



# „Tömeges” import tapasztalatok és a Kont-tréner

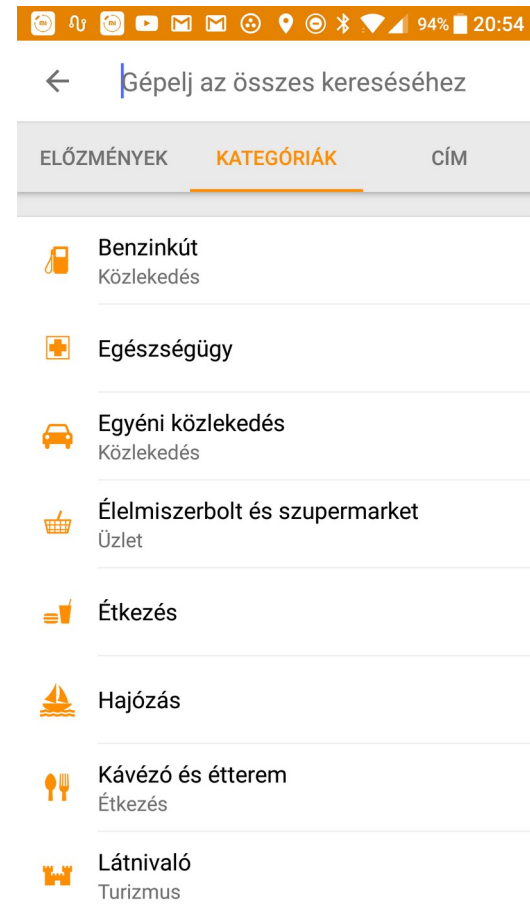
#POI #webscraping #adatok #Python

OpenStreetMap Meetup  
2019. szeptember 16.

Szala Kálmán  
@kami911kami

# A POI-k fontosak

- Sokszor van rá szükség
- Keresés
- Meglévő POI-k
  - Sokszor kevés adat
    - Nyitvatartás
  - Nincs fent minden POI
  - Nem egységesek a nevek
  - Elavulhatnak, elavulnak



# Ezeket terveztem



- Új adatok gyűjtése (POI)
  - Boltok, áruházláncok
  - Bankok
  - Éttermek, és étteremláncok
- Meglévő POI-k egységesítése
  - Azonos írásmód
  - Azonos kategória
  - OSM Wiki oldal, közös címkék → dolgozzunk rajta együtt!
    - Boltok gyűjteménye
    - Éttermek gyűjteménye
    - Töltőállomások gyűjteménye





