# StatBot.Swiss

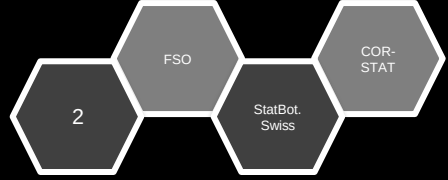
**Thomas Knecht**
**Statistical Office Canton Zürich**
**CORSTAT**

**Dr. Christian Ruiz**
**Statistical Office Canton Zürich**
**CORSTAT**


Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Eidgenössisches Departement des Innern EDI
Département fédéral de l'intérieur DFI
Federal Department of Home Affairs FDHA
**Bundesamt für Statistik BFS**
**Office fédéral de la statistique OFS**
**Federal Statistical Office FSO**




# What is StatBot?

## Aim:

Simplify for a range of users  access to information in OGD

## Technical goal:

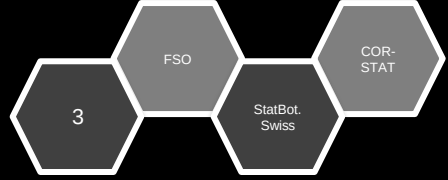
StatBot is supposed to convert natural language questions into DB-queries (SQL for now)

## What it is not:

StatBot is not a "chatty" chatbot for conversation

## How:

Open-source solution and replicable for everyone



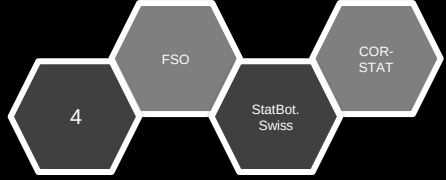

# "The way is the aim"

## Technical goals:

- Feasibility of the endeavour
- Technical limitations
- Realistic assessment

## Broader impact:

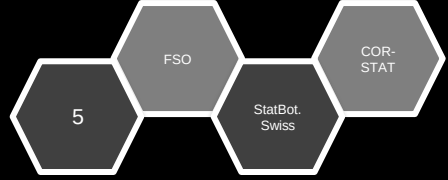
- Improve common diffusion
- build competence and knowledge
- Increase harmonization and standardization




# Main parts of the Project

Part I
"Data Warehouse"
(FSO + CORSTAT + SDSC + ETC.)

Part II
Machine Learning (academic partner)




# Data requirements

– ML-Part needs a certain data foundation

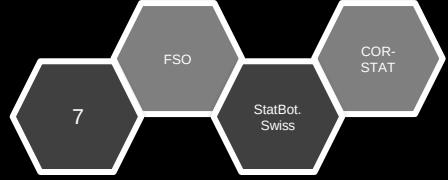
– ML-Part has certain requirements on data structure

→ Human readable

→ long format (one value column)

→ same structure of the datasets

| spatialunit_name | time_value | value | population_type | gender | age_group_1 | citizenship_category |
|---|---|---|---|---|---|---|
| Altstadt Grossbasel | 1980-12-31 | 4 | total | Male | 2 years | Switzerland |
| Altstadt Grossbasel | 1980-12-31 | 1 | total | Male | 9 years | Foreign country |
| Altstadt Grossbasel | 1980-12-31 | 1 | total | Male | 11 years | Foreign country |
| Altstadt Grossbasel | 1980-12-31 | 6 | total | Male | 13 years | Switzerland |
| Altstadt Grossbasel | 1980-12-31 | 7 | total | Male | 25 years | Foreign country |
| Altstadt Grossbasel | 1980-12-31 | 18 | total | Male | 25 years | Switzerland |
| Altstadt Grossbasel | 1980-12-31 | 8 | total | Male | 30 years | Foreign country |
| Altstadt Grossbasel | 1980-12-31 | 5 | total | Male | 33 years | Foreign country |
| Altstadt Grossbasel | 1980-12-31 | 27 | total | Male | 34 years | Switzerland |
| Altstadt Grossbasel | 1980-12-31 | 8 | total | Male | 35 years | Foreign country |
| Altstadt Grossbasel | 1980-12-31 | 24 | total | Male | 36 years | Switzerland |
| Altstadt Grossbasel | 1980-12-31 | 9 | total | Male | 39 years | Switzerland |
| Altstadt Grossbasel | 1980-12-31 | 4 | total | Male | 43 years | Foreign country |
| Altstadt Grossbasel | 1980-12-31 | 3 | total | Male | 46 years | Foreign country |
| Altstadt Grossbasel | 1980-12-31 | 13 | total | Male | 47 years | Switzerland |
| Altstadt Grossbasel | 1980-12-31 | 1 | total | Male | 48 years | Foreign country |
| Altstadt Grossbasel | 1980-12-31 | 2 | total | Male | 51 years | Foreign country |
| Altstadt Grossbasel | 1980-12-31 | 8 | total | Male | 52 years | Switzerland |




