



To be FAIR, what is missing in Official Statistics?

Statistics Netherlands

Olav ten Bosch, Edwin de Jonge, Henk Laloli

COSMOS 2024, 11-12 April Paris

Contents

- The official statistics landscape, what is that?
- Software perspective
- Adding linked data to the landscape
- What about FAIR digital objects (FDO)?
- Wrap-up



Statistics Netherlands output

The collage displays several key outputs from Statistics Netherlands:

- Population dashboard:** Shows population figures with a counter displaying 17,778,852. It includes a 'Population by sex' table and a 'Regional' facts page showing 189 people per sq km.
- News article:** 'Inflation rate up to 14.5 percent in September' with a photo of a gas stove.
- Consumer prices table:** 'Consumer prices; price index 2015=100' showing a table of CPI values and year-on-year changes from June 2021 to September 2022.
- Regional facts:** A page for '189 people per sq km' with various demographic and economic indicators.

Layered Open Data architecture

News articles



Thematic pages & visualisations



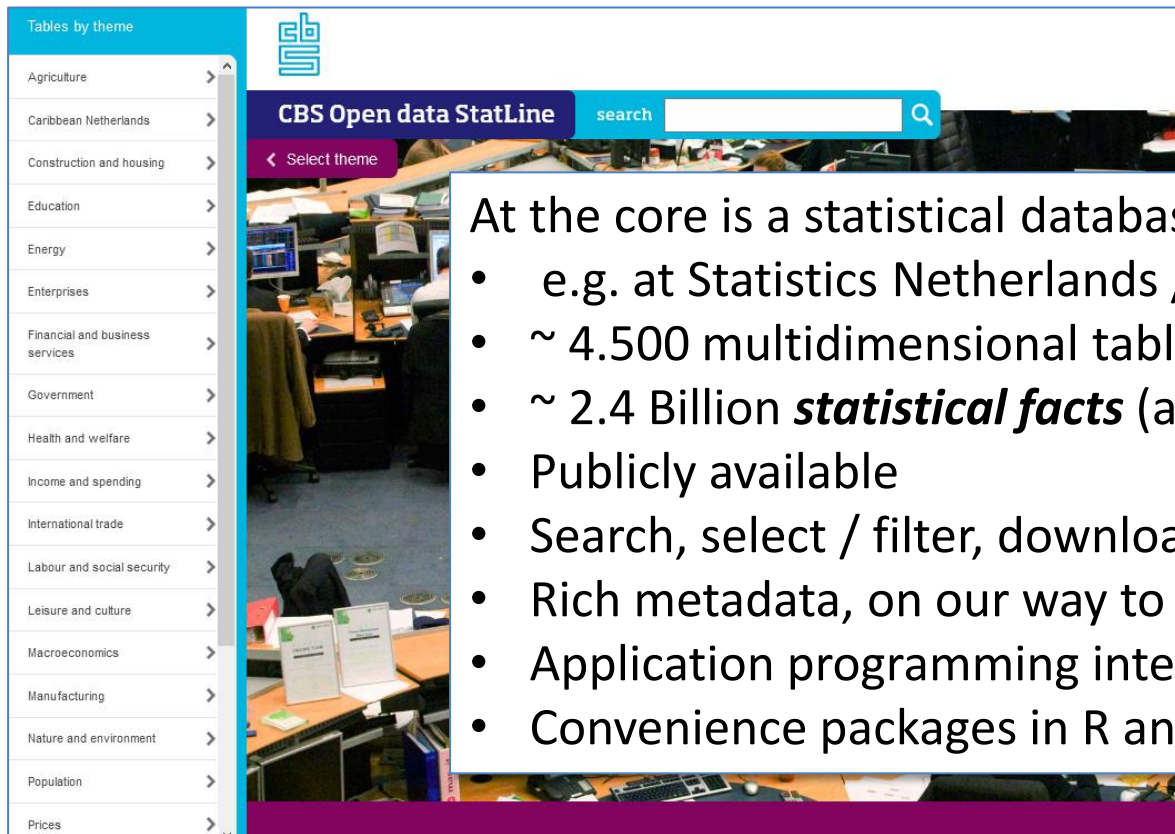
Selected data slices



Statistical Facts



Statistical dissemination database



Tables by theme

- Agriculture
- Caribbean Netherlands
- Construction and housing
- Education
- Energy
- Enterprises
- Financial and business services
- Government
- Health and welfare
- Income and spending
- International trade
- Labour and social security
- Leisure and culture
- Macroeconomics
- Manufacturing
- Nature and environment
- Population
- Prices

