connections.

Questions?

How open is data in Austria?
What's your idea?

hello@minio.sk