SDMX in a FAIR official statistics landscape

Statistics Netherlands

Olav ten Bosch, Edwin de Jonge, Henk Laloli

12th SDMX Experts workshop, 7-11 Oct. Amsterdam

uRos23

Access to official statistics from R: an overview

Statistics Netherlands

Olav ten Bosch, Edwin de Jonge
uRos2023 12-14 December 2023

COSMOS24

To be FAIR, what is missing in Official Statistics?

Statistics Netherlands

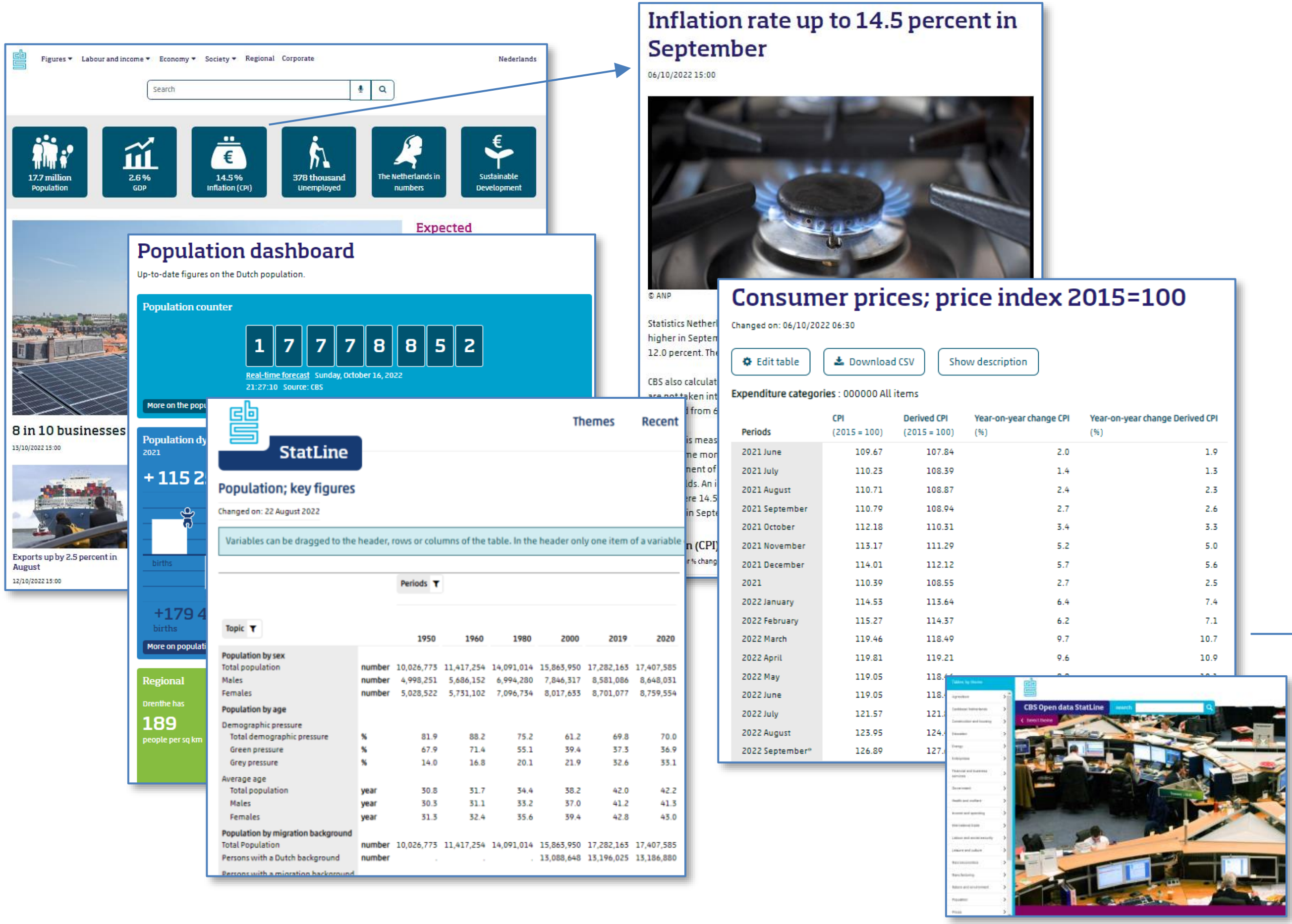
Olav ten Bosch, Edwin de Jonge, Henk Laloli
COSMOS 2024, 11-12 April Paris

CC BY 4.0

Contents

- The official statistics landscape
- Open source software for access; awesome list
- Features commonly offered: towards FAIRness
- And what about linked data?
- Wrap-up

Statistics Netherlands output



Inflation rate up to 14.5 percent in September

06/10/2022 15:00

Statistics Nether
higher in Septem
12.0 percent. The

CBS also calculat
are not taken int
from 6

is meas
me mor
nent of
lds. An i
re 14.5
in Septe

n (CPI)
r % chang

Population dashboard

Up-to-date figures on the Dutch population.

Population counter

1 7 7 7 8 8 5 2

Real-time forecast Sunday, October 16, 2022
21:27:10 Source: CBS

StatLine

Population; key figures

Changed on: 22 August 2022

Variables can be dragged to the header, rows or columns of the table. In the header only one item of a variable

Periods

Topic

Topic		1950	1960	1980	2000	2019	2020
Population by sex							
Total population	number	10,026,773	11,417,254	14,091,014	15,863,950	17,282,163	17,407,585
Males	number	4,998,251	5,686,152	6,994,280	7,846,317	8,581,086	8,648,031
Females	number	5,028,522	5,731,102	7,096,734	8,017,633	8,701,077	8,759,554
Population by age							
Demographic pressure							
Total demographic pressure	%	81.9	88.2	75.2	61.2	69.8	70.0
Green pressure	%	67.9	71.4	55.1	39.4	37.3	36.9
Grey pressure	%	14.0	16.8	20.1	21.9	32.6	33.1
Average age							
Total population	year	30.8	31.7	34.4	38.2	42.0	42.2
Males	year	30.3	31.1	33.2	37.0	41.2	41.3
Females	year	31.3	32.4	35.6	39.4	42.8	43.0
Population by migration background							
Total Population	number	10,026,773	11,417,254	14,091,014	15,863,950	17,282,163	17,407,585
Persons with a Dutch background	number	.	.	.	13,088,648	13,196,025	13,186,880
Persons with a migration background							

Consumer prices; price index 2015=100

Changed on: 06/10/2022 06:30

Edit table | Download CSV | Show description

Expenditure categories : 000000 All items

Periods	CPI (2015 = 100)	Derived CPI (2015 = 100)	Year-on-year change CPI (%)	Year-on-year change Derived CPI (%)
2021 June	109.67	107.84	2.0	1.9
2021 July	110.23	108.39	1.4	1.3
2021 August	110.71	108.87	2.4	2.3
2021 September	110.79	108.94	2.7	2.6
2021 October	112.18	110.31	3.4	3.3
2021 November	113.17	111.29	5.2	5.0
2021 December	114.01	112.12	5.7	5.6
2021	110.39	108.55	2.7	2.5
2022 January	114.53	113.64	6.4	7.4
2022 February	115.27	114.37	6.2	7.1
2022 March	119.46	118.49	9.7	10.7
2022 April	119.81	119.21	9.6	10.9
2022 May	119.05	118.4	[illegible]	[illegible]
2022 June	119.05	118.4	[illegible]	[illegible]
2022 July	121.57	121.5	[illegible]	[illegible]
2022 August	123.95	124.4	[illegible]	[illegible]
2022 September*	126.89	127.	[illegible]	[illegible]

