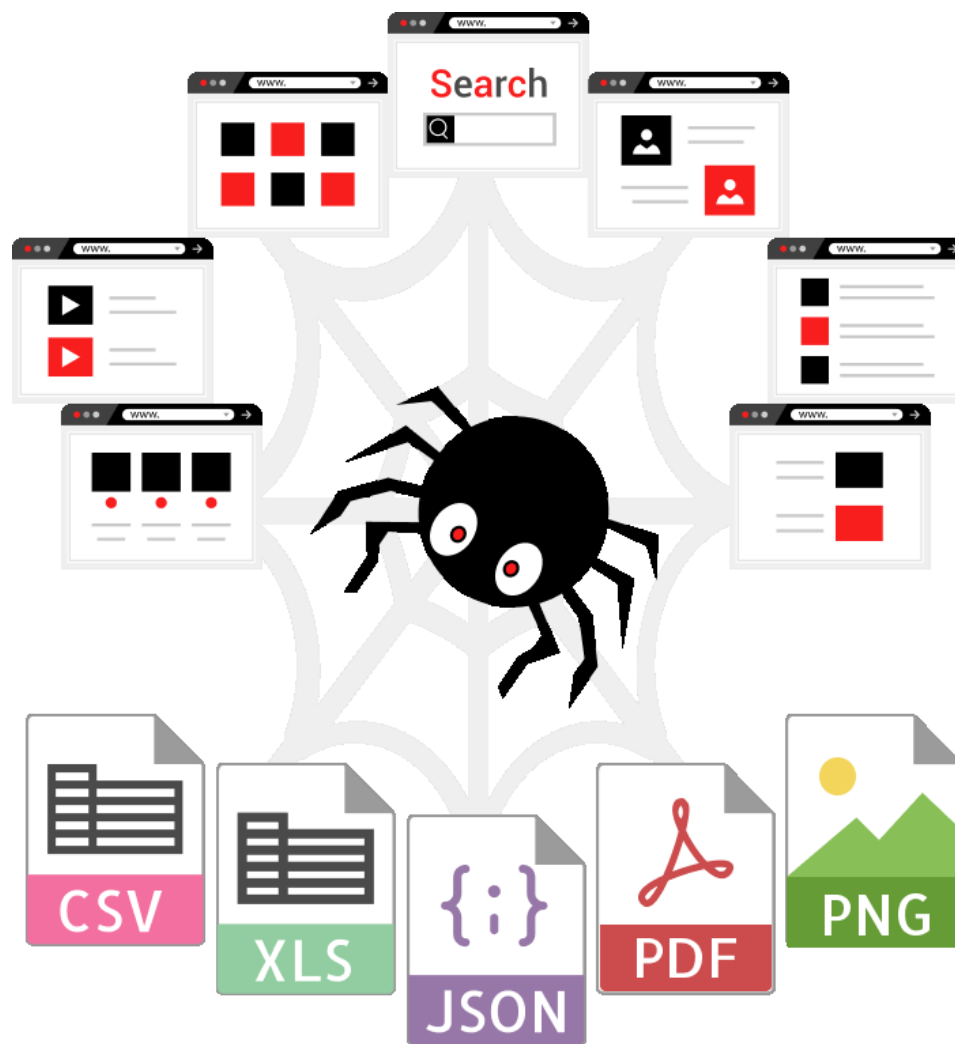


# Z7894 Geoinformační technologie v sociální geografii



4. cvičení  
21. 10. 2024

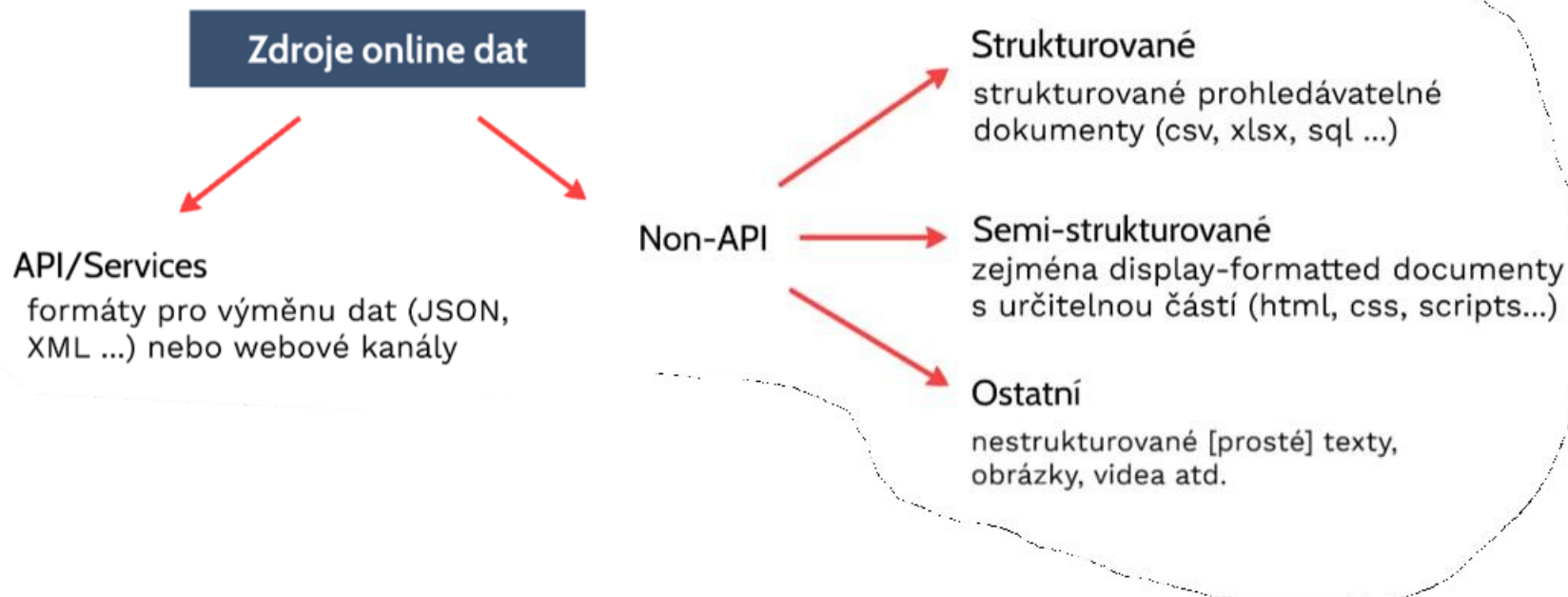
# Web scraping: teorie



# Web scraping

- **Získávání dat** z webových stránek (ale i webových aplikací a jejich součástí) a jejich **ukládání do strojově zpracovatelné podoby** (např. tabulky, samostatná databáze apod.).
- Související pojmy: **Extrakce dat** a **automatizace**
- Sporná je otázka (i)legality web scrapingu – někdy jde o **šedou zónu**
- Možné použití pro geography:
  - kopírování obsáhlých tabulek s listingem (např. data ČSÚ, ERU...),
  - ukládání webových seznamů do tabulky (např. seznam chystaných akcí v Brně...),