**POI aratás**

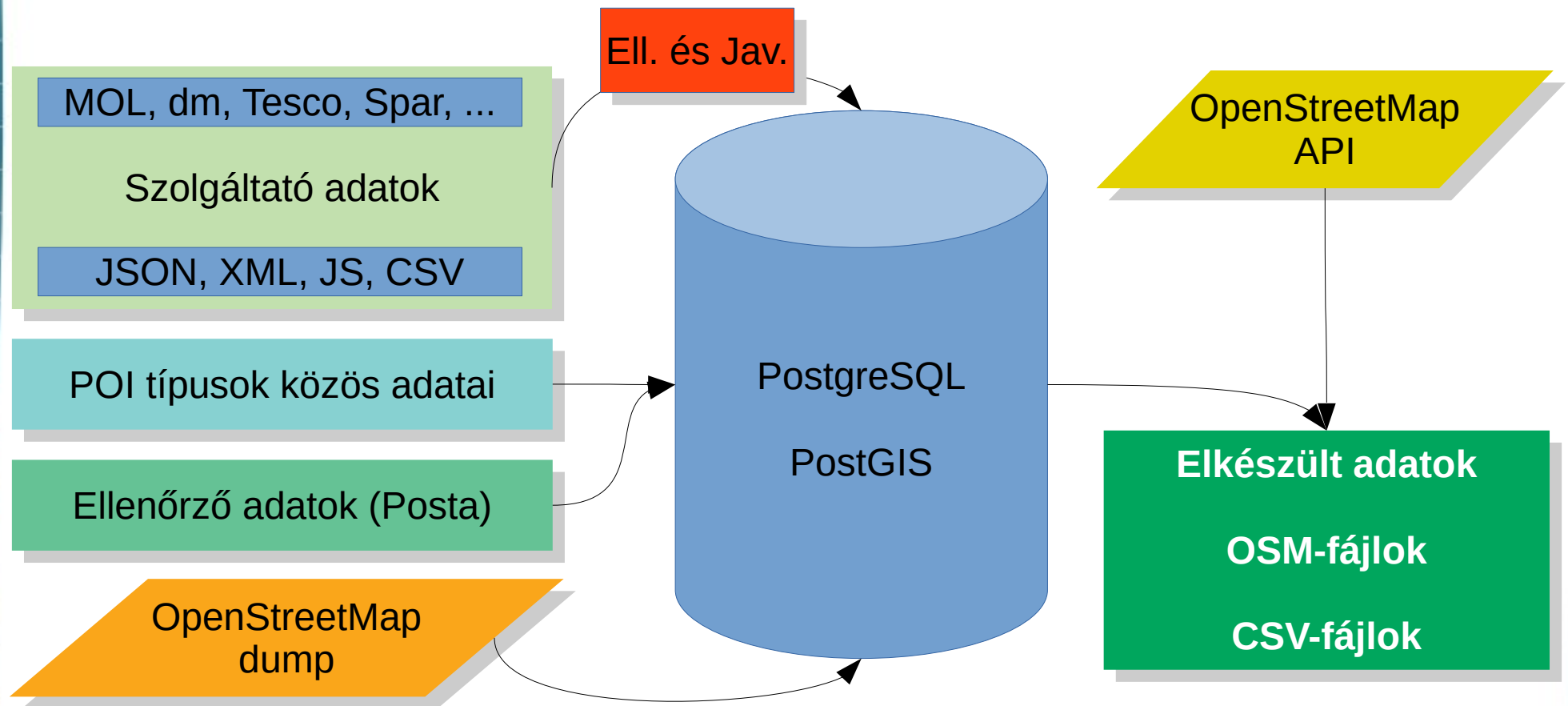


# Szolgáltatói adatok



- Benzinkutak
  - Avia, Shell, OMV, MOL
- Gyógyszertárak:
  - Benu, Kulcs
- Posták, Posta Pontok, csomagautomaták, csekkautomaták
- Boltok
  - CBA, Príma, **dm**, Rossmann, **Spar**, **Tesco**, **S-Market**
- Bankok és ATM-ek
  - K&H Bank és ATM, **Budapest Bank** és ATM
- Egyebek
  - Foxpost, MOL Bubi

# POI kereső



# Hibalehetőségek: OSM



- Már megszűnt POI
  - Példa: DM – Nyugati tér
    - <https://www.openstreetmap.org/node/2694483503>
- Rossz címkézés
  - Például shop címke eltér a várttól
    - shop=chemist helyett shop=cosmetics van
    - Shop exclusively selling cosmetics.
    - For shops selling also articles of personal hygiene and household cleaning products use shop=chemist.
    - <https://www.openstreetmap.org/node/3330070357/history>
- Nem jó a POI helye