# BUT!

Human readable is quite difficult to manage

Particularly for a large amount of datasets

# Underlying data warehouse



Use of codes instead of human readable names

| spatialunit_ontology | spatialunit_hist_id | time_value | period_value | value | gender_of_child | citizenship_category_of_child | age_class_of_mother |
|---|---|---|---|---|---|---|---|
| CH | 0 | 1985-12-31 | NA | 26939 | -111 | 1 | 2 |
| CH | 0 | 1985-12-31 | NA | 17809 | -111 | 1 | 3 |
| CH | 0 | 1985-12-31 | NA | 5424 | -111 | 1 | 4 |
| CH | 0 | 1985-12-31 | NA | 846 | -111 | 1 | 5 |
| CH | 0 | 1985-12-31 | NA | 11675 | -111 | 2 | -111 |
| CH | 0 | 1985-12-31 | NA | 3768 | -111 | 2 | 1 |
| CH | 0 | 1985-12-31 | NA | 4065 | -111 | 2 | 2 |
| CH | 0 | 1985-12-31 | NA | 2614 | -111 | 2 | 3 |
| CH | 0 | 1985-12-31 | NA | 1036 | -111 | 2 | 4 |
| CH | 0 | 1985-12-31 | NA | 192 | -111 | 2 | 5 |
| CH | 0 | 1985-12-31 | NA | 37965 | 1 | -111 | -111 |
| CH | 0 | 1985-12-31 | NA | 7918 | 1 | -111 | 1 |
| CH | 0 | 1985-12-31 | NA | 15908 | 1 | -111 | 2 |
| CH | 0 | 1985-12-31 | NA | 10315 | 1 | -111 | 3 |
| CH | 0 | 1985-12-31 | NA | 3317 | 1 | -111 | 4 |
| CH | 0 | 1985-12-31 | NA | 507 | 1 | -111 | 5 |
| CH | 0 | 1985-12-31 | NA | 32023 | 1 | 1 | -111 |

Note: Source, definition and "flags" are missing here as columns

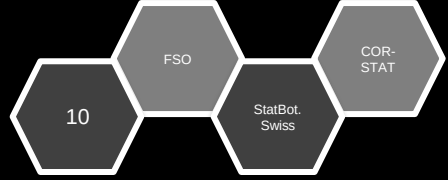
# Underlying data warehouse

→ Dictionaries are used for a translation into names

→ Single source of truth: needs to be adjusted only once!

