Dokumentace výstupu

Obsah dat

V datové tabulce je záznam o původní geolokaci projektu spolu s informací o nově odvozené geolokaci. Tam, kde je nová geolokace odvozena příliš extenzivně (např. rozpad projektu původně lokalizovaného na kraj do všech jeho obcí), je tedy možné vrátit se k původní geolokaci. (Prakticky by se to provedlo tak, že by se vybraly pro daný projekt (`prj_id`) řádky se všemi kombinacemi hodnot sloupce `geo_id_orig` a `level_orig`).

Kvůli velikosti souborů výstupová tabulka neobsahuje všechna metadata o projektech a příjemcích. Ta lze podle potřeby připojit z jiných zdrojů.

Jak vznikaly rozpady na obce

Projekty, kde geolokace byla vnitřně konzistentní Toto byla velká většina projektů.

Pro většinu projektů platilo několik jednoduchých pravidel, pokud projekt v datech neměl geolokaci na úrovni obce. Všechna pravidla byla ověřena alespoň na části dat, abychom měli jistotu, že aplikace pravidla nepřinese nesprávné výsledky u většího množství projektů.

- vycházíme z nejjemnější v datech uvedené lokalizace, tj. pokud byl projekt alokován zároveň na kraj a obec, vycházíme z obce, pokud zároveň na kraj a ORP, vycházíme z ORP atd.
- pokud je žadatel kraj nebo jeho příspěvkovka, je projekt rozpadnut do všech obcí kraje
- pokud je žadatel městská část Prahy nebo její příspěvkovka, projekt je alokován do této ZÚJ
- obec je alokována do sídla žadatele u těchto projektů:
  - výzvy pro školy
  - výzvy na zateplení
  - žadatel je obec
  - výzvy pro VŠ
  - sídlo žadatele je zmíněno v názvu projektu nebo je součástí názvu žadatele
  - specifické výzvy pro firmy (firemní vzdělávání)
  - výzvy pro dětské skupiny, předškolní vzdělávání apod.
  - výzvy na budování kapacit NNO
  - výzvy pro sociálně vyloučené lokality
- všechny verze Prahy (okres, ORP, kraj) jsou převedeny na obec Praha, též u všech projektů ve výzvách pouze pro Prahu
- projekty technické pomoci jsou převedeny do obce sídla příjemce
- rozpad na obce MAS, pokud je MAS zmíněna v názvu nebo anotaci projektu
- sídlo MAS u výzev na podporu kapaci MAS
- rozpad na obce v okrese/ORP/kraji, pokud:
- je okres/ORP zmíněn v názvu projektu
- je projekt původně lokalizován pouze na jeden okres, ORP nebo kraj
- projekty na krajské úrovni ve více než 11 kraji jsou rozpadnuty na celou ČR

Projekty bez geolokace v původních datech Část projektů měla geolokaci zadanou na úrovni chráněné oblasti, což ale nebylo v původním exportu. Z dodatečných dat se podařilo většinu z nich přiřadit k CHKO/NP a pomocí geodat poté rozpadnout na jednotlivé obce překrývající se s těmito územími. Současný postup zahrnuje do projektu všechny obce, do kterých dané chráněné území zasahuje.

Projekty s nekonzistentní lokalizací To jsou projekty, kde je např. jako místo realizace uveden zároveň Karlovarský kraj a město Třebíč.

V datech z léta 2020 jich asi 119.

U těch je potřeba úvaha a ruční zadání. Prakticky to vypadá tak, že skript vygeneruje excelový soubor, kde se do novéhou sloupečku zadá TRUE u řádků s tím geografickým údajem, kterých chceme použít. Pokud u nějakého projektu nezadáme TRUE u žádného řádku, použijou se všechny údaje.

Výsledný excelový soubor je potřeba uložit do adresáře `data-manual` a znovu spustit skripty `09_resolve-complicated.Rmd` a `10_compile-export.Rmd`.

Následně skript na základě tohoto zadání všechny projekty rozpadne na obce - tj. pokud jsme u nějakého projektu zadali, že se má použít geolokace na uvedený ORP, skript projekt rozpadne do všech obcí daného ORP.

Rozložení projektů

```
## # A tibble: 40 x 4
##    obec_puvod                                        rozpad_typ    rozpad_duvod
##    <chr>                                             <chr>         <chr>                                 <
##  1 doplnění obce nebo ZÚJ podle chráněných území     nic           <NA>
##  2 dovození obce nebo ZÚJ, kde tato úroveň chyběla   detekce_okres okres XY v názvu nebo anotaci
##  3 dovození obce nebo ZÚJ, kde tato úroveň chyběla   detekce_orp   ORP XY v názvu nebo anotaci
##  4 dovození obce nebo ZÚJ, kde tato úroveň chyběla   detekce_uzemi Název obce/MČ v názvu nebo anotaci
##  5 dovození obce nebo ZÚJ, kde tato úroveň chyběla   mc            žadatel je MČ
##  6 dovození obce nebo ZÚJ, kde tato úroveň chyběla   mc            žadatel je PO MČ
##  7 dovození obce nebo ZÚJ, kde tato úroveň chyběla   obce_v_kraj   1 kraj
##  8 dovození obce nebo ZÚJ, kde tato úroveň chyběla   obce_v_kraj   příjemce je kraj nebo jeho PO
##  9 dovození obce nebo ZÚJ, kde tato úroveň chyběla   obce_v_kraj   více krajů
## 10 dovození obce nebo ZÚJ, kde tato úroveň chyběla   obce_v_mas    MAS je příjemce
## # ... with 30 more rows
```

