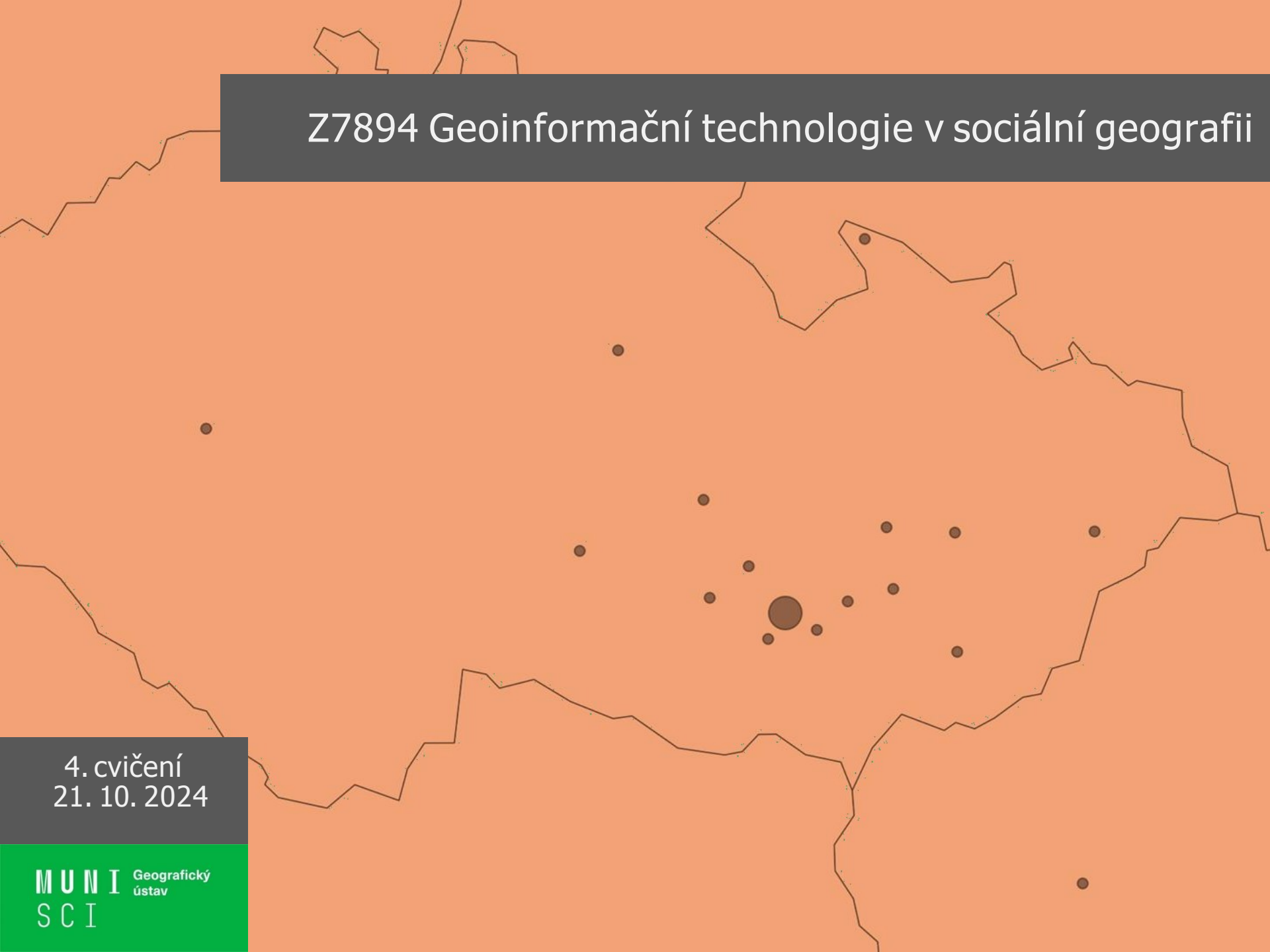
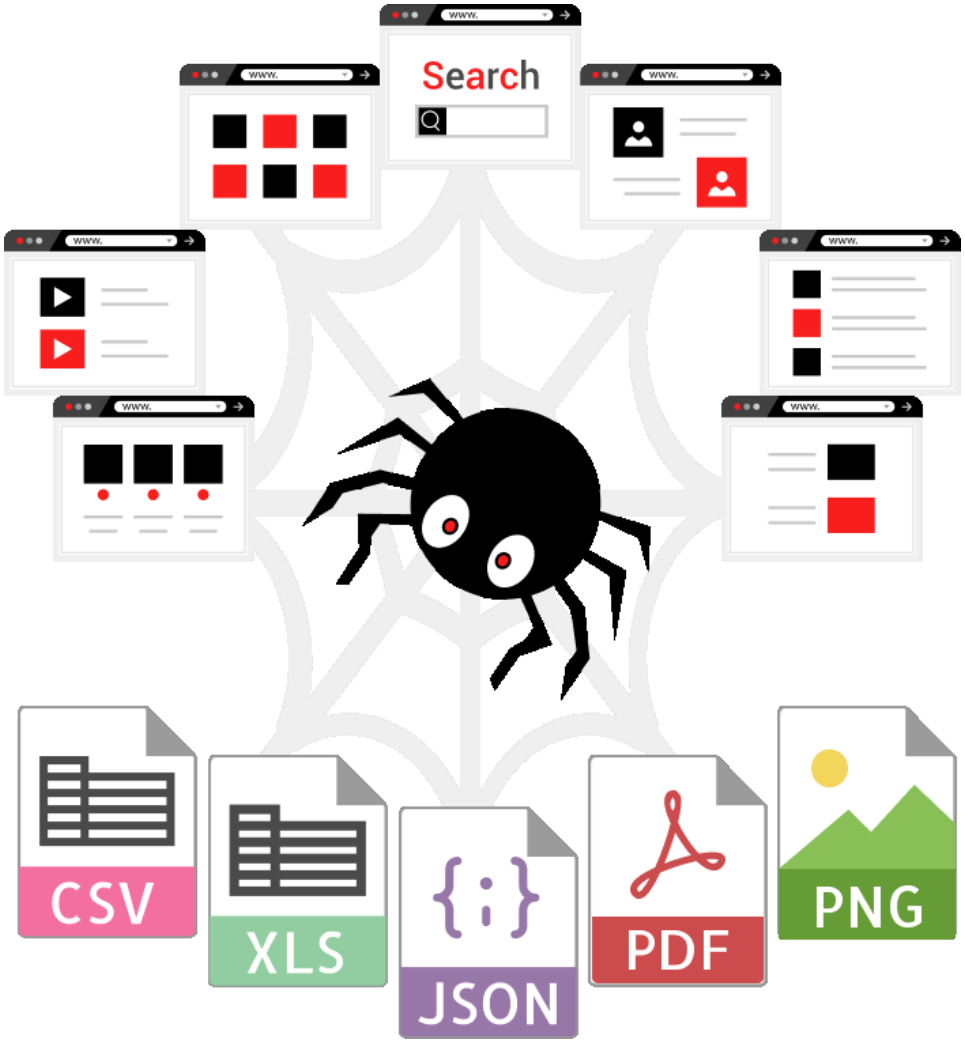
Z7894 Geoinformační technologie v sociální geografii