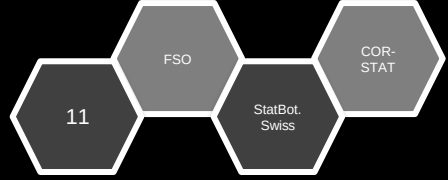
→ Makes it scalable

| dim_id | unique_name | dim_name_de | dim_name_fr | dim_name_it | dim_name_en | value_id | def_value_name_de | def_value_name_fr | def_value_name_it | def_value_name_en |
|---|---|---|---|---|---|---|---|---|---|---|
| 97 | citizenship_category | Staatsangehörigkei... | Nationalité catégo... | Nazionalità categ... | Citizenship categ... | -111 | Staatsangehörigkeit (... | Nationalité (catégo... | Nazionalità (catego... | Citizenship (category) - tota |
| 97 | citizenship_category | Staatsangehörigkei... | Nationalité catégo... | Nazionalità categ... | Citizenship categ... | 1 | Schweiz | Suisse | Svizzera | Switzerland |
| 97 | citizenship_category | Staatsangehörigkei... | Nationalité catégo... | Nazionalità categ... | Citizenship categ... | 2 | Ausland | Etranger | Straniera | Foreign country |




# Challenges

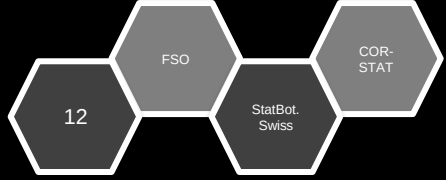
- Pull approach → we collect the data from different institutions
- Different input data structures
- Different definitions of dimensions
- Different spatial levels
- Different file formats
- Sometimes very large datasets (on renku we only have 8GB RAM)
- Etc.




# Import Pipeline

- Programmed in R
- Run in renku
- Object oriented (S3 variant)
- For large objects use disk and not R-RAM
- Parametrized
- Highly modular and standardized

⇒ Code can be reused
⇒ Very large datasets can be processed with relatively small RAM
⇒ New institutions can be easily added with only little adjustment




Describe Dataset

(parametrize)