  - automatické načítání údajů (např. „excelovský sheet“ napojený aktuální teplotu),
  - periodické ukládání určitého údaje (např. návštěvnost bazénu),
  - ...

# Typy zdrojů online dat na webu ke scrapování



Web scraping se nejčastěji považuje za odchyťávání elementů html (resp. **parsování html**), ale může jít i o další formy spojené s **reverzním inženýrstvím** aplikací a extrakcí dat.

# Typy (časoprostorových) geografických dat

## Typ kolekce

historická

současná

předpověď

## Typ záznamu

agregovaný

jednoduchý

## Typ vyjádření

absolutní

relativní

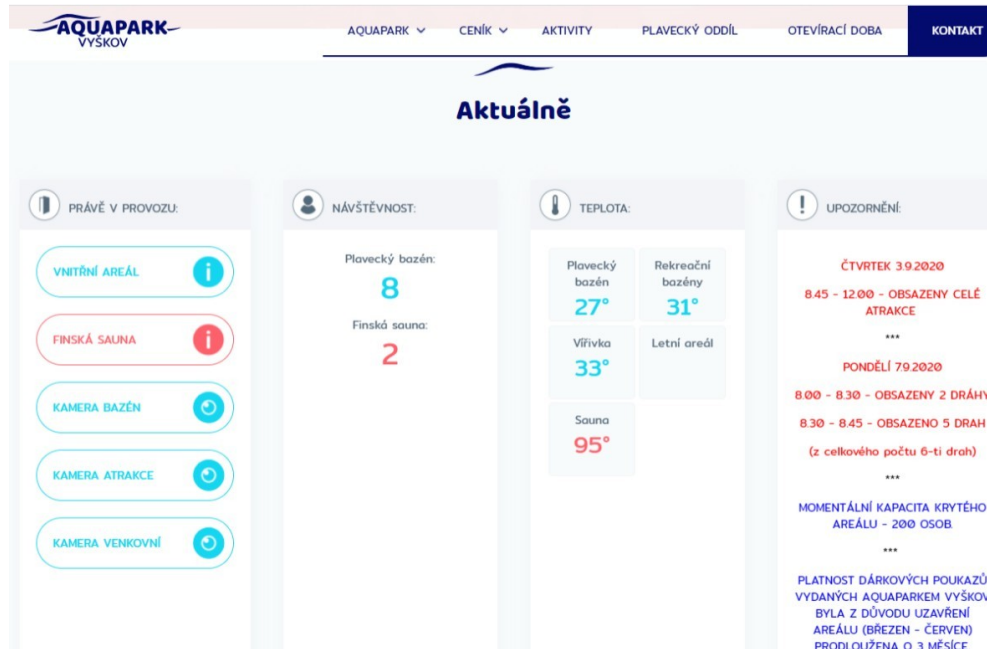
již zaznamenaná měření teploty  
vs  
aktuální teplota  
vs  
predikce budoucí teploty

aktuální obsazenost parkoviště  
vs  
google popular times

návštěvnost v procentech  
vs  
návštěvnost přesná

Některé scrapování je nutné **automatizovat** (např. pokud chceme sbírat údaj o návštěvnosti v čase a sestavit časovou řadu).

# Typy časoprostorových dat - ukázka



## HANGAR lezecké centrum Brno



Web Trasa Uložit Volat

4,9 ★★★★★ 323 recenzí Google

Tělocvična s horolezeckou stěnou v Brně

Adresa: Pražákova 1027/53, 619 00 Brno-střed

Navštívili jste v den: Středa

Otevírací doba: Otevřeno · Zavírá: 22

Telefon: 608 987 910

Navrhněte úpravu · Vlastněte tuto firmu?

Znáte toto místo? Podělte se o nejnovější informace

### Otázky a odpovědi

Zobrazit všechny otázky (2)

Zeptejte se

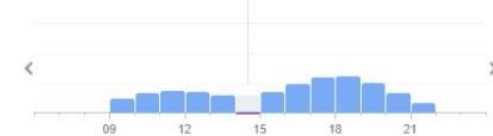
### Recenze z webu

100 % Firmy.cz · 2 hlasy

### Oblíbené časy

pátky ↕

Živě: Nizké vytížení



### Plánování návštěvy

Lidé zde obvykle stráví 1,5–3 h

## Základní rozdělení

- **Jednoduché s GUI** – např. Web Scraper plugin pro webový prohlížeč
- **Složitější pro programovací jazyky** – např. knihovny pro python/R (např. rvest)
- **Vlastní skripty/programy**
- **Komerční cloudové platformy** – např. Apify, Octoparse, Scrapestack

**Jaké jsou výhody a nevýhody jednotlivých řešení?**

# Projekt č. 2 – Příprava na zadání Prostudování nového tématu

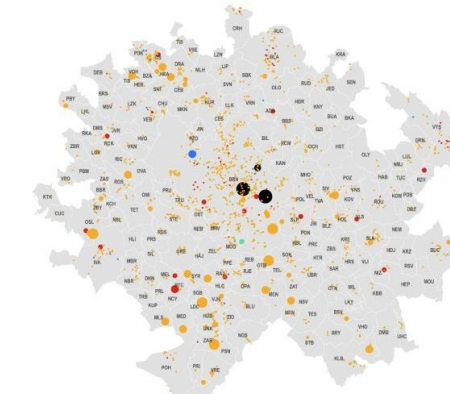
**Úkol na cvičení:** Prostudujte následující datové analýzy (publikace) týkající se energetiky a zamyslete se nad možnými tématy a zpracováním

## 7 MODERNÍ A BEZPEČNÁ ENERGETIKA

### 7.1 Energetický mix BMO

Diskuse o vhodném energetickém mixu roste na významu především se zájmem snižovat využívání fosilních paliv a produkci emisí skleníkových plynů. Z hlediska hodnocení zvoleného energetického mixu je důležité brát ohled nejenom na zvyšování podílu obnovitelných zdrojů energie (OZE) a diverzifikaci zdrojů, ale také na jejich celkovou vyváženost a flexibilitu. Snahou je snižování energetické závislosti a výstavba vhodných lokálních zdrojů, které maximalizují využívání přírodních podmínek spolu s bezpečným a nepřetržitým zajištěním dodávek elektrické energie.

07-1 INSTALOVANÉ ZDROJE ELEKTRICKÉ ENERGIE NA ÚZEMÍ BMO (únor 2020)



● sluneční ● vodní ● pami ● paroplynové ● plynové a spalovací ● větrné ● kogenerační  
● do 0,50 MW ● 0,51-2,00 MW ● 2,01-5,00 MW ● 5,01-10,00 MW ● nad 10,00 MW



[https://www.irozhlas.cz/zpravy-domov/fotovoltaika-energetika-obnovitelne-zdroje\\_1912040600\\_jab](https://www.irozhlas.cz/zpravy-domov/fotovoltaika-energetika-obnovitelne-zdroje_1912040600_jab)

[https://is.muni.cz/auth/el/sci/jaro2023/Z7894/um/cviceni/cv\\_01/inspirace/Analyticka-ychodiska-ISR-BMO-21-1.pdf\(kapitola 7\)](https://is.muni.cz/auth/el/sci/jaro2023/Z7894/um/cviceni/cv_01/inspirace/Analyticka-ychodiska-ISR-BMO-21-1.pdf(kapitola 7))

data.brno TĚMATA FORMÁT PŘÍSPĚVKŮ DATASETY BRNO V ČÍSLECH O WEBU

### Rozvoj fotovoltaiky v Brně

Martin Dvořák | 31.01.2023

Využití obnovitelných zdrojů je trendem, který má charakter energetického průmyslu. Místo náročná velkých elektřin, které léta zajišťují stabilní dodávku energie, se začínají objevovat desítky menších výroben, získávajících energii z obnovitelných zdrojů. Ty jsou geograficky roztroušené a jejich dodávky jsou značně závislé na počasí. Je to výhoda z hlediska funkčnosti a stability celé rozvodné sítě. V tomto článku se zaměříme primárně na fotovoltaické elektrárny (FVE), jejich prostorovou distribuci v Brně a úskali, která nás provázejí při získávání a zpracování dat.

#### DATA DATA DATA

Na Magistrátu města Brna vzniklo od 1/12/2023 nové oddělení energetiky, jež bude vykonávat energetický management na území města. Ke správnému řízení jsou vždy zapotřebí data. Obrátit jsme se na Energetický regulační úřad (ERÚ), což je ústřední orgán státní správy ČR pro oblast regulace energetiky.

Premisa: K výrobě elektřiny je zapotřebí získat licenci od Energetického regulačního úřadu.

Tyto licence jsou k dispozici na [webu ERÚ ve veřejném listině](https://www.eru.cz/aktuality/licence). Zde je možné vyhledat všechny právníky subjekty disponující licencí k výrobě, a to za celou ČR. Ještě jsme chtěli získat všechny subjekty (tedy i fyzické osoby), bylo nutné se obrátit na ERÚ se žádostí podle zákona 106/1999 S., o svobodném přístupu k informacím. Zde je na místě otázka, proč tato data nepublikuje ERÚ (anonymizované) jako otevřená data v [národním katalogu otevřených dat](https://www.eru.cz/aktuality/licence). Dostali jsme odpověď, že na tom již pracují. Držíme palce.

Data jsme v každém případě dostali obratem a vrhli se na jejich zpracování. Pojďme se podívat na některé chyby, které se v dodaném exportu dat vyskytly:

- použít oddělovač – ... data jsou ve formátu Královo Pole-612484-Brno-3426/3 (standardně se používá čárka či středník)
- cca 15 způsobů zápisu parcely (a to pouze na 600 řádcích týkajících se Brna), ilustrativně:

par. č. 3426, parc. č. 3426, p.č. 3426, St. 3426, Parc.č. 3426, par.č.3426, parc.č. 3426, č.p.3426, Pt. 3426, p. č. 3426, par.č.st.

- pokud FVE leží na více parcelách, chyby se kumulují, což je ještě horší.

Zajímavé evidenci parcel bychom doporučili **sjednotit**, velmi to usnadní další zpracování.

Výsledek snad však stáří za námahy.



<https://data.brno.cz/pages/rozvoj-fotovoltaiky-v-brne>



## Projekt č. 2

- Vytvořte analýzu **energetických zařízení a jejich energetických údajů** vybraný kraj.
- Využijte tabulku udělených licencí pro provoz elektrárny v ČR (<https://licence.eru.cz/index.php>)
- Součástí odevzdaného dokumentu budou části zabývající se **explorací** (tj. průzkumem dat), **analýzou, syntézou dat** a (zejména kartografickou) **vizualizací** výsledků a **vývojový diagram** (postup).
- Aspoň **dva výstupy** budou mapové
- Využít lze všech relevantních datových zdrojů i softwarových nástrojů.