Layered Open Data architecture

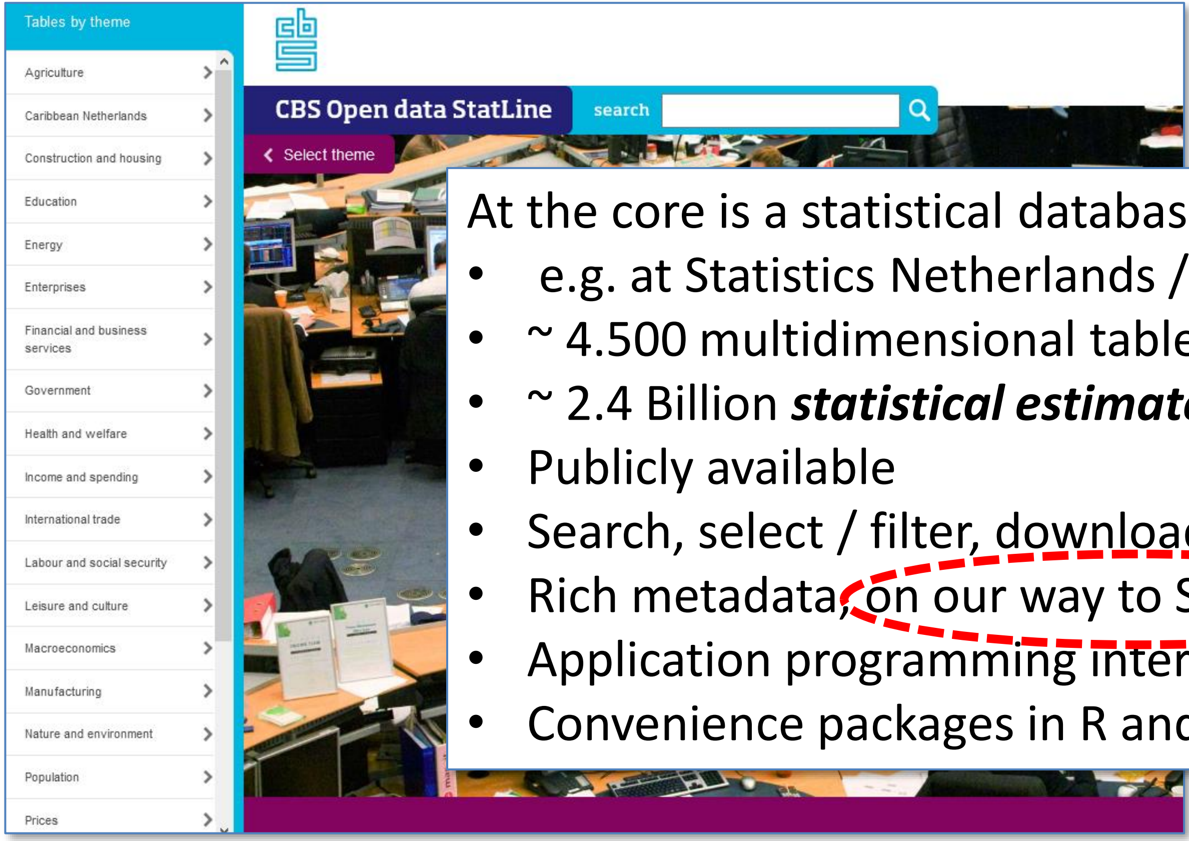
News articles

Thematic pages & visualisations

Selected data slices

Statistical Estimates

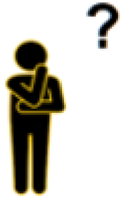
Statistical dissemination database



At the core is a statistical database:

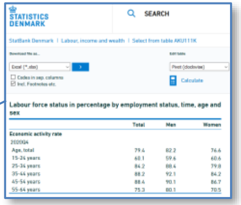
- e.g. at Statistics Netherlands / CBS:
- ~ 4.500 multidimensional tables
- ~ 2.4 Billion ***statistical estimates*** (active tables)
- Publicly available
- Search, select / filter, download
- Rich metadata, on our way to SDMX
- Application programming interface (API): Odata
- Convenience packages in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

Official Statistics landscape (OSL)



User

OSL = collective output of all official statistics organizations



STATISTICS DENMARK

SEARCH

StatBank Denmark | Labour, income and wealth | Select from table AKU111K

Download file as...

Excel (*.xlsx)

□ Codes in sep. columns

☑ Incl. Footnotes etc.

Edit table

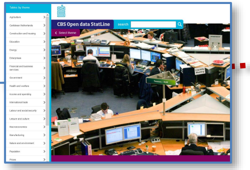
Pivot (clockwise)

Calculate

Labour force status in percentage by employment status, time, age and sex

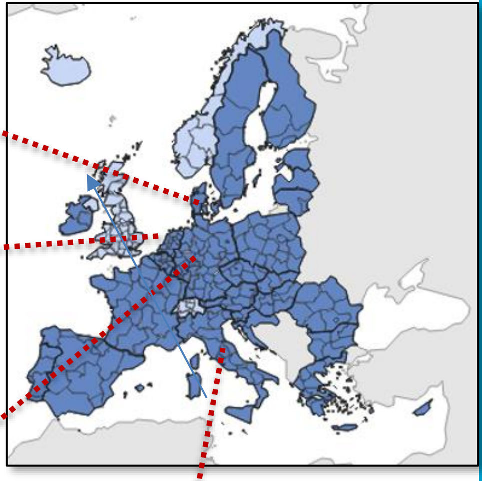
	Total	Men	Women
Economic activity rate			
2020Q4			
Age, total	79.4	82.2	76.6
15-24 years	60.1	59.6	60.6
25-34 years	84.2	88.4	79.8
35-44 years	88.2	92.1	84.2
45-54 years	88.4	90.1	86.7
55-64 years	75.3	80.1	70.5

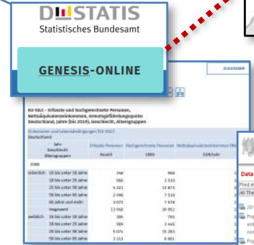
DK



Tables by theme

NL





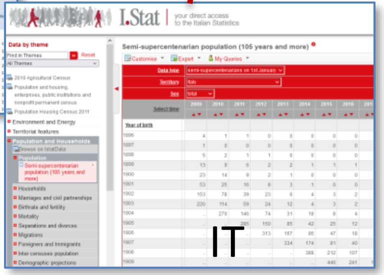
DESTATIS
Statistisches Bundesamt

GENESIS-ONLINE

EU-SILC - Erfasste und hochgerechnete Personen, Nettoäquivalenzeinkommen, Armutsgefährdungsquote Deutschland, Jahre (bis 2019), Geschlecht, Altersgruppen

Jahre Geschlecht Altersgruppen		Erfasste Personen	Hochgerechnete Personen	[illegible]
		Anzahl	1000	EUR/Jahr
2008				
männlich	16 bis unter 18 Jahre	348	968	[illegible]
	18 bis unter 25 Jahre	950	3 510	[illegible]
	25 bis unter 50 Jahre	4 321	13 873	[illegible]
	50 bis unter 65 Jahre	2 996	7 518	[illegible]
	65 Jahre und mehr	3 072	7 578	[illegible]
	Insgesamt	13 940	39 952	[illegible]
weiblich	16 bis unter 18 Jahre	306	766	[illegible]
	18 bis unter 25 Jahre	969	3 445	[illegible]
	25 bis unter 50 Jahre	5 074	15 283	[illegible]
	50 bis unter 65 Jahre	3 313	8 001	[illegible]

DE



I.Stat | your direct access to the Italian Statistics

Data by theme

Find in Themes
Reset
All Themes
2010 Agricultural Census
Population and housing, enterprises, public institutions and nonprofit permanent census
Population Housing Census 2011
Environment and Energy
Territorial features
Population and Households
Browse on IstatData
Population
Semi-supercentenarian population (105 years and more)
Households
Marriages and civil partnerships
Birthrate and fertility
Mortality
Separations and divorces
Migrations
Foreigners and Immigrants
Inter censuses population
Demographic projections
Customise
Export
My Queries

Semi-supercentenarian population (105 years and more)