# Hibalehetőségek: külső adatok

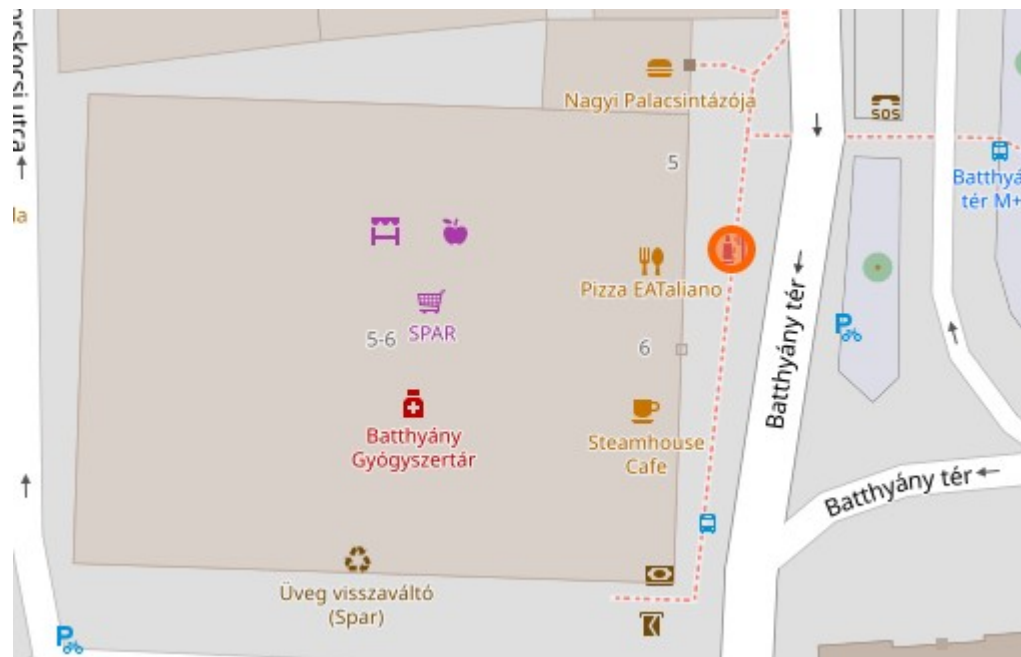


- Nem jó a POI címe
  - Más az közterület neve:
    - Teljes név van, vagy éppen hiányzik egy-egy rész a névből
    - Elgépelések
    - Az utca és az út szinonimaként történő használata
  - Eltér a házszám
- Nem jó a POI helye
  - Valami hiba miatt máshol van
    - Koordináta hiba, más geokódoló használata
  - Ha jó is, akkor is az út mellett van, nem az épületben
    - <https://www.openstreetmap.org/node/6752849469>



# DM az járdán

- DM, Batthyány tér




# Hibakeresés és javítás



- Életciklus menedzsment
  - Jelenleg nincs lehetőség a POI-k félautomatikus törlésére
  - Talán jobb ez kézzel csinálni → Ellenőrzés
- Samu Imre geojson scriptje
- Egyszerű megoldás – Overpass Turbo-ban:
  - dm and "source:www.dm.hu:date" != "2019-08-27" in Hungary
- <https://overpass-turbo.eu/s/Mk4>

# Megszűnt vagy átalakult POI-k



2016. augusztus 1-én az Eni Hungaria Zrt.-t - a kenőanyag üzletág kivételével - megvásárolta a MOL Nyrt., így többé nem része az Eni csoportnak.

- 80 darab (2018. június)
  - <https://taginfo.openstreetmap.hu/tags/brand=Agip>
- 50 darab (2019. szeptember)
  - <https://taginfo.openstreetmap.hu/tags/brand=Agip>





**POI-k felkutatása**



# POI-k felkutatása régen



- Koordináta alapján (hibalehetőségek)
- Azonos kategória (hibalehetőségek)
- Eredetileg 70 méteres körzetben
  - Konfigurációs fájlból konfigurálható
- Node felvétele, ha még nincs ilyen POI
- Ha létezik, akkor a már meglévő frissítése
  - Node
  - Way
  - Relation

# POI-k felkutatása jelenleg



- Egyezőségi modell
  - Tökéletes (Egyelőre ez fix 2000 m)
    - Ha van, akkor a közterület neve és a közelben van ilyen nevű utca
    - Ha van, akkor a házszám
    - POI típus (shop, amenity, egyéb címkék)
    - POI neve, esetleg több lehetőségből
      - `príma abc|prima abc|príma|prima`
  - Biztonságos (70-200 m; kód, konfigurációs fájl, adatbázis)
    - POI típus (shop, amenity, egyéb címkék)
    - POI neve, esetleg több lehetőségből
  - Nem biztonságos (5-23 m; kód, konfigurációs fájl, adatbázis)
    - POI típus (shop, amenity, egyéb címkék)