Create Dataset Object

Download data

Prep data

Dataset specific data tansformation

Write data into DWH

# Describe dataset

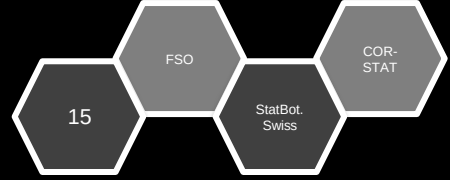
| id | Version/Status | class_name | name_de | name_en | description_de | unit_short_de | unit_long_de | unique_dimension_names |
|---|---|---|---|---|---|---|---|---|
| 1_01_001_CH | 4.2 | resident_population | Ständige Wohnbevölkerung | Resident Population | Ständige Wohnbevölkerung | Pers. | Personen | c("population_type", "citizenship_category", "gender", "age_group_1") |
| 1_01_001_CH_OLD | OLD | resident_population | Ständige Wohnbevölkerung | Resident Population | Ständige Wohnbevölkerung | Pers. | Personen | c("gender","age_group_1_until_99" ) |
| 1_01_004_CH | 4.2 | nativity | Geburten | Nativity | Geburten | Pers. | Personen | c("gender_of_child", "citizenship_category_of_child", "age_class_of_mother") |
| 1_01_005_CH | 4.2 | deaths_citizenships_maritalstatus_age | Todesfälle | Deaths | Todesfälle | Pers. | Personen | c("gender", "citizenship_death_deleteme","marital_status_7","age_group_1") |
| 1_01_006_CH | 4.2 | death | Todesfälle | Deaths | Todesfälle | Pers. | Personen | c("gender","citizenship_category_alt_names", "marital_status", "age_group_10") |
| 1_01_008_CH | 4.2 | marriages | Eheschliessungen | Marriages | Eheschliessungen | Ehen | Eheschliessungen | c("citizenship_category_husband","citizenship_category_wife" ) |
| 1_01_009_CH | 4.2 | marriages_ageclass_citizenship | Eheschliessungen | Marriages | Heiraten nach gegenseitiger Altersklasse, Staatsan | Ehen | Eheschliessungen | c("age_class_husband_marriages","citizenship_selection_husband_alt_names","marita |
| 1_01_010_CH | 4.2 | divorces_durationofmarriage_citizenshipcategories | Scheidungen | Divorces | Scheidungen nach institutionellen Gliederungen, Eh | Sch. | Scheidungen | c("duration_of_marriage","citizenship_category_husband", "citizenship_category_wife |
| 1_01_011_CH | 4.2 | divorces_durationofmarriage_ageclasses | Scheidungen | Divorces | Scheidungen nach Kanton, Ehedauer und Altersklas | Sch. | Scheidungen | c("duration_of_marriage", "age_class_husband", "age_class_wife") |
| 1_01_012_CH | 4.2 | divorces_citizenships | Scheidungen | Divorces | Scheidungen nach Ehedauer, gegenseitiger Staatsa | Sch. | Scheidungen | c("duration_of_marriage", "citizenship_selection_husband", "age_class_husband", "citi |
| 1_01_013_CH | 4.2 | civil_unions | Eingetragene Partnerschafter | Civil unions | Eingetragene Partnerschaften nach Kanton, gegens | Eing. Partnersch | Eingetragene Partn | c("civil_union_type", "marital_status_first_partner", "marital_status_second_partner", |
| 1_01_014_CH | 4.2 | civil_unions_dissolved | Aufgelöste Partnerschaften | Dissolved civil unions | Aufgelöste Partnerschaften | Aufg. Partnersch | Aufgelöste Partner | c("civil_union_type","duration_of_civil_union", "citizenship_category_partners","age_c |
| 1_01_015_CH | 4.2 | stillbirth | Todgeburten | Stillbirths | Todgeburten | Tode | Todgeburten | c("demographic_characteristic_stillbirth","gender_of_child_stillbirth") |
| 1_01_016_CH | 4.2 | favorite_firstname_rank_girl | Babynamen Rang weiblich | Favorite first name rank g | Weibliche Vornamen der Neugeborenen nach Sprac | Rang | Rang | c("first_name_girl") |
| 1_01_017_CH | 4.2 | favorite_firstname_amount_girl | Babynamen Anzahl weiblich | Favorite first name amou | Weibliche Vornamen der Neugeborenen nach Sprac | Anzahl | Anzahl | c("first_name_girl") |
| 1_01_018_CH | 4.2 | favorite_firstname_rank_boy | Babynamen Rang männlich | Favorite first name rank b | Männliche Vornamen der Neugeborenen nach Sprac | Rang | Rang | c("first_name_boy") |
| 1_01_019_CH | 4.2 | favorite_firstname_amount_boy | Babynamen Anzahl männlich | Favorite first name amou | Männliche Vornamen der Neugeborenen nach Sprac | Anzahl | Anzahl | c("first_name_boy") |
| 1_01_020_CH | 4.2 | immigration_incl_change_of_population_type | Einwanderung inkl. Änderun | Immigration incl. change | Wanderung der ständigen Wohnbevölkerung nach in | Pers. | Personen | c( "citizenship_selection_alt_names","gender") |
| 1_01_021_CH | 4.2 | emigration | Auswanderung | Emigration | Auswanderung | Pers. | Personen | c( "citizenship_selection_alt_names","gender") |
| 1_01_022_CH | 4.2 | internal_migration_intercantonal_arrival | Interkantonaler Zuzug | Internal migration interca | Interkantonaler Zuzug | Pers. | Personen | c( "citizenship_selection_alt_names","gender") |
| 1_01_023_CH | 4.2 | internal_migration_intercantonal_departure | Interkantonaler Wegzug | Internal migration interca | Interkantonaler Wegzug | Pers. | Personen | c( "citizenship_selection_alt_names","gender") |
| 1_01_024_CH | 4.2 | internal_migration_intracantonal_arrival | Intrakantonaler Zuzug | Internal migration intraca | Intrakantonaler Zuzug | Pers. | Personen | c( "citizenship_selection_alt_names","gender") |
| 1_01_025_CH | 4.2 | internal_migration_intracantonal_departure | Intrakantonaler Wegzug | Internal migration intraca | Intrakantonaler Wegzug | Pers. | Personen | c( "citizenship_selection_alt_names","gender") |