4. cvičení
21. 10. 2024

MUNI SCI Geografický ústav

Web scraping: teorie

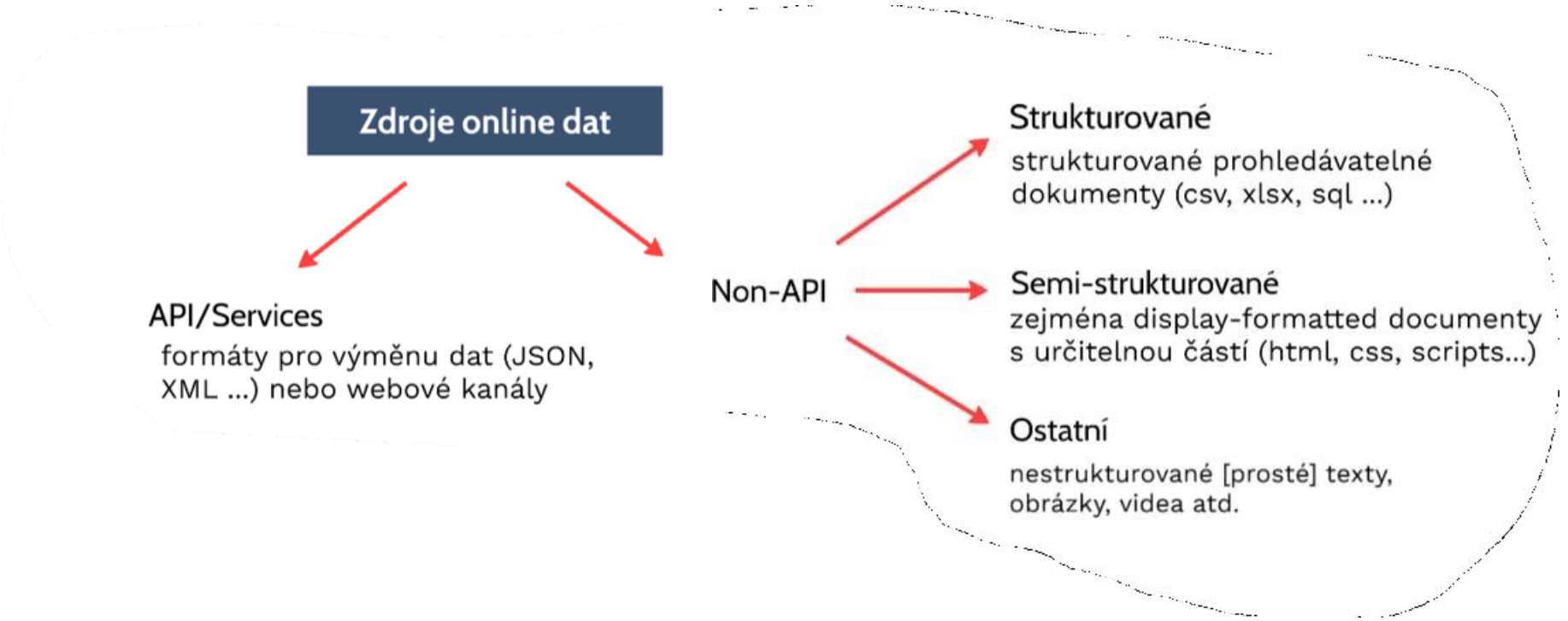


Web scraping

- **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**

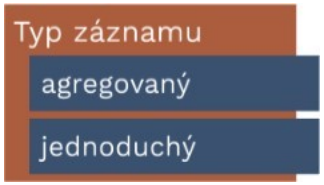
- Možné použití pro geografy:
  - kopírování obsáhlých tabulek s listingem (např. data ČSÚ, ERU…),
  - ukládání webových seznamů do tabulky (např. seznam chystaných akcí v Brně…),
  - automatické načítání údajů (např. „excelovský sheet" napojený aktuální teplotu),
  - periodické ukládání určitého údaje (např. návštěvnost bazénu),
  - …