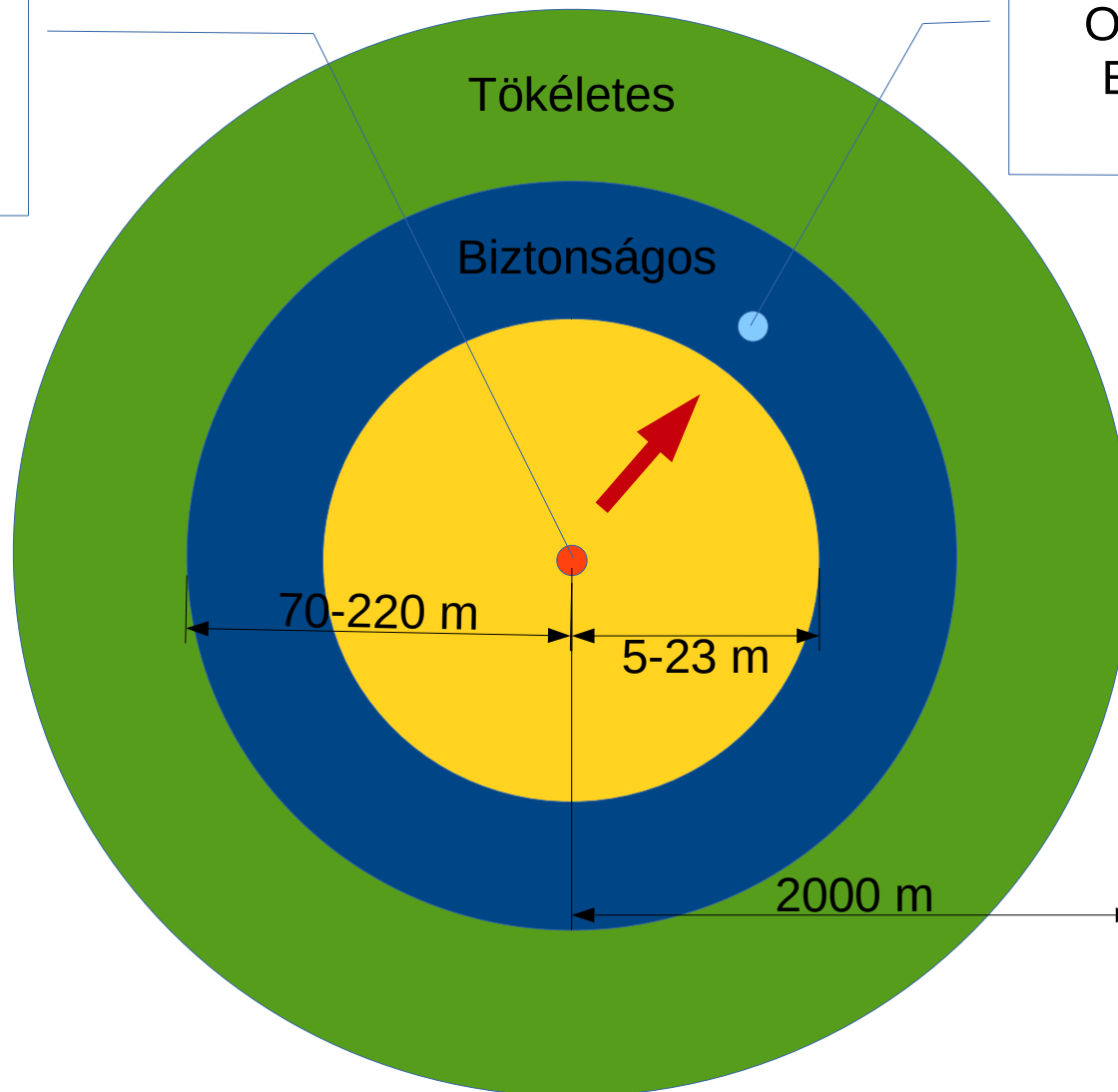


# Egyezés



Külső adatforrás  
alapján a POI

OSM POI helye  
Egyezik a név  
és típus







# Motiváció



- Egyszerűbb telepítés
  - A laptopom újrategelítése után határoztam el
- Sokkal könnyebb Nektek is kipróbálni
- Akár kiszolgálóra is költöztethető