CBS Open data StatLine search

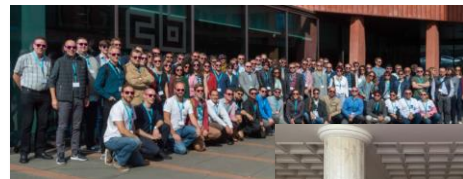
Select theme

At the core is a statistical database:

- e.g. at Statistics Netherlands / CBS:
- ~ 4.500 multidimensional tables
- ~ 2.4 Billion ***statistical facts*** (active tables)
- Publicly available
- Search, select / filter, download
- Rich metadata, on our way to SDMX
- Application programming interface (API): Odata
- Convenience packages in R and Python

Explore via OS Software

- Using the “awesome list of official statistics software”
- A *community approach* to knowledge management
- To *collectively remember useful software* in official statistics
- Maintained by *statistical community*
- *Conferences*

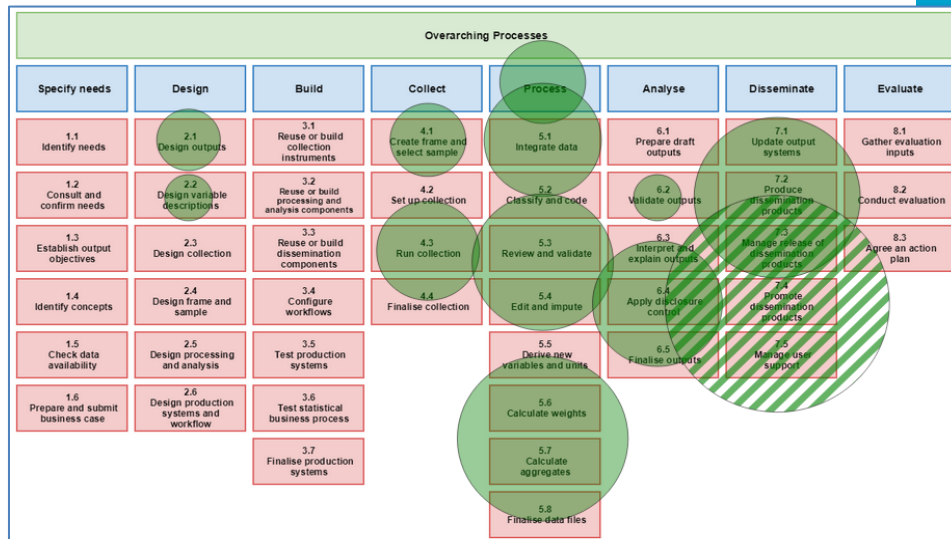


uRos17-23

• How:

- Using [awesome concept](#)
- Github repo
- [awesomeofficialstatistics.org](https://www.awesomeofficialstatistics.org)

awesome packages by GSBPM



Statistical disclosure control (GSBPM 6.4)

- GitHub v5.1.7b3 last commit march license EUPL-1.2

Java and C++ application [Mu-ARGUS](#). Tool to create safe micro-

- GitHub v4.2.4.2 license EUPL-1.2

Java C++ Fortran and Delphi application [T-ARGUS](#). Tool to protect

- CRAN 5.7.6 – 2 months ago license GPL-2

R package [sdcMicro](#). Disclosure control for statistical microdata.

- CRAN 0.32.6 – 4 months ago

R package [sdcTable](#). Disclosure control for tabulated data.