Typy zdrojů online dat na webu ke scrapování



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

Typy (časoprostorových) geografických dat



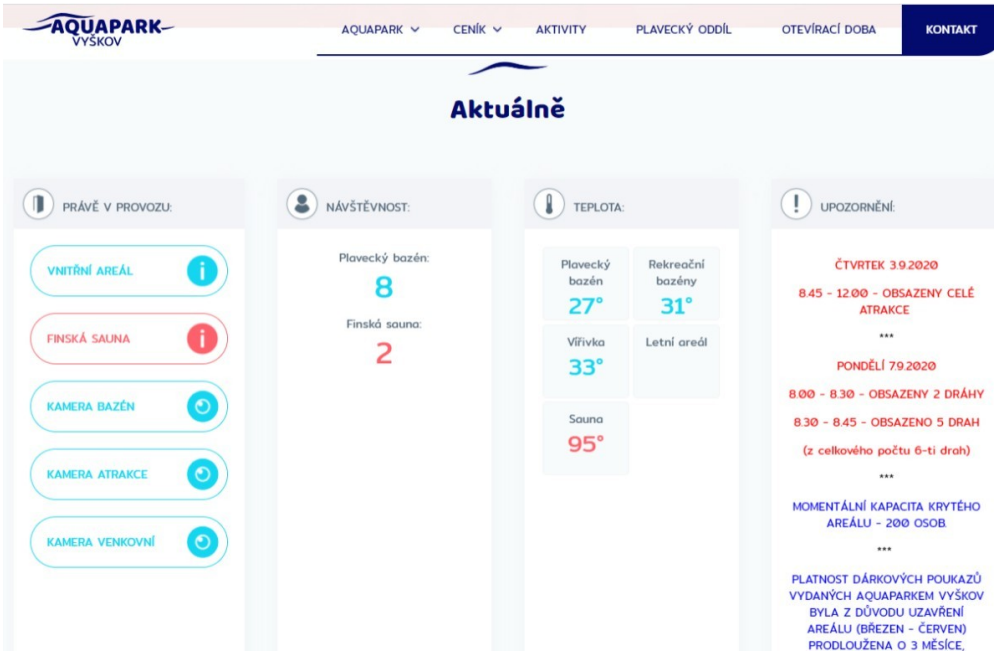
již zaznamenaná měření teploty
vs
aktuální teplota
vs
predikce budoucí teploty

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



Web scraping: Nástroje

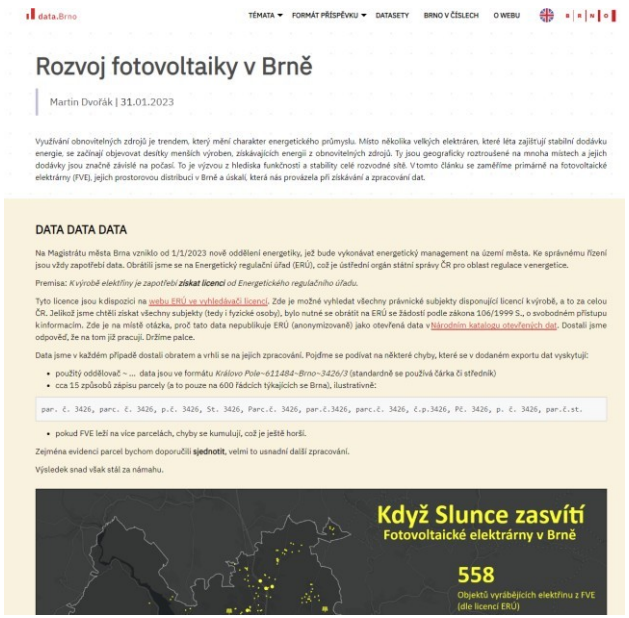
Základní rozdělení

- **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

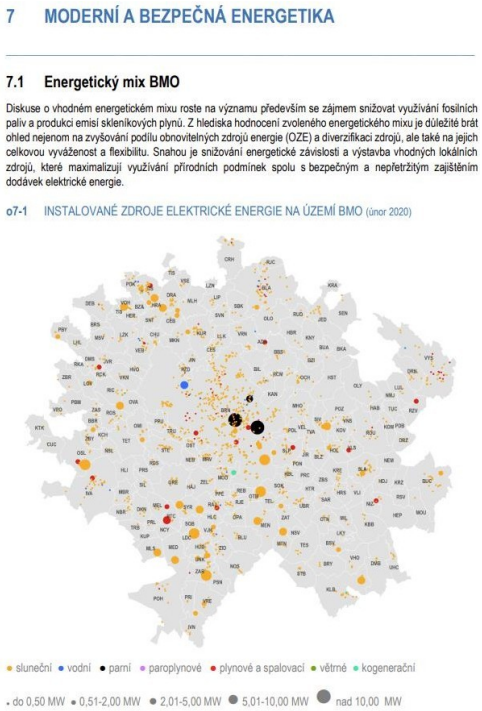
Jaké jsou výhody a nevýhody jednotlivých řešení?