# Dataset definition as yaml

- Can be checked (structure and input)
- Human readable
- Can be easily transformed into an R-Object
- Easier to write than json

```yaml
id: 1_01_011_CH
source_id:
- px-x-0102020203_104
name:
  en: Divorces
  de: Scheidungen
unit_short:
  de: Sch.
read_class: px
unique_dimension_names:
- duration_of_marriage
- age_class_husband
- age_class_wife
modules:
- etl_spatialunit
- join_dimension_values
- final
institution: bfs
smallest_spatial_entity: canton
time_col: Jahr
```

# Dataset Object




Metadata of dataset-object

- ds
  - id
  - source_id
  - name
  - unit_short
  - read_class
  - unique_dimension_names
  - modules
  - institution
  - smallest_spatial_entity
  - time_col
  - time_interval
  - institution_dim_names
  - spatial_col
  - obs_value_col
  - description
  - unit_long
  - class_name
  - source
  - separator_var
  - melt_vars
  - arrow_pointer
  - basename_template
  - metadata
  - arrow_query

Specific metadata of dataset

Data-related content

```
Browse[1]> class(ds)
[1] "dataset"     "px"          "1_01_011_CH" "bfs"         "canton"      "year"
```

# Dataset download

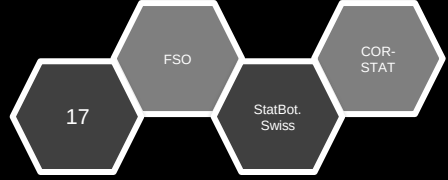
– Based on institution, a specific download link is generated

⇒ Different methods for different institutions

```r
generate_urls <- function(ds) UseMethod("generate_urls", ds)

#' Generate URLs for downloading datasets from BS
#' @param ds The dataset object
#' @export
generate_urls.bs <- function(ds) {
  paste0("https://data.bs.ch/api/v2/catalog/datasets/", ds$source_id, "/exports/csv")
}

#' Generate URLs for downloading datasets from BL
#' @param ds The dataset object
#' @export
generate_urls.bl <- function(ds) {
  paste0("https://data.bl.ch/explore/dataset/", ds$source_id, "/download/?format=csv")
}
```




# Data Preparation

file to parquet

toparquet.csv
toparquet.px
toparquet.json
etc.

Rename Data columns

rename.default
rename.bfs
rename.gpzh
etc.

Add time column and extract the spatial reference

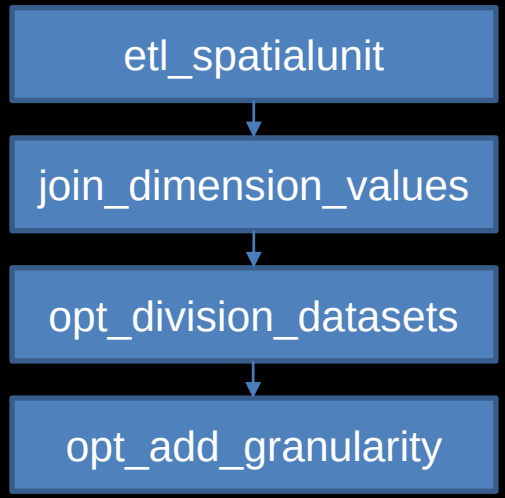
add_time.year
add_time_year.bfs
add_time_year.bs
add_time.month
add_time.day

# Transform data

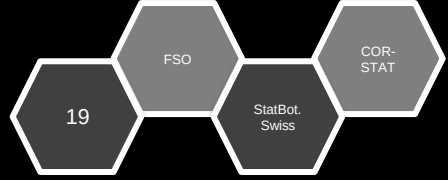
For each dataset a different pipeline is constructed.