Sampling (GSBPM 4.1)

- CRAN 2.10 – a month ago license GPL (>= 2)

R package [sampling](#). Several algorithms for drawing survey samples, including sampling designs (high entropy, systematic, Rao-Sampford, etc.), and calibration

- CRAN 4.0 – 4 years ago license GPL (>= 2)

R package [surveyplanning](#). Tools for sample survey planning, including sample expected precision for the estimates of totals, and calculation of optimal sample

Data integration and record linkage (GSBPM 5.1)

- CRAN 0.3.4 – 5 months ago license GPL-3

R package [reclin2](#). Functions to assist in performing probabilistic record linkage of pairs, comparing records, em-algorithm for estimating m- and u-probabilities, for also be used for pre- and post-processing for machine learning methods for record

- CRAN 0.4-12.4 – a year ago license GPL (>= 2)

R package [RecordLinkage](#). Implementation of the Fellegi-Sunter method for record

- CRAN 1.4.1 – 2 years ago license GPL (>= 2)

R package [StatMatch](#). Statistical Matching or Data Fusion

- CRAN 0.6.1 – 24 days ago license GPL (>= 3)

R package [fastLink](#). Implements a Fellegi-Sunter probabilistic record linkage model

Over 30 software packages, giving access to > 60 data providers majority are R-packages

Access to official statistics (GSBPM 7.4)

- CRAN 0.6-3 – 7 months ago license GPL (>= 2)

R package [rdsdmx](#). Access to data or metadata from statistical organisations that support SDMX. The package contains a list of SDMX access points of various national and international statisti

- CRAN 0.3.1 – 7 months ago license GPL-3

R package [readsdmx](#). Read SDMX into dataframes from local SDMX-ML file or web-service. Pa OECD.

- GitHub v2.14.0 last commit last wednesday license Apache-2.0

Python [sdmx](#). Python library that implements SDMX 2.1 to explore data from SDMX data provi data and metadata and convert it into Pandas objects.

- CRAN 0.4.3 – 7 months ago license MIT + file LICENSE

R package [rjstat](#). Read and write data sets in the JSON-stat format.

- PyPI v2.4.0 license Apache License 2.0

Python [pyjstat](#). Read and write JSON-stat.

- GitHub v0.2.8 last commit march 2023 license MIT

Java application [json-stat.java](#). Read and write JSON-stat. By Statistics Norway.

- CRAN 0.2.5 – 2 years ago license CC0

R package [oecd](#). Search and Extract Data from the OECD

- CRAN 0.8.21 – 7 months ago license BSD_2_clause + file LICENSE

R package [sorvi](#). Finnish Open Government Data Toolkit

- CRAN 4.0.0 – 3 months ago license BSD_2_clause + file LICENSE

R package [eurostat](#). Tools to download data from the Eurostat database together with search manipulation utilities.

- CRAN 0.22.5 – 3 months ago license EUPL

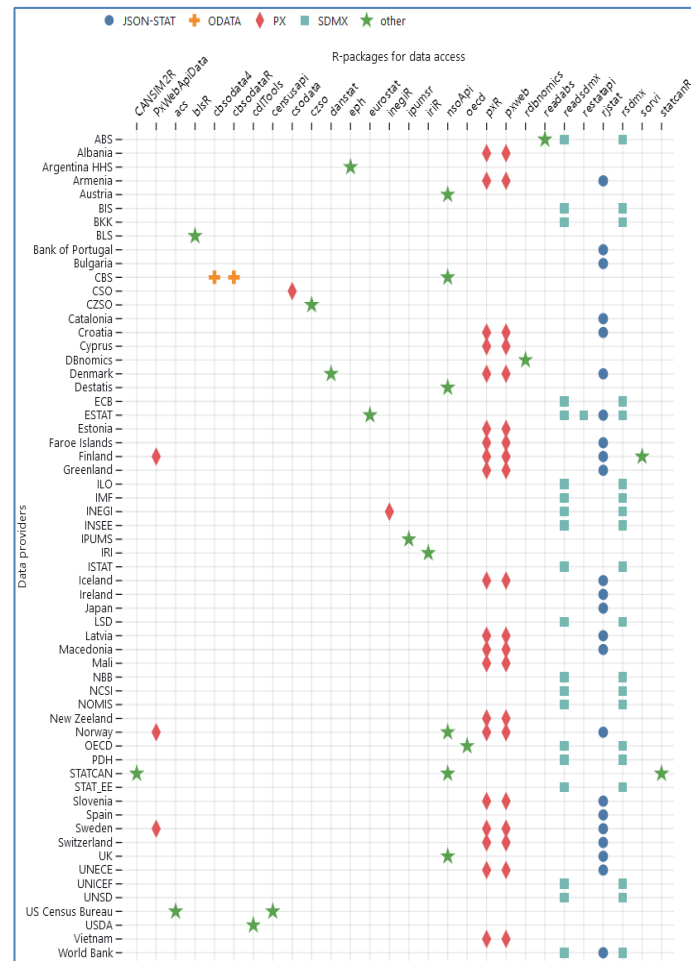
R package [restatapi](#). Search and retrieve data from Eurostat database, by Eurostat.