# Kialakítás



- 3 konténer
  - PostgreSQL, PostGIS, fuzzystmatch konténer
    - Hstore használata
  - OSM adatok letöltése, betöltése
  - OSM\_POI\_matchmaker
- Docker compose fájlból konfigurálva

# Megfelelő-e konténer?

- Majdnem

- kartoza/postgis:11.0-2.5

- Alap adatbázis encoding nem UTF-8 volt.
    - shm\_size megemelése
      - Import során elfogyott a memória

- osm2pgsql

- Debian 9 (Stretch) alapú, régi verzió volt benne
      - Import során elfogyott a memória
    - Új konténer használata, amiben lefordítjuk a osm2pgsql-t
      - Megjelent a osm2pgsql 1.0.0 verzió \o/
      - Egyedi stílus és le- és betöltő script




Default tablespace	pg_default
Encoding	SQL_ASCII
Collation	C



Összefoglalás



- 
- Az adatok összegzése nehéz
    - Külső adathibák
    - OSM hibák
      - Lassú a folyamat
  - Mégis érdemes
    - Meghosszabbított kezünk lehet
      - Ellenőrizni is nehezebb
    - Jól növelhető a lefedettség – látszik, hogy az OSM ilyen téren el van maradva
    - Célrányosan hibák fedezhetőek fel

# Hogyan segíthetsz?



- Kimeneti fájlok átnézése, ellenőrzése, javítása
  - <https://github.com/KAMI911/osm-import-data/tree/master/hu-poi-data>
  - <https://github.com/KAMI911/osm-import-request>
- Hibás bemeneti adatok felkutatása
- Javaslatok további közös kulcs-érték párokra
- Hibák jelentése
- Újabb bemeneti adatok felkutatása
  - Könyvtárak
  - <http://kozadat.hu/kereso/intezmeny>

# Hogyan segíthetsz?



- További importálók készítése
  - Egységes formátumú bemeneti feldolgozó készítése, például GeoJSON vagy CSVű
- További engedélyek kérése
  - [https://wiki.openstreetmap.org/wiki/WikiProject\\_Hungary/Import%C3%A1l%C3%A1s/POI\\_adatok](https://wiki.openstreetmap.org/wiki/WikiProject_Hungary/Import%C3%A1l%C3%A1s/POI_adatok)
- Javítások és tesztek készítése
- Továbbfejlesztés
- Előállított adatok, egyeztetett POI címkék használata más projektben
  - <https://raw.githubusercontent.com/openstreetmap/id/master/data/presets/presets.json>



# Fejlesztési ötletek



- Tovább finomítani a keresést
  - Maximum távolságok meghatározása az adatok alapján
  - Még több megoldást a biztosabb találatok érdekében
  - Patch-tábla a hibás bejövő adatok számára
- POI pontok elhelyezése az azonos című (addr) épületbe
- Kimenő OSM-fájlok rendezése és szűrése
  - Irányítószám szerint (területi ellenőrzés)
  - Csak a létezők (meglévő POI-k pontosítása)
- Webes felület
  - Az adatbázisban lévő adatok javításához
  - Az OSM-mel történő összehasonlításhoz
  - OSM-be történő tömeges importáláshoz
- All the places használata: <https://www.alltheplaces.xyz/>



**Kérdések**





**Köszönöm a figyelmet!**

**Szalai Kálmán (KAMI)  
@kami911kami**