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 n  
## <chr> <chr> <chr> <int>  
## 1 doplnění obce nebo ZÚJ podle chráněných území nic <NA> 26  
## 2 dovození obce nebo ZÚJ, kde tato úroveň chyběla detekce\_okres okres XY v názvu nebo anotaci 155  
## 3 dovození obce nebo ZÚJ, kde tato úroveň chyběla detekce\_orp ORP XY v názvu nebo anotaci 427  
## 4 dovození obce nebo ZÚJ, kde tato úroveň chyběla detekce\_uzemi Název obce/MČ v názvu nebo anotaci 916  
## 5 dovození obce nebo ZÚJ, kde tato úroveň chyběla mc žadatel je MČ 120  
## 6 dovození obce nebo ZÚJ, kde tato úroveň chyběla mc žadatel je PO MČ 533  
## 7 dovození obce nebo ZÚJ, kde tato úroveň chyběla obce\_v\_kraj 1 kraj 2578  
## 8 dovození obce nebo ZÚJ, kde tato úroveň chyběla obce\_v\_kraj příjemce je kraj nebo jeho PO 213  
## 9 dovození obce nebo ZÚJ, kde tato úroveň chyběla obce\_v\_kraj více krajů 866  
## 10 dovození obce nebo ZÚJ, kde tato úroveň chyběla obce\_v\_mas MAS je příjemce 206  
## # … 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.)](opz-one-chunk.xlsx)
* [vzorek projektů zahrnující různé OP a typy řešených mezer v datech](sample_export.xlsx)

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

## Dokumentace výstupního exportu