- CRAN 2.1.4 – 5 years ago license GPL-3

“access to official statistics” software landscape

- Matrix from docs, links to web pages and packages execution: packages (38) x dataproviders (60) x standards (5)

- Standards:
 JSON-STAT
 ODATA
 PX
 SDMX
 other

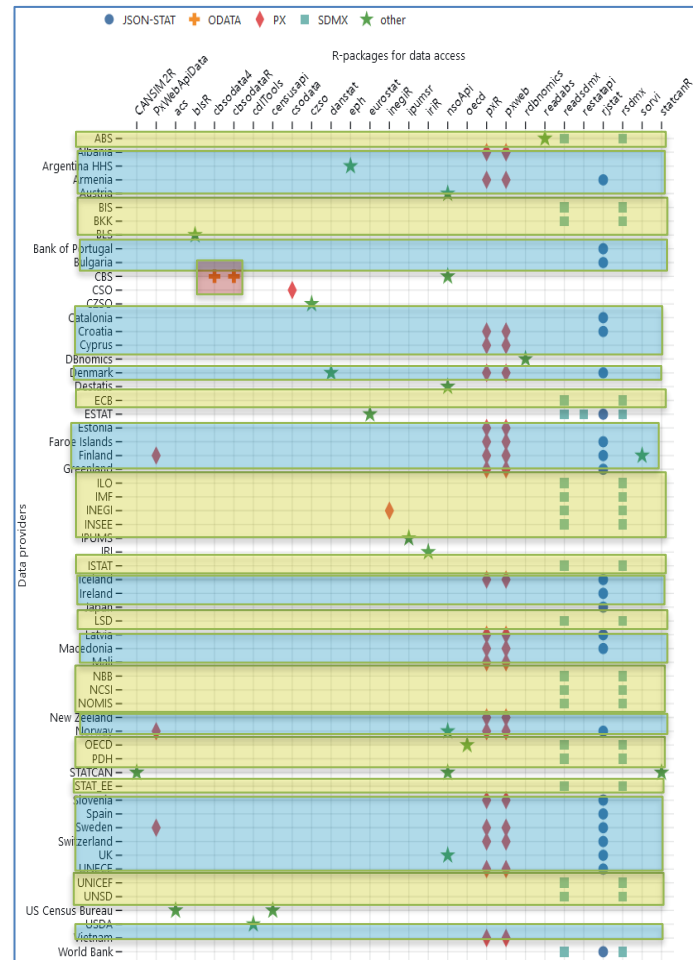


“access to official statistics” software landscape

- Matrix from docs, links to web pages and packages execution: packages (38) x dataproviders (60) x standards (5)

- Standards:

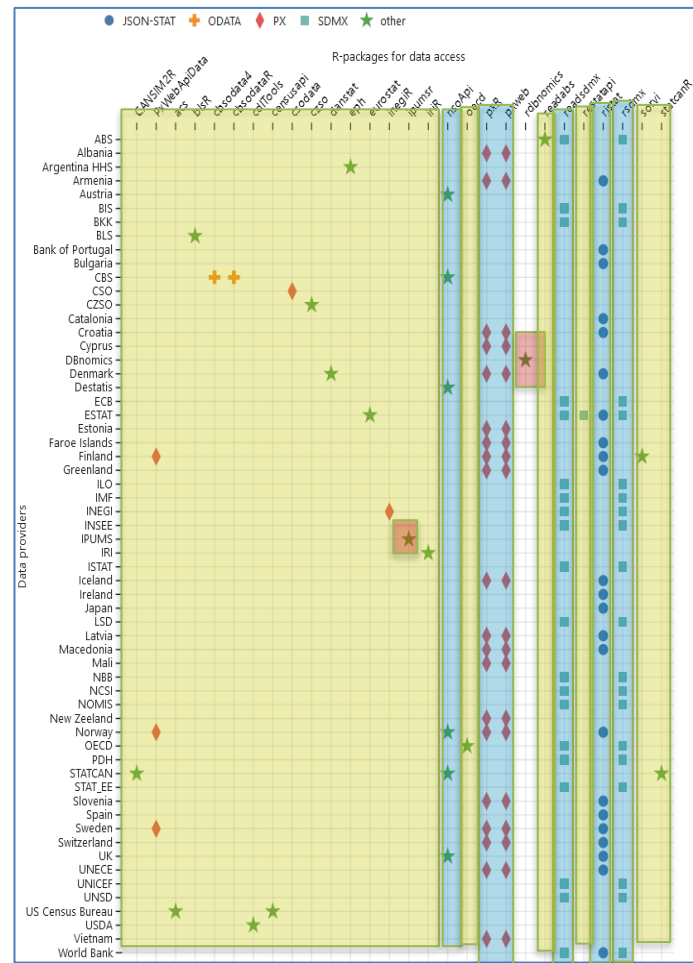
- JSON-STAT/PX v. SDMX: almost disjunct worlds
- ODATA: CBS only