**Currently existing modules:**

- opt_spatial_filter
- etl_spatialunit
- join_dimension_values
- opt_division_datasets
- opt_add_granulatiry
- opt_set_totals
- opt_add_totals
- opt_melt




```
df <- initiator(ds) %>%
  purrr::reduce(modules, ~ module_wrapper(..2, ..1), .init = .)
```




# Write data

Data is written to a parquet file

The data is also validated

- check if NA values in dimension columns
- check if date column has the correct type
- …

# Package "Arrow"

- Could be a presentation for itself

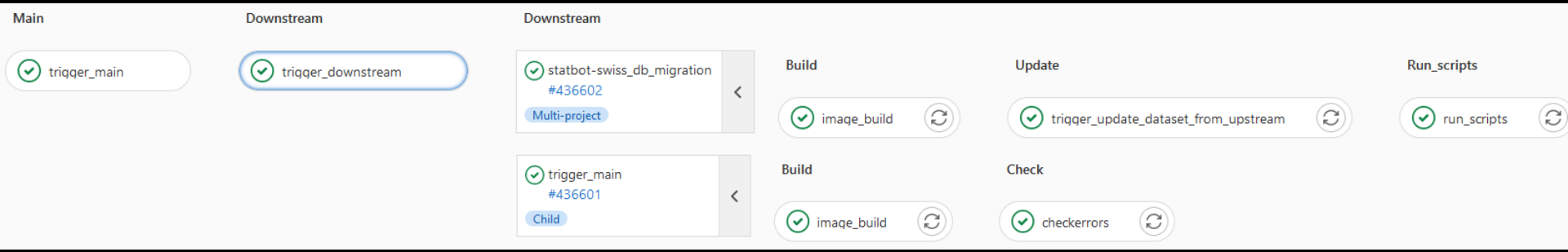
- The ETL is not run step by step
- Instead:
  - A recipe is constructed
  - "When bow is streched":
    "arrow is shot" and ETL executed
  - For large files:
    "Split into several arrows"

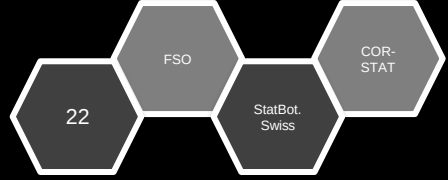
APACHE ARROW

A cross-language development platform for in-memory analytics

# CI/CD as a useful helper

- Continuous integration allows to automatically trigger processes
- We defined certain processes for different steps of a data integration to run checks or to autom. launch ETL
- Multi-Repo

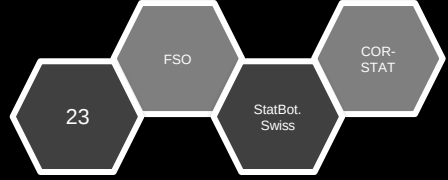





# Current data status

- 244 generated "Statbot"-Datasets as parquet files
- Approx 71 regional datasets
- Over 80% of these datasets in topic population
- Only 3 exceptions (important due to scaling)

- Data from FSO,BS,TG,AG,ZH,BL, and SSZ

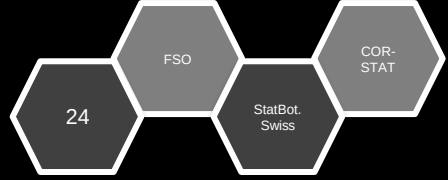



# Technical Learnings

- S3-Object oriented programming has the flexibility to define the pipelines per dataset without producing too much code

- "arrow" is a powerful interface to work with large datasets that would exeed the available RAM

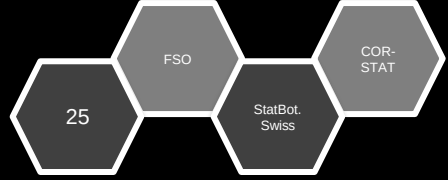
- Parquet file is a very efficient (around factor 30) data storage and can be accessed by multiple interfaces.