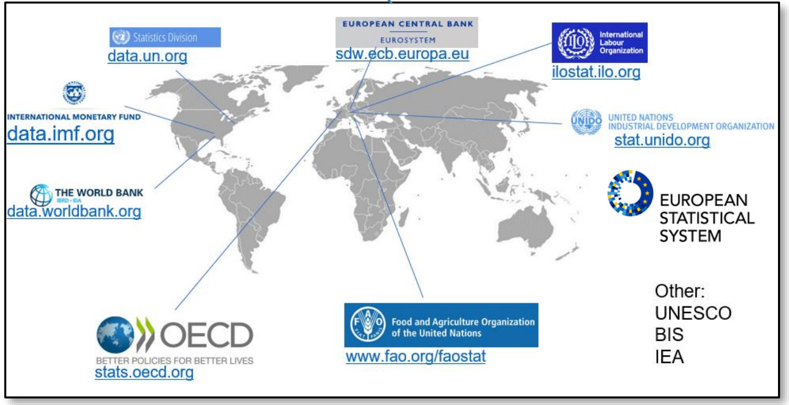
Data type: semi-supercentenarians on 1st January

Territory: Italy

Sex: total

Select time	2009	2010	2011	2012	2013	2014	2015	2016
Year of birth								
1896	4	1	1	0	0	0	0	0
1897	1	0	0	0	0	0	0	0
1898	5	2	1	1	0	0	0	0
1899	13	9	6	2	2	1	1	1
1900	23	14	9	2	1	0	0	0
1901	53	25	16	8	3	1	0	0
1902	153	78	39	23	6	4	3	2
1903	220	114	59	24	12	4	3	2
1904	..	279	146	74	31	18	8	4
1905	..	..	285	150	85	42	25	12
1906	..	..	..	313	167	95	47	18
1907	..	..	..	..	334	174	91	40
1908	..	..	..	..	..	388	212	107
1909	..	..	..	..	..	..	445	241

IT



Statistics Division
data.un.org

EUROPEAN CENTRAL BANK
EUROSYSTEM
sdw.ecb.europa.eu

International Labour Organization
ilostat.ilo.org

INTERNATIONAL MONETARY FUND
data.imf.org

UNITED NATIONS INDUSTRIAL DEVELOPMENT ORGANIZATION
stat.unido.org

THE WORLD BANK
IBRD · IDA
data.worldbank.org

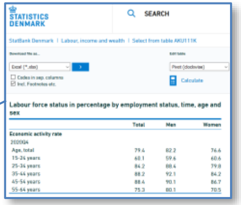
EUROPEAN STATISTICAL SYSTEM

OECD
BETTER POLICIES FOR BETTER LIVES
stats.oecd.org

Food and Agriculture Organization of the United Nations
www.fao.org/faostat

Other:
UNESCO
BIS
IEA

Official Statistics landscape (OSL)



STATISTICS DENMARK

SEARCH

StatBank Denmark | Labour, income and wealth | Select from table AKU111K

Download file as...

Excel (*.xlsx)

□ Codes in sep. columns
☑ Incl. Footnotes etc.

Edit table

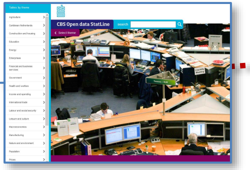
Pivot (clockwise)

Calculate

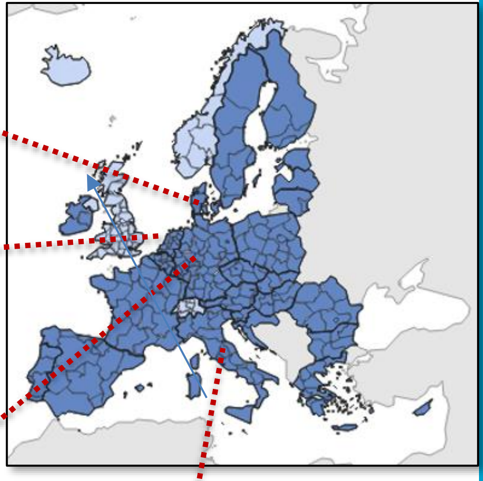
Labour force status in percentage by employment status, time, age and sex

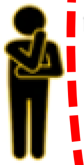
	Total	Men	Women
Economic activity rate			
2020Q4			
Age, total	79.4	82.2	76.6
15-24 years	60.1	59.6	60.6
25-34 years	84.2	88.4	79.8
35-44 years	88.2	92.1	84.2
45-54 years	88.4	90.1	86.7
55-64 years	75.3	80.1	70.5

DK



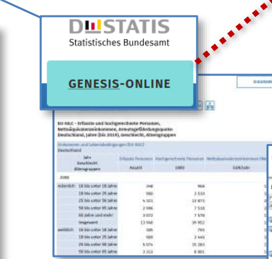
NL





Happy user?

OSL = collective output of all official statistics organizations

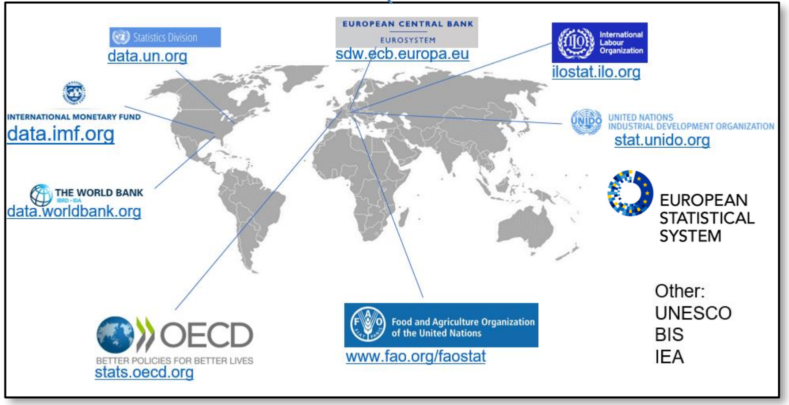


DESTATIS
Statistisches Bundesamt

GENESIS-ONLINE

EU-SILC - Erfasste und hochgerechnete Personen, Nettoäquivalenzeinkommen, Armutsgefährdungsquote
Deutschland, Jahre (bis 2019), Geschlecht, Altersgruppen

DE



Statistics Division
data.un.org

EUROPEAN CENTRAL BANK
EUROSYSTEM
sdw.ecb.europa.eu

International Labour Organization
ilostat.ilo.org

INTERNATIONAL MONETARY FUND
data.imf.org

UNIDO UNITED NATIONS INDUSTRIAL DEVELOPMENT ORGANIZATION
stat.unido.org

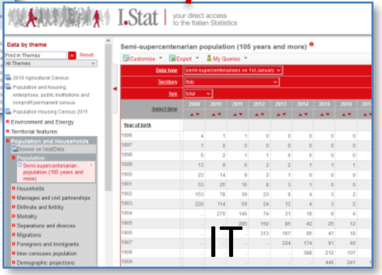
THE WORLD BANK
IBRD · IDA
data.worldbank.org

EUROPEAN STATISTICAL SYSTEM

OECD
BETTER POLICIES FOR BETTER LIVES
stats.oecd.org

Food and Agriculture Organization of the United Nations
www.fao.org/faostat

Other:
UNESCO
BIS
IEA



I.Stat | your direct access to the Italian Statistics

Data by theme

Find in Themes | Reset

All Themes

- 2010 Agricultural Census
- Population and housing, enterprises, public institutions and nonprofit permanent census
- Population Housing Census 2011
- Environment and Energy
- Territorial features
- Population and Households
  - Browse on IstatData
  - Population
    - Semi-supercentenarian population (105 years and more)
  - Households
  - Marriages and civil partnerships
  - Birthrate and fertility
  - Mortality
  - Separations and divorces
  - Migrations
  - Foreigners and Immigrants
  - Inter censuses population
  - Demographic projections

Semi-supercentenarian population (105 years and more)