[https://observablehq.com/@olavtenbosch/access to official statistics](https://observablehq.com/@olavtenbosch/access-to-official-statistics)

“access to official statistics” software landscape


- Standards-oriented packages:
rsdmx, *readsdmx*, *rjstat*, *px**
- Data provider-centric packages:
inegiR, *readabs*, *statcanR*, *eurostat*
- Official statistics aggregator sites:
rdbnomics: economic data
ipumsr: census & survey data
time&space harmonised



[https://observablehq.com/@olavtenbosch/access to official statistics](https://observablehq.com/@olavtenbosch/access%20to%20official%20statistics)

Features commonly offered

- **endpoint hiding**: wrapping the preconfigured endpoint(s) in a function
- **catalogue retrieval**: to list the availability datasets on the endpoint(s)
- **search**: to search for datasets or within datasets
- **endpoint queries**: query for subsets / slices on the endpoint(s) side
- **local queries**: the ability to easily slice or filter on the client
- **caching**: preventing unnecessary roundtrips
- **cartographic queries**: retrieve a geo data or a map with the data
- **registry access**: access to coordinated metadata in registries



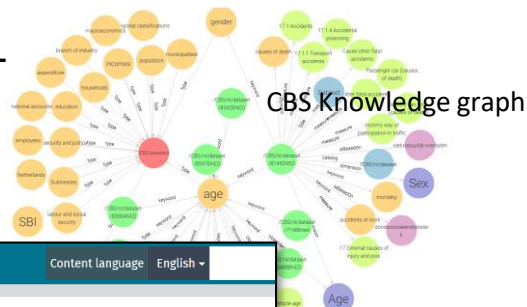
Could we build a “one-for-all” solution with all features?

Software features supporting FAIR principles

Software feature	Findability	Accessibility	Interoperability	Reusability
endpoint hiding	yes	yes		
catalogue retrieval	yes	yes		
search	yes			
endpoint queries		yes		
local queries		yes		
caching		yes		
cartographic queries	yes		yes	yes
registry access	yes	yes	yes	

Official statistics and linked data (1)

- Multiple stat. organisations offer (meta)data as Linked Data (LD)
- URIs, SKOS, XKOS, schema.org, SPARQL



Publications Office of the European Union

Search nace2

Law European data Public procurement EU Publications Research & Innovation

EU Vocabularies

Home Controlled vocabularies Models Business collections

EU Vocabularies > Controlled vocabularies > Taxonomies > nace2

Permanent link

Asset

Statistical Classification of Economic Activities in the European Union

Version: 2 **LATEST**

URI: <http://publications.europa.eu/resource/dataset/nace2>

Type of dataset: Taxonomy

About Downloads Documentation Alignments Release notes Diff files Advanced

Published: 2022-07-15 NACE Rev. 2 (Statistical Classification of Economic Activities in the European Union (EU). It is the European Industrial Classification of All Economic Activities), Rev. 4, which is maintained by the United Nations.

Author: Eurostat

Publisher: Publications Office of the European Union

Legal basis: Regulation (EC) No 1893/2006 of the European Parliament and of the Council of 20 December 2006

op.europa.eu/en/web/eu-vocabularies/eurostat

CBS taxonomy

Content language English

education > diploma

PREFERRED TERM **diploma**

BROADER CONCEPT education

NARROWER CONCEPTS bachelor degree diploma higher professional programme Master's degree professional qualifications propaedeutic certificate

IDENTIFIER 2468

IN OTHER LANGUAGES diploma

URI <https://taxonomie.cbs.nl/vocab/2468>

DOWNLOAD THIS CONCEPT: RDF/XML TURTLE JSON-LD CSV

Alphabetical Hierarchy New

- agriculture and fisheries
- businesses
- CBS corporate terms
- construction and housing
- education
 - classification of education
 - curriculum
 - denomination
 - diploma
 - bachelor degree
 - diploma higher professional education
 - continuation programme
 - master's degree
 - Master's degree
 - professional qualifications
 - propaedeutic certificate
 - education legislation
 - education policy
 - education programme
 - education statistics
 - educational attainment level
 - educational institutions
 - government expenditure on education
 - programme stage
 - student population
 - student progress

vocabs.cbs.nl/taxonomie

Scottish Government
Riaghaidh na h-Alba
gov.scot

STATISTICS.GOV.SCOT

ATLAS DATA

Explore Tools

developer documentation

Linked Data Vocabularies

For brevity, this list does not include third-party vocabularies. To view an exhaustive list of all vocabularies in use on statistics.gov.scot including third-party vocabularies, please see the [linked data vocabularies](#) page.

DATA

CONCEPT SCHEMES

A concept scheme is a collection of concepts about a topic. A concept scheme often acts as a list of possible values for a certain property of a resource (i.e. possible objects for RDF triples).

- 4-fold RESEAS classification
- 6-fold urban/rural classification
- Access to Car
- Accident Status
- Accommodation Type and Occupancy
- Admission Type
- Admissions, Discharges and Deaths
- Adult Disability Payment Indicators
- Affordable Housing Supply Programme Category

statistics.gov.scot

Official statistics and linked data (2)

- Statistical metadata as LD helps modeling *changes in metadata & linkability*

Geographical area Netherlands and country schemes

Content language: English

Alphabetical | Hierarchy | Groups

There is no term for this concept in this language.

NUTS 3 Areas 2021
Countries and areas
Geographic areas Netherlands
Netherlands
NUTS 1 areas
Provinces
Regional classifications
NUTS 3 areas
CR14 Achterhoek
CR26 Agglomeratie 's-Gravenhage
CR21 Agglomeratie Haarlem
CR25 Agglomeratie Leiden en Bollenstreek
CR19 Alkmaar en omgeving
CR15 Arnhem/Nijmegen
CR27 Delft en Westland
CR02 Delfzijl en omgeving
CR40 Flevoland
CR23 Groot-Amsterdam
CR29 Groot-Rijnmond
CR24 Het Gooi en Vechtstreek
CR20 IJmond
CR18 Kop van Noord-Holland
CR38 Midden-Limburg
CR34 Midden-Noord-Brabant
CR07 Noord-Drenthe
CR04 Noord-Friesland
CR37 Noord-Limburg
CR10 Noord-Overijssel
CR35 Noordoost-Noord-Brabant
CR01 Oost-Groningen
CR28 Oost-Zuid-Holland
CR03 Overig Groningen
CR32 Overig Zeeland
GM0654 Borsele
WK065413 's-Gravenpolder
WK065414 's-Heer Abtskerke
WK065415 's-Heerenhoek
WK065401 Baarland
WK065402 Borssele
WK065403 Driewegen wijk (Borsele)

PREFERRED TERM: GM0654 Borsele (nl)

TYPE: <http://rdf.histograph.io/Municipality>
<https://schema.org/AdministrativeArea>

BROADER CONCEPT: CR32 Overig Zeeland (nl)
PV29 Zeeland (nl)

NARROWER CONCEPTS: WK065401 Baarland (nl)
WK065402 Borssele (nl)
WK065403 Driewegen wijk (Borsele) (nl)
WK065404 Ellewoutsdijk (nl)
WK065405 Heinkenszand (nl)
WK065406 Hoedekenskerke (nl)
WK065407 Kwadendamme (nl)
WK065408 Lewedorp (nl)
WK065409 Nieuwdorp wijk (Borsele) (nl)
WK065410 Nisse (nl)
WK065411 Oudelande (nl)
WK065412 Ovezande (nl)
WK065413 's-Gravenpolder (nl)
WK065414 's-Heer Abtskerke (nl)
WK065415 's-Heerenhoek (nl)

REPLACES: GM1254 Borssele (nl)

TEMPORAL COVERAGE: 1970-01-01/..

START DATE: 1970-01-01

NOTATION: GM0654

Linking within ESS

ShowVoc

ESTAT_Nomenclature_of_Territorial_Units_for_Statistics

Concept | Collection | Scheme | Property | Alignments

- LU Luxembourg
- LV Latvija
- ME Црна Гора
- MK Северна Македонија
- MT Malta
- NL Nederland
 - NL1 Noord-Nederland
 - NL2 Oost-Nederland

WIKIDATA

Item Discussion | Read | View history | Set

Wikidata is turning 10 in October 2022, and the community is organizing plenty of decentralized birthday events all around the world. You can join one of them or organize your own!

Borsele (Q10071)

municipality in the Netherlands

In more languages

Language	Label	Description
English	Borsele	municipality in the Netherlands
Dutch	Borsele	gemeente in Zeeland
German	Borsele	Gemeinde in den Niederlanden
French	Borsele	commune de Zélande, Pays-Bas

All entered languages

Statements

instance of: municipality of the Netherlands

start time: 1 January 1970

+ 2 references

Linking outside ESS

Reflections on linked data (LD) in offstats

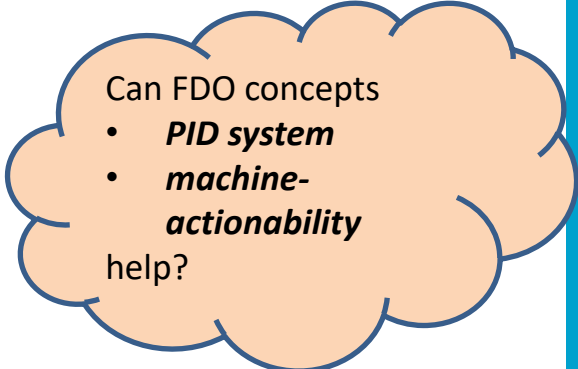
- LD is very well suited to *link* statistical metadata *across ESS* and to *wider communities*
- Scottish LD shows how every *statistical fact* can be *identifiable* and *linkable*
- Dutch LD shows modeling *complex statistical metadata dependencies*, a complete historical graph of Dutch geo-changes on municipality level for over two centuries
- Challenges:
 - LD can be perceived *complex* to end-users => need for *data stories*
 - *Flexibility* of LD model gives room for different implementation choices among ESS => need for *harmonization*
- Food for thought: with the current *AI training data hunger*: could statistical LD content more easily be picked up than ESS-specific standards?



How FAIR is this?



open	Access	Digital objects	Standards
		 News, charts, thematic	Schema.org DOI Dublin Core (DC)
		 Metadata / registries	SDMX (STAT-) DCAT SIMS Linked Data SPARQL SKOS Schema.org RDF XKOS
		 Statistical estimates	SDMX OData JSON-stat CSV JSON R Python GEO standards GeoJSON OGC
restricted		 microdata	DDI / DC Linked Data



> **A1: (Meta)data are retrievable by their identifier using a standardised communication protocol**



Wrap-up

- Official statistics software landscape:
 - Grows towards standardisation on SDMX, JSON-STAT and PX
 - Common features identified from awesome list: endpoint hiding, catalogue retrieval, search, caching, endpoint / local cartographic queries, registry access
 - Weakness on interoperability and reusability: No 'one-for-all' software that provides access to all offstats data => Can we do this
- Linked data:
 - Good examples at CBS, Eurostat, Scotland and more...
 - Flexible, linking metadata within and outside ESS, modeling complex metadata dependencies
 - Need for harmonisation and data stories
- FDOs:
 - Openness is not enough, FAIRness should hold on every layer for every digital object
 - Minimize variety in standards and identifiers and add machine-actionability

To be (more) FAIR there is **work to do on all perspectives**. Only then we can continue our role of trusted data partner in a future data-intense AI-aware society.