Výstupní soubory a formáty

Arrow datasety

Zpracovaná data jsou exportována do datasetu Arrow ve formátu Parquet. Arrow je knihovna pro efektivní skladování a načítání dat využitelná v různých prostředích (R, Python, Java, JavaScript, Ruby, Rust, Go, Julia, Matlab aj.) Parquet je konkrétní formát skladování dat v souborech na disku.

Výstupem je adresář souborů zanořených ve struktuře podadresářů - v našem případě adresář `data-output/dtl-all-arrow`. Z této struktury lze strojově odvodit datovou strukturu; členění dat do mnoha souborů umožňuje rychlé načítání části dat, např. pro jednotlivé OP nebo podle tzv. chunks (oddílů, na které jsou data rozdělena pro snadný export do většího množství Excel souborů.)

Náhledové Excel soubory

- jeden kus OPZ (cca 500 000 řádků z celkem cca 7 mil.)
- vzorek projektů zahrnující různé OP a typy řešených mezer v datech

Jejich schéma (názvy sloupců, jejich obsah a datový typ) odpovídají schématu níže.

Dokumentace výstupního exportu

```
## Warning: HTML tags found, and they will be removed.
##  * set `options(gt.html_tag_check = FALSE)` to disable this check
```

```
## Warning: HTML tags found, and they will be removed.
##  * set `options(gt.html_tag_check = FALSE)` to disable this check

## Warning: HTML tags found, and they will be removed.
##  * set `options(gt.html_tag_check = FALSE)` to disable this check

## Warning: HTML tags found, and they will be removed.
##  * set `options(gt.html_tag_check = FALSE)` to disable this check

## Warning: HTML tags found, and they will be removed.
##  * set `options(gt.html_tag_check = FALSE)` to disable this check

## Warning: HTML tags found, and they will be removed.
##  * set `options(gt.html_tag_check = FALSE)` to disable this check

## Warning: HTML tags found, and they will be removed.
##  * set `options(gt.html_tag_check = FALSE)` to disable this check

## Warning: HTML tags found, and they will be removed.
##  * set `options(gt.html_tag_check = FALSE)` to disable this check

## Warning: HTML tags found, and they will be removed.
##  * set `options(gt.html_tag_check = FALSE)` to disable this check

## Warning: HTML tags found, and they will be removed.
##  * set `options(gt.html_tag_check = FALSE)` to disable this check

## Warning: HTML tags found, and they will be removed.
##  * set `options(gt.html_tag_check = FALSE)` to disable this check

## Warning: HTML tags found, and they will be removed.
##  * set `options(gt.html_tag_check = FALSE)` to disable this check

## Warning: HTML tags found, and they will be removed.
##  * set `options(gt.html_tag_check = FALSE)` to disable this check

## Warning: HTML tags found, and they will be removed.
##  * set `options(gt.html_tag_check = FALSE)` to disable this check

## Warning: HTML tags found, and they will be removed.
##  * set `options(gt.html_tag_check = FALSE)` to disable this check
```

Pointblank Informa
[2021-01-14|12:25:15]

item
Table
Export všech upravených dat v jedné tabulce - pouze kódy území, bez metadat o projektech a příjemcích Arrow dataset, členěný do parquet souborů podle OP, typu rozpadu, původu ID obce a oddílu pro export (chunk).
Columns
prj_id character VYZNAM číslo projektu ČÍSELNÍK NA ZDROJ původní data, neupraveno radek integer VYZNAM pořadové číslo řádku v projektu ZDROJ generováno po rozpadu na obce level ordered, factor VYZNAM úroveň dovozené geolokace ZDROJ dovozeno rozpadem na obce a ZÚJ HODNOTY obec,

geo_id character VYZNAM kód obce nebo ZÚJ (pro MČ) bez prefixu NUTS ČÍSELNÍK ČSÚ číselník obcí (43), ZÚJ (51
level_orig character VYZNAM původní úrove HODNOTY kraj, orp, okres, obec, zuj
geo_id_orig character VYZNAM původní geolokace - kód území FORMA obce a ZÚJ včetně NUTS prefixu, ORP a okre
rozpad_duvod character VÝZNAM podrobná informace o způsobu dovození obce
obec_puvod character VYZNAM zdroj informace o obci
op_id character VYZNAM zkratka OP ve formě OP Z atd. (mezera, ne '_') ZDROJ převzato z původních dat
chunk integer VYZNAM číslo oddílu; oddíly čítají cca 500 000 řádků pro export do jednotlivých excelových souborů
rozpad_typ character VÝZNAM základní informace o způsobu dovození obce

---

2021-01-14 12:25:16 CET1.4 s2021-01-14 12:25:17 CET