Customise | Export | My Queries

Data type: semi-supercentenarians on 1st January

Territory: Italy

Sex: total

Select time	2009	2010	2011	2012	2013	2014	2015	2016
Year of birth								
1896	4	1	1	0	0	0	0	0
1897	1	0	0	0	0	0	0	0
1898	5	2	1	1	0	0	0	0
1899	13	8	6	2	2	1	1	1
1900	23	14	9	2	1	0	0	0
1901	53	25	16	8	3	1	0	0
1902	153	78	39	23	6	4	3	2
1903	220	114	69	24	12	4	3	2
1904	..	279	146	74	31	18	8	4
1905	..	..	285	150	85	42	25	12
1906	..	..	..	313	167	95	47	18
1907	..	..	..	..	334	174	91	40
1908	..	..	..	..	..	388	212	107
1909	..	..	..	..	..	..	445	241

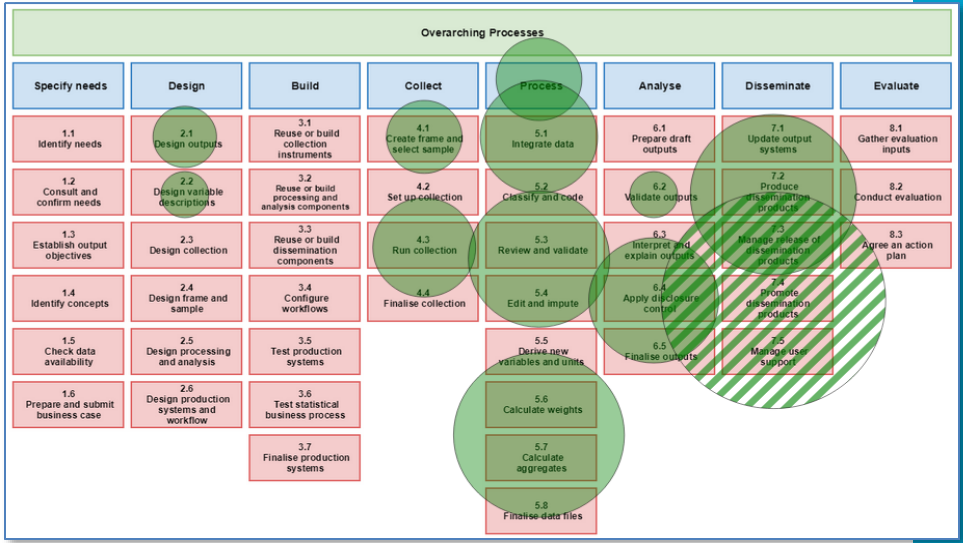
IT

Awesome list of statistical software

uRos17-23

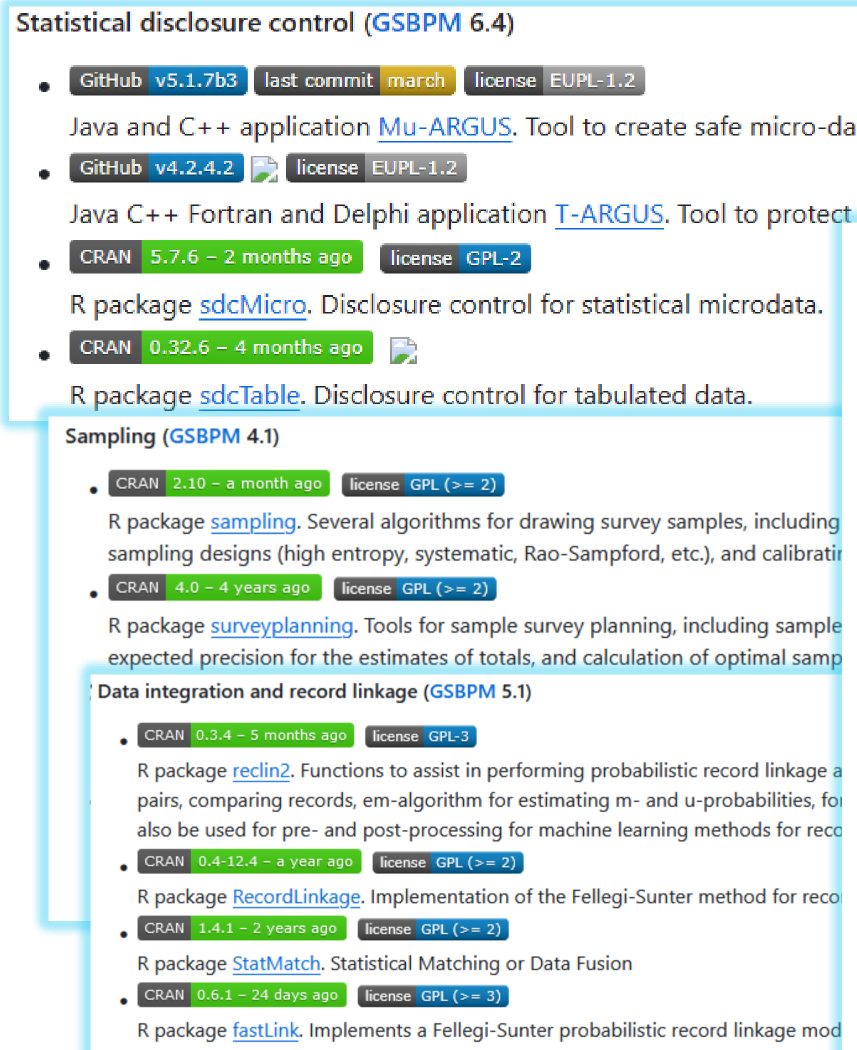
- A ***community approach*** to knowledge management
- To ***collectively remember useful software*** in official statistics
- Started UNECE SDE 2017
- Maintained by ***statistical community***
- How:
  - Using ***awesome concept***
  - ***awesomeofficialstatistics.org***
  - ***Largest category: "Access to official statistics"***

# awesome packages by GSBPM



Overarching Processes

Specify needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Reuse or build collection instruments	4.1 Create frame and select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult and confirm needs	2.2 Design variable descriptions	3.2 Reuse or build processing and analysis components	4.2 Set up collection	5.2 Classify and code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collection	3.3 Reuse or build dissemination components	4.3 Run collection	5.3 Review and validate	6.3 Interpret and explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
1.4 Identify concepts	2.4 Design frame and sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit and impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing and analysis	3.5 Test production systems		5.5 Derive new variables and units	6.5 Finalise outputs	7.5 Manage user support	
1.6 Prepare and submit business case	2.6 Design production systems and workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production systems		5.7 Calculate aggregates			
				5.8 Finalise data files			



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-data files. S
- GitHub v4.2.4.2 license EUPL-1.2

  Java C++ Fortran and Delphi application T-ARGUS. Tool to protect statistica
- 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 calibratir
- 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 samp

Data integration and record linkage (GSBPM 5.1)