# Data Learnings

- Currently harmonization of the data only in the structure
- Ground definitions not harmonized since it is too big of a task for the statbot project

- BUT:
  - Through the integration of national and regional data we get an overview
  - a mapping of the dimension names as well as the definitions is generated

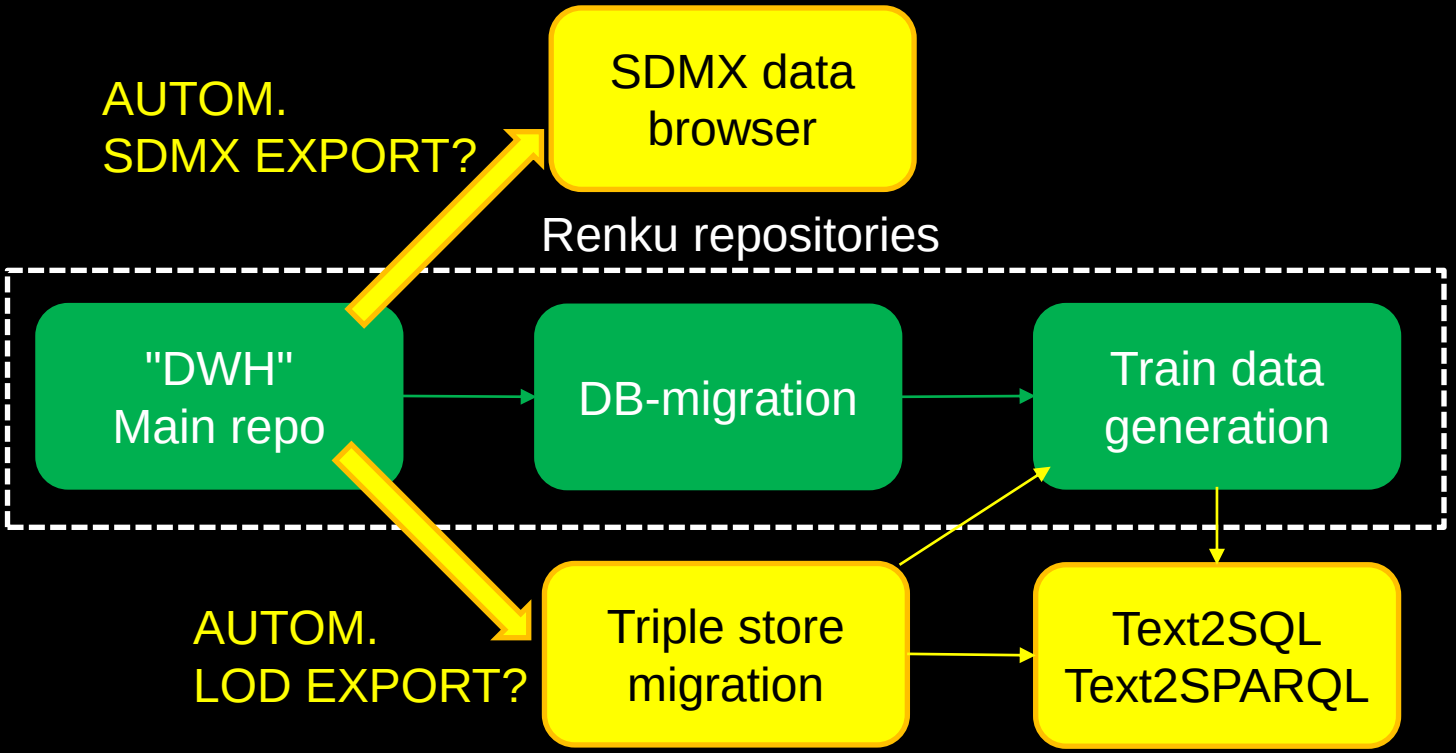



# Work in progress

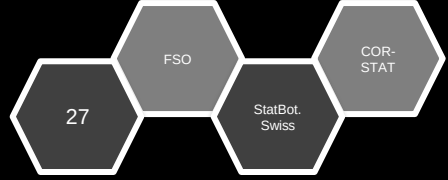
- We are still developing modules for special input data structures
- A shiny-app to easily add new dimensions and change existing dimension files
- Adding additional institutions
- The link to the DB (for part 2) is under construction

# Opportunities





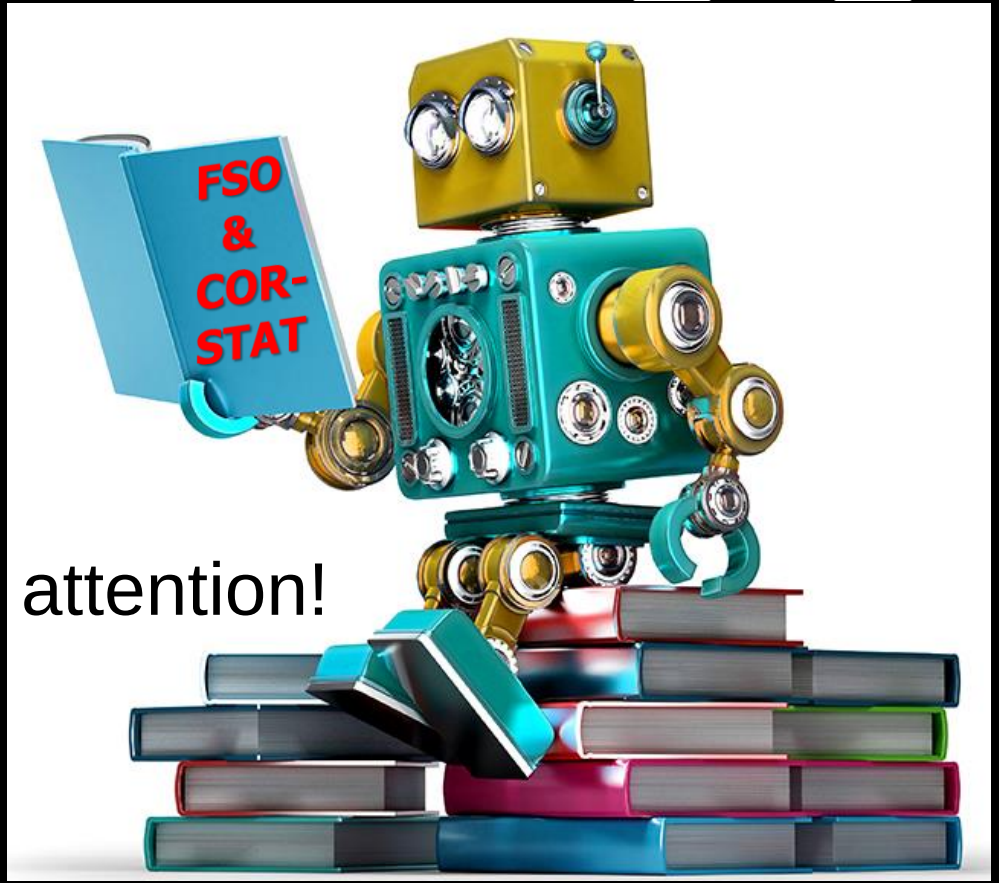

Claim: We can now, having a defined data structure, automatically export/generate SDMX and LOD




# Importance of DWH

"Even after the project statbot one day will formally be finished, I think that the DWH will remain. Maybe under another name and not focused on ML, but it seems to me to be the first database in Switzerland where data of different data producers is in a common, harmonized structure."

Thank you for your attention!




Source: CC0 Public Domain with modifications