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



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



https://is.muni.cz/auth/el/sci/jaro2023/Z7894/um/cviceni/cv_01/inspirace/Analyticka-vychodiska-ISR-BMO-21-1.pdf (kapitola 7)

https://www.irozhlas.cz/zpravy-domov/fotovoltaika-energetika-obnovitelne-zdroje_1912040600_jab

Zadání cvičení: Projekt

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…

Prezetace: 18.11.2024 v čase přednášky

Deadline: 24.11.2024

Zadání cvičení
Optimální postup

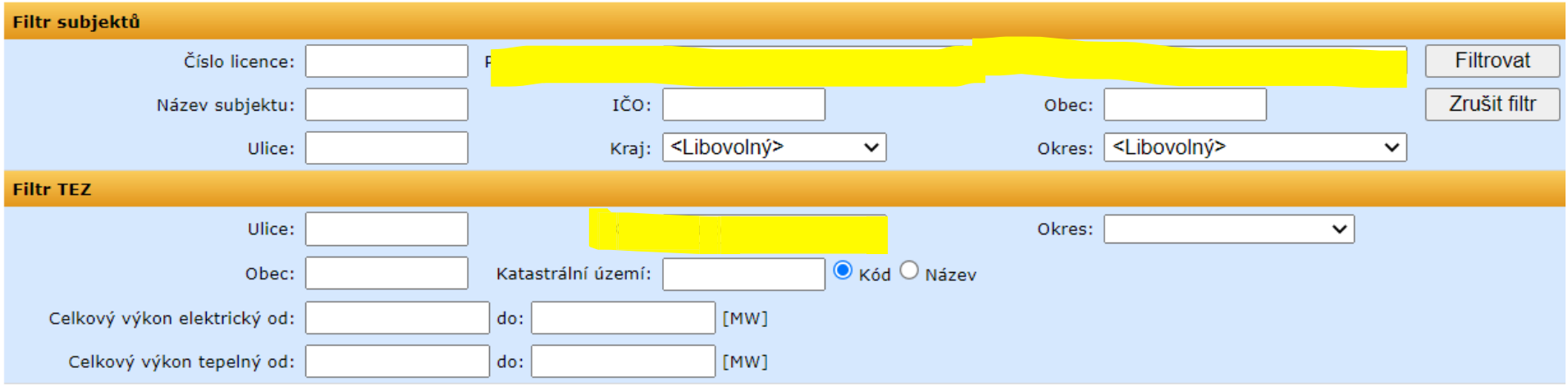
1. web scraping z ERU (tabulka udělených licencí pro provoz elektrárny v ČR)
2. normalizace dat (sjednocení č. parcely)
3. propojování přes data RÚIAN (přes č. parcely)
4. explorace, analýza, syntéza dat a vizualizace
5. Dál už to znáte ☺…

Zadání cvičení
Data ERU

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

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

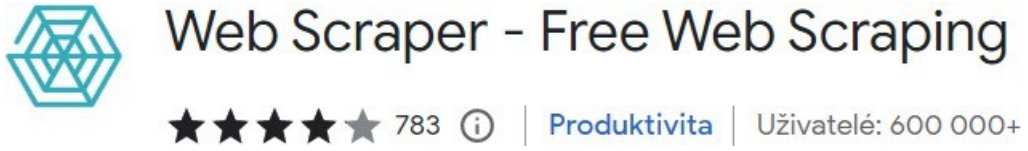
Zadání cvičení
Nastavení filtrů



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 s GUI ve vývojářském režimu Chrome

https://chrome.google.com/webstore/detail/web-scraper-free-web-scra/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 kmenovecislo a pododdelenicisla)

ogc_fid	gml_id	id	nespravny	arcely	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