- CRAN 0.3.4 – 5 months ago license GPL-3

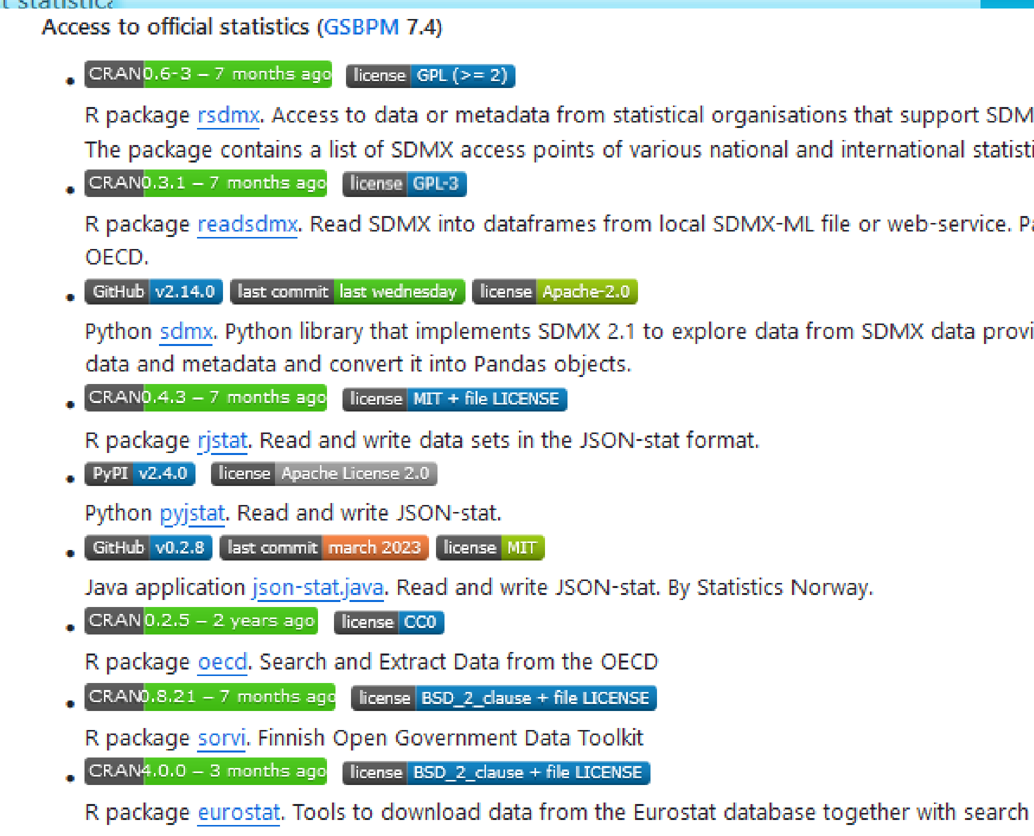
  R package reclin2. Functions to assist in performing probabilistic record linkage a pairs, comparing records, em-algorithm for estimating m- and u-probabilities, fo also be used for pre- and post-processing for machine learning methods for reco
- CRAN 0.4-12.4 – a year ago license GPL (>= 2)

  R package RecordLinkage. Implementation of the Fellegi-Sunter method for reco
- 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 mod

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



Access to official statistics (GSBPM 7.4)

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

  R package rsdmx. Access to data or metadata from statistical organisations that support SDM 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

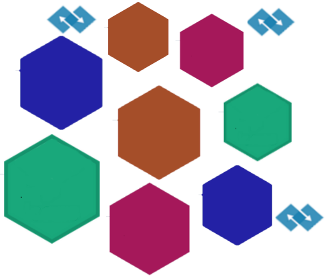
  R package sorvi. Finnish Open Government Data Toolkit
- CRAN 4.0.0 – 3 months ago license BSD_2_clause + file LICENSE

  R package eurostat. Tools to download data from the Eurostat database together with search

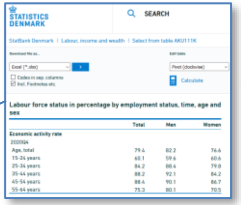
Software layer surrounding OSL

> 30 software packages
> 60 dataproviders
> majority R-packages
> Some use SDMX

Data user



OSL



STATISTICS DENMARK

SEARCH

StatBank Denmark | Labour, income and wealth | Select from table AKU111K

Download file as... Excel (*.xlsx)

□ Codes in sep. columns
☑ Incl. Footnotes etc.

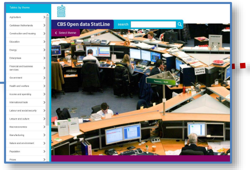
Edit table: Pivot (clockwise)

Calculate

Labour force status in percentage by employment status, time, age and sex

	Total	Men	Women
Economic activity rate			
2020Q4			
Age, total	79.4	82.2	76.6
15-24 years	60.1	59.6	60.6
25-34 years	84.2	88.4	79.8
35-44 years	88.2	92.1	84.2
45-54 years	88.4	90.1	86.7
55-64 years	75.3	80.1	70.5

DK



Tables by theme

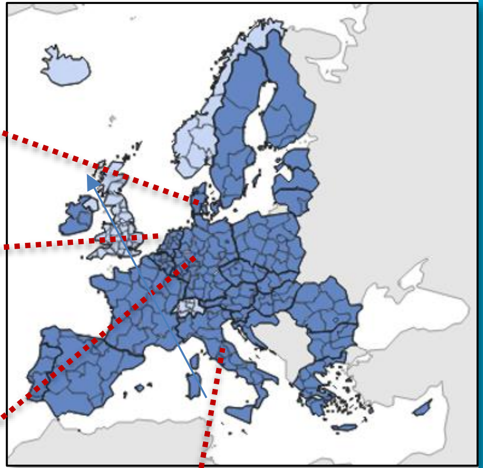
- Agriculture
- Caribbean Netherlands
- Construction and housing
- Education
- Energy
- Enterprises
- Financial and business services
- Government
- Health and welfare
- Income and spending
- International trade
- Labour and social security
- Leisure and culture
- Macroeconomics
- Manufacturing
- Nature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

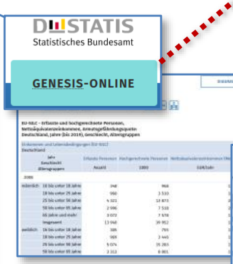
CBS Open data StatLine

search

Select theme

NL





DESTATIS

Statistisches Bundesamt

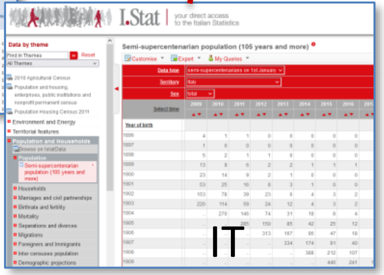
GENESIS-ONLINE

EU-SILC - Erfasste und hochgerechnete Personen, Nettoäquivalenzeinkommen, Armutsgefährdungsquote: Deutschland, Jahre (bis 2019), Geschlecht, Altersgruppen

[illegible]
Deutschland

Jahr Geschlecht Altersgruppen		Erfasste Personen	Hochgerechnete Personen	Nettoäquivalenzeinkommen [illegible]
		Anzahl	1000	EUR/Jahr
2005				
männlich	16 bis unter 18 Jahre	348	968	[illegible]
	18 bis unter 25 Jahre	950	3 510	[illegible]
	25 bis unter 50 Jahre	4 321	13 873	[illegible]
	50 bis unter 65 Jahre	2 996	7 518	[illegible]
	65 Jahre und mehr	3 072	7 578	[illegible]
	Insgesamt	13 940	39 952	[illegible]
weiblich	16 bis unter 18 Jahre	305	795	[illegible]
	18 bis unter 25 Jahre	969	3 440	[illegible]
	25 bis unter 50 Jahre	5 076	15 265	[illegible]
	50 bis unter 65 Jahre	3 313	8 001	[illegible]

DE



I.Stat | your direct access to the Italian Statistics

Data by theme

Find in Themes | Reset

All Themes

- 2010 Agricultural Census
- Population and housing, enterprises, public institutions and nonprofit permanent census
- Population Housing Census 2011
- Environment and Energy
- Territorial features
- Population and Households
  - Browse on IstatData
  - Population
    - Semi-supercentenarian population (105 years and more)
  - Households
  - Marriages and civil partnerships
  - Birthrate and fertility
  - Mortality
  - Separations and divorces
  - Migrations
  - Foreigners and Immigrants
  - Inter censuses population
  - Demographic projections

Semi-supercentenarian population (105 years and more)

Customise | Export | My Queries

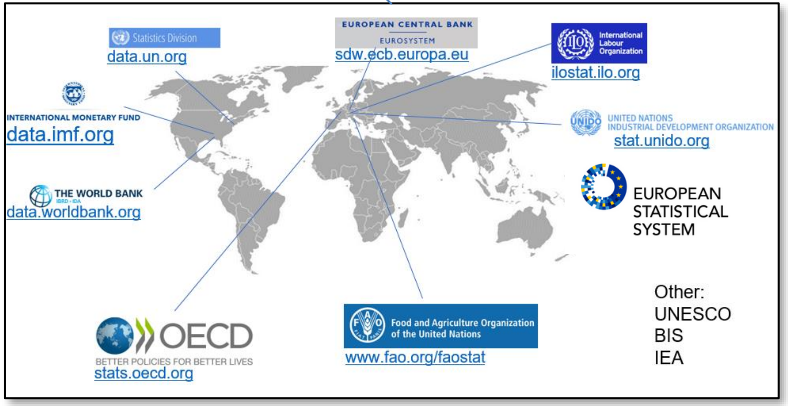
Data type: semi-supercentenarians on 1st January