# Zadání cvičení

## Výsledky

- Odevzdaný dokument (protokol) se všemi náležitostmi (viz. formální splnění).
- Kladen důraz na **průzkum a analýzu dat, vizualizaci výsledků**.
- Lze se zaměřit na užší téma (např. udržitelná energie, energetická soběstačnost...)
- Data je vhodné doplnit např.: další data ERU (např. tabulky spotřeby/výroby), energetický potenciál území (solární, informace o vodních tocích a o převládajícím větru...), hustota zalidnění, blízkost velkých odběratelů elektrické energie, klima, OSM...

**Prezence: 18.11.2024 v čase přednášky**

**Deadline: 24.11.2024**

# Zadání cvičení

## Optimální postup

1. web scraping z ERU ([tabulka udělených licencí pro provoz elektrárny v ČR](#))
2. normalizace dat (sjednocení č. parcely)
3. propojování přes data RÚIAN (přes č. parcely)
4. explorace, analýza, syntéza dat a vizualizace
5. Dál už to znáte 😊...

# Zadání cvičení

## Data ERU

<https://licence.eru.cz/index.php>

<https://www.eru.cz/zpristupnena-data>

# Zadání cvičení

## Nastavení filtrů

**Filtr subjektů**

Číslo licence:

Název subjektu:  IČO:  Obec:

Ulice:  Kraj:  Okres:

**Filtr TEZ**

Ulice:   Okres:

Obec:  Katastrální území:   Kód  Název

Celkový výkon elektrický od:  do:  [MW]

Celkový výkon tepelný od:  do:  [MW]

Pro projekt nás budou zajímat výrobci elektřiny (předmět podnikání: výroba elektřiny) s udělenou licencí od ERU (stav žádosti: udělená licence).

Filtr TEZ zapnout pro vybraný kraj.

# Zadání cvičení

## Praktická ukázka



Web Scraper - Free Web Scraping

★★★★★ 783 ⓘ | Produktivita | Uživatelé: 600 000+

Web Scraper s GUI  
ve vývojářském  
režimu Chrome

<https://chrome.google.com/webstore/detail/web-scraper-free-web-scr/jnhgnonknehpejjnehehllkliplmbmhn>

Vytvoření schémat, resp. jejich import a zahájení scrapování.

# Zadání cvičení

## Úkol do příště

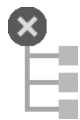
Všechny skupiny budou mít nascrapovaná data a pokud možno normalizované číslo parcely.

Datová struktura dat o parcelách z RÚIAN (atributy kmenovecíslo a pododdelenicíslo)

ogc_fid	gml_id	id	nespravny	parcely	usobyvyuzitipozem	druhcislovanikod	druhpozemkukod	atastralniuzemikod	platiod	platido	idtransakce
1	PA.72779565010	72779565010	NULL	205	27	2	14	600041	2019-06-27T00:...	NULL	2922468
2	PA.32933771010	32933771010	NULL	301	26	2	14	600041	2013-12-19T00:...	NULL	439084
3	PA.32412513010	32412513010	NULL	10	26	2	14	600041	2023-02-27T00:...	NULL	4819824
4	PA.65561479010	65561479010	NULL	145	NULL	1	13	600041	2017-12-15T00:...	NULL	2199303
5	PA.1296794701	1296794701	NULL	295	NULL	1	13	600041	2012-04-03T00:...	NULL	0
6	PA.1296795701	1296795701	NULL	235	NULL	1	13	600041	2012-04-03T00:...	NULL	0
7	PA.1501049701	1501049701	NULL	150	NULL	1	13	600041	2012-05-04T00:...	NULL	0
8	PA.1298281701	1298281701	NULL	957	NULL	2	5	600041	2012-05-04T00:...	NULL	0
9	PA.1297703701	1297703701	NULL	2	NULL	1	13	600041	2012-05-04T00:...	NULL	0
10	PA.18681985010	18681985010	NULL	5	26	2	14	600041	2012-05-04T00:...	NULL	0
11	PA.18681984010	18681984010	NULL	2	NULL	2	5	600041	2012-05-04T00:...	NULL	0

Dr. Herman se bude příští hodinu věnovat propojením dat z ERU s daty z RÚIAN.

# Dotazy?



Děkuji za pozornost