Territory: Italy

Sex: total

Select time	2009	2010	2011	2012	2013	2014	2015	2016
Year of birth								
1896	4	1	1	0	0	0	0	0
1897	1	0	0	0	0	0	0	0
1898	5	2	1	1	0	0	0	0
1899	13	8	6	2	2	1	1	1
1900	23	14	9	2	1	0	0	0
1901	53	25	16	8	3	1	0	0
1902	153	78	39	23	6	4	3	2
1903	220	114	59	24	12	4	3	2
1904	..	279	146	74	31	18	8	4
1905	..	..	285	150	85	42	25	12
1906	..	..	..	313	187	95	47	18
1907	..	..	..	..	334	174	91	40
1908	..	..	..	..	..	388	212	107
1909	..	..	..	..	..	..	445	241

IT



Statistics Division
data.un.org

EUROPEAN CENTRAL BANK
EUROSYSTEM
sdw.ecb.europa.eu

International Labour Organization
ilostat.ilo.org

INTERNATIONAL MONETARY FUND
data.imf.org

UNIDO
UNITED NATIONS INDUSTRIAL DEVELOPMENT ORGANIZATION
stat.unido.org

THE WORLD BANK
IBRD · IDA
data.worldbank.org

EUROPEAN STATISTICAL SYSTEM

OECD
BETTER POLICIES FOR BETTER LIVES
stats.oecd.org

Food and Agriculture Organization of the United Nations
www.fao.org/faostat

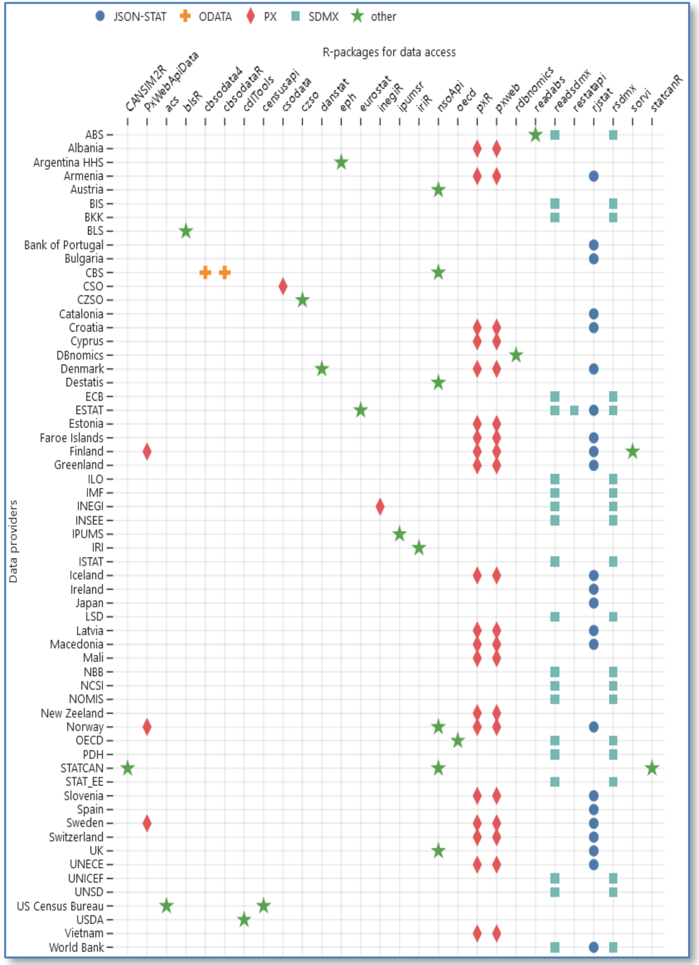
Other:
UNESCO
BIS
IEA

“access to official statistics” software landscape

- Matrix derived from documentation and links to web pages and packages execution:

packages (38) x data ● JSON-STAT ✚ ODATA ♦ PX ■ SDMX ★ other

- Standards:



https://observablehq.com/@olavtenbosch/access_to_official_statistics

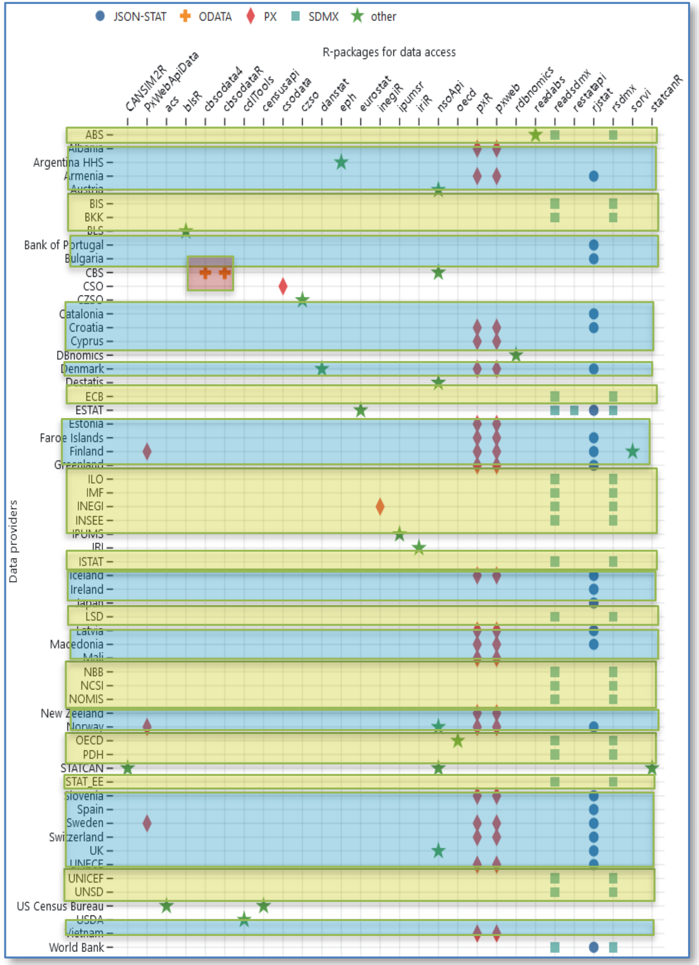
"access to official statistics" software landscape

- 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

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



https://observablehq.com/@olavtenbosch/access_to_official_statistics

“access to official statistics” software landscape

- Standards-oriented packages:

  *rsdmx, readsdmx, rjstat, px**

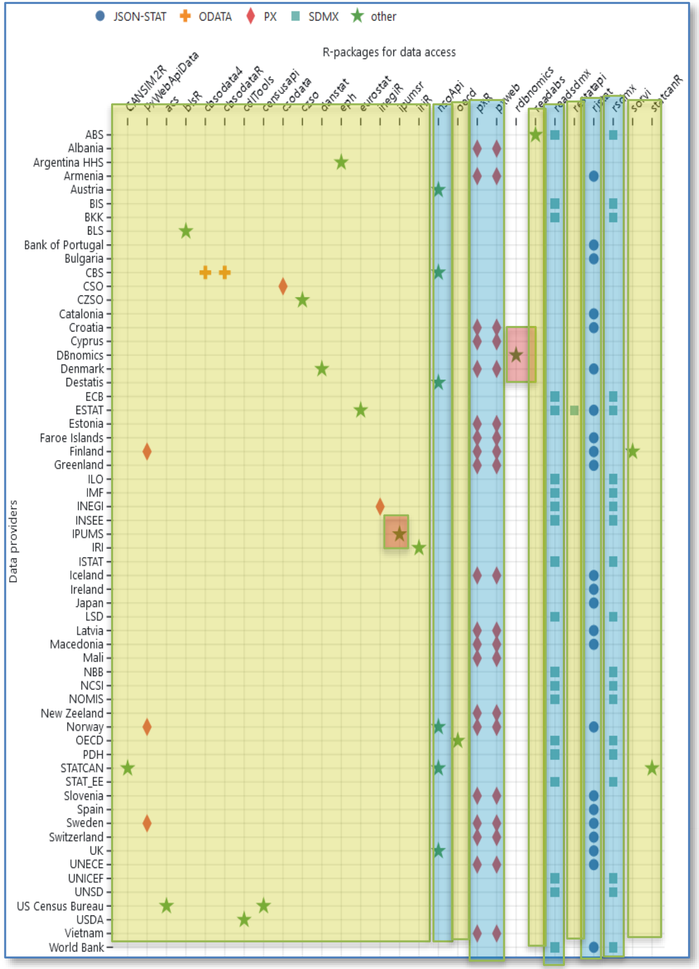
- Data provider-centric packages:

  inegiR, readabs, statcanR, eurostat

- Official statistics aggregator sites:

  rdbnomics: economic data

  ipumsr: census & survey data
  time&space harmonised

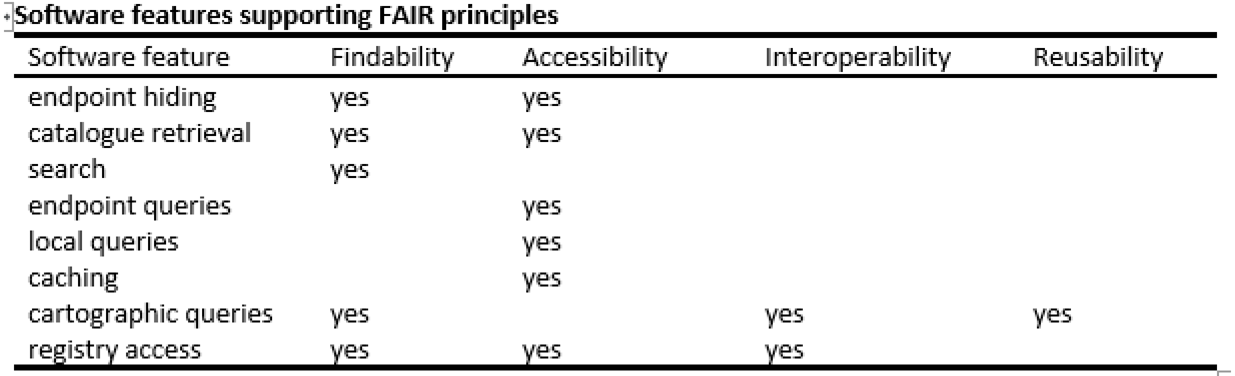


https://observablehq.com/@olavtenbosch/access_to_official_statistics

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?

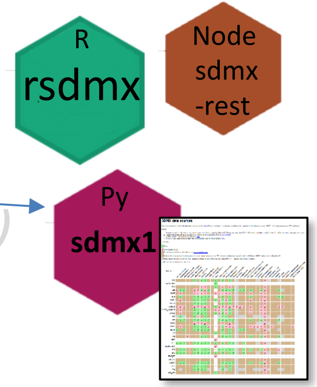


Software features supporting FAIR principles

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	

Features powerful in SDMX (1)

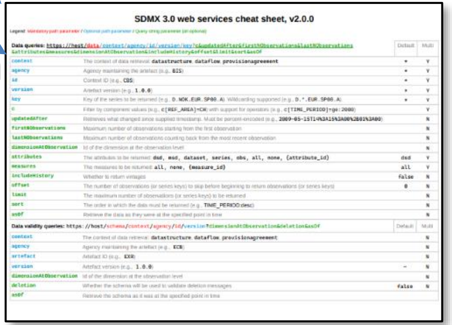
- ***endpoint hiding***: wrapping the preconfigured endpoint(s) in a function
- *catalogue retrieval: to list the availability datasets on the endpoint(s)*
- *search: to search for datasets or within datasets*
- *endpoint queries: query for subsets / slices on the endpoint(s) side*
- *local queries: 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*



khaeru.github.io/sdmx/

Features powerful in SDMX (2)

- ***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



SDMX 3.0 web services cheat sheet, v2.0.0

Legend: Mandatory path parameter / Optional path parameter / Query string parameter (all optional)

Data queries: https://host/data/context/agency/id/version/key?c&updatedAfter&firstNObservations&lastNObservations&attributes&measures&dimensionAtObservation&includeHistory&offset&limit&sort&asOf		Default	Multi
context	The context of data retrieval: datastructure, dataflow, provisionagreement	*	Y
agency	Agency maintaining the artefact (e.g., BIS)	*	Y
id	Context ID (e.g., CBS)	*	Y
version	Artefact version (e.g., 1.0.0)	*	Y
key	Key of the series to be returned (e.g., D.NOK.EUR.SP00.A). Wildcarding supported (e.g., D.*.EUR.SP00.A)	*	Y
c	Filter by component values (e.g., c[REF_AREA]=CH) with support for operators (e.g., c[TIME_PERIOD]=ge:2000)		Y
updatedAfter	Retrieves what changed since supplied timestamp. Must be percent-encoded (e.g., 2009-05-15T14%3A15%3A00%2B01%3A00)		N
firstNObservations	Maximum number of observations starting from the first observation		N
lastNObservations	Maximum number of observations counting back from the most recent observation		N
dimensionAtObservation	Id of the dimension at the observation level		N
attributes	The attributes to be returned: dsd, msd, dataset, series, obs, all, none, {attribute_id}	dsd	Y
measures	The measures to be returned: all, none, {measure_id}	all	Y
includeHistory	Whether to return vintages	false	N
offset	The number of observations (or series keys) to skip before beginning to return observations (or series keys)	0	N
limit	The maximum number of observations (or series keys) to be returned		N
sort	The order in which the data must be returned (e.g., TIME_PERIOD:desc)		N
asOf	Retrieve the data as they were at the specified point in time		N

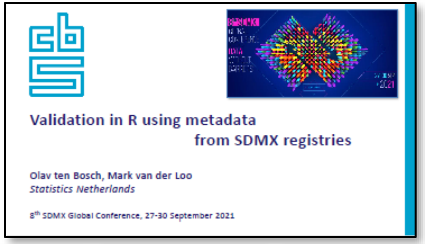
Data validity queries: https://host/schema/context/agency/id/version?dimensionAtObservation&deletion&asOf		Default	Multi
context	The context of data retrieval: datastructure, dataflow, provisionagreement		N
agency	Agency maintaining the artefact (e.g., ECB)		N
artefact	Artefact ID (e.g., EXR)		N
version	Artefact version (e.g., 1.0.0)	~	N
dimensionAtObservation	Id of the dimension at the observation level		N
deletion	Whether the schema will be used to validate deletion messages	false	N
asOf	Retrieve this schema as it was at the specified point in time		

Rich API &
cheat sheets

Features powerful in SDMX (3)

- ***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

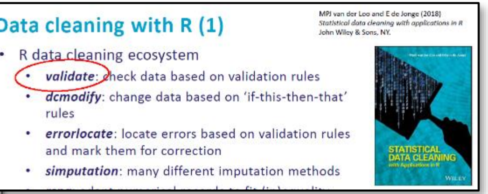
2021



Validation in R using metadata
from SDMX registries

Olav ten Bosch, Mark van der Loo
Statistics Netherlands

8th SDMX Global Conference, 27-30 September 2021



Data cleaning with R (1)

MPJ van der Loo and E de Jonge (2018)
Statistical data cleaning with applications in R
John Wiley & Sons, NY.

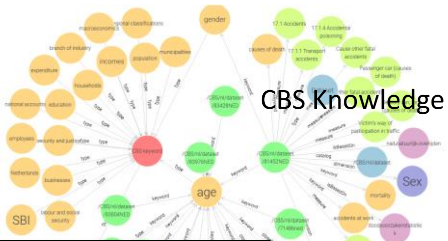
- R data cleaning ecosystem
  - ***validate***: check data based on validation rules
  - ***dcmodify***: change data based on 'if-this-then-that' rules
  - ***errorlocate***: locate errors based on validation rules and mark them for correction
  - ***simputation***: many different imputation methods

Validating data in R using SDMX registry

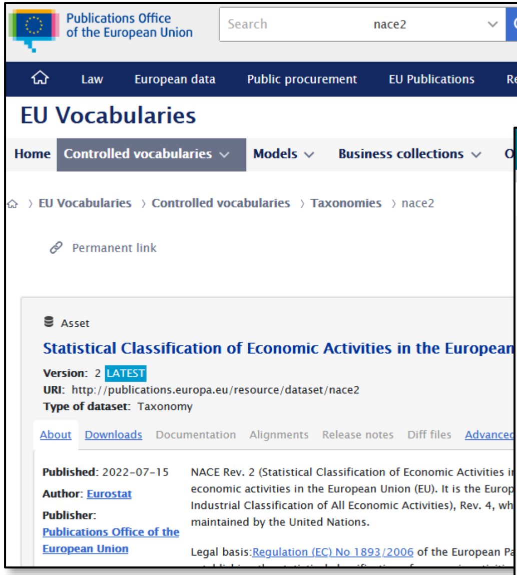
SDMX registries

And what about linked data? (1)

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



CBS Knowledge graph


Publications Office of the European Union

Search nace2

Law | European data | Public procurement | EU Publications | Research & Innovat

EU Vocabularies

Home | Controlled vocabularies | Models | Business collections | O

EU Vocabularies › Controlled vocabularies › Taxonomies › nace2

Permanent link

Asset

Statistical Classification of Economic Activities in the European

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

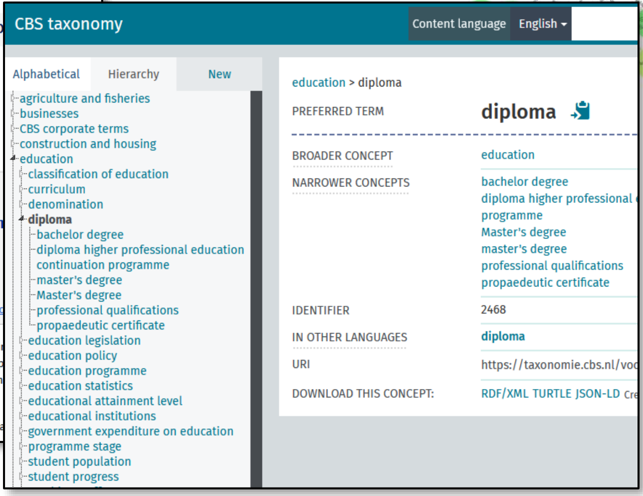
Author: Eurostat

Publisher: Publications Office of the European Union

NACE Rev. 2 (Statistical Classification of Economic Activities i economic activities in the European Union (EU). It is the Europ Industrial Classification of All Economic Activities), Rev. 4, wh maintained by the United Nations.

Legal basis: Regulation (EC) No 1893/2006 of the European Pa

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


CBS taxonomy | Content language | English

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

education > diploma

PREFERRED TERM: **diploma**

BROADER CONCEPT: education

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

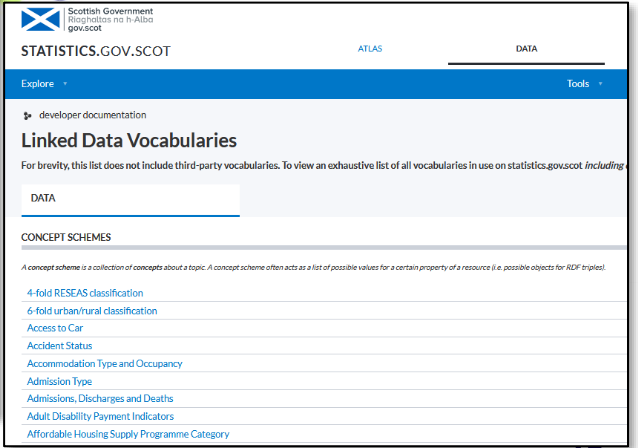
IDENTIFIER: 2468

IN OTHER LANGUAGES: diploma

URI: https://taxonomie.cbs.nl/voc

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

vocabs.cbs.nl/taxonomie


Scottish Government
Riaghaltas 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*

DATA

CONCEPT SCHEMES

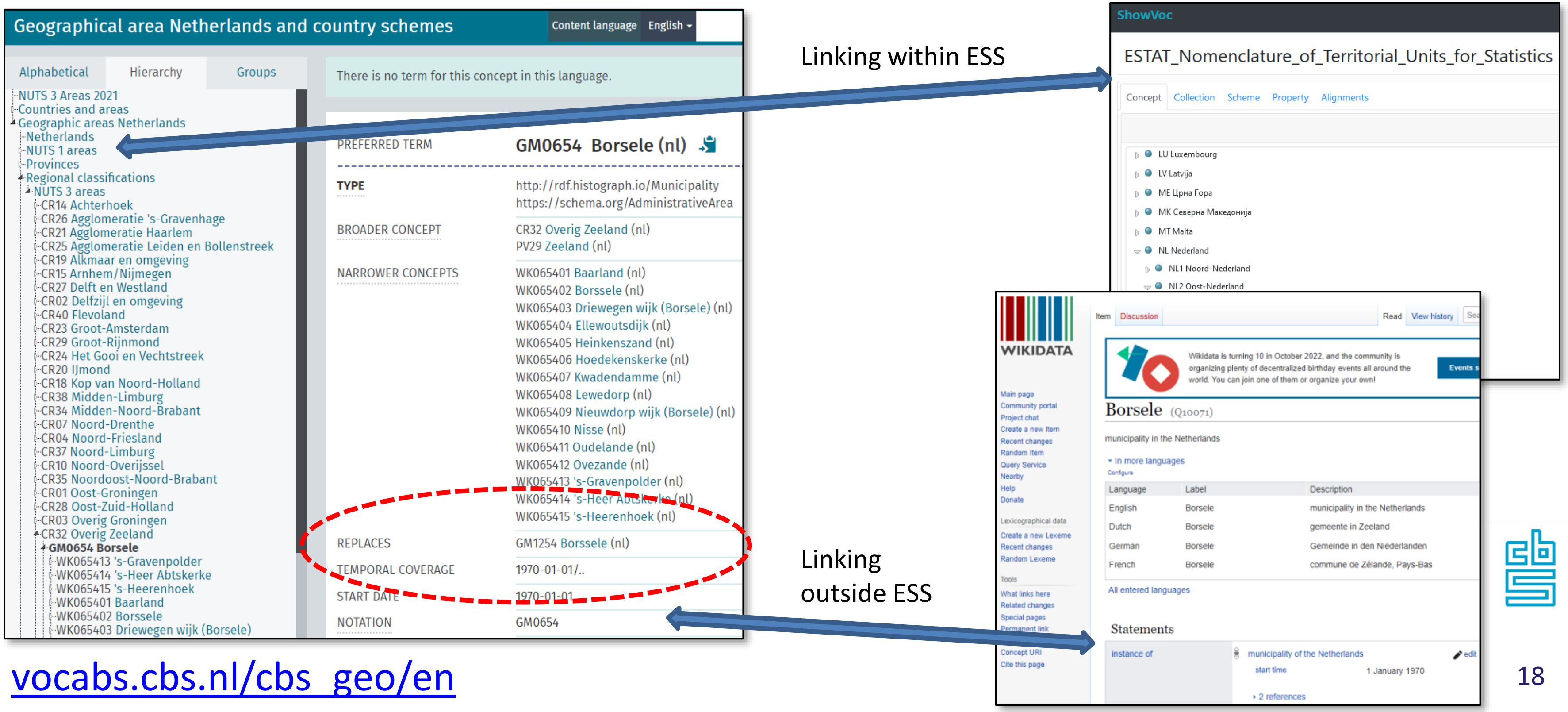
A concept scheme is a collection of concepts about a topic. A concept scheme often acts as a list of possible values for a certain property of a resource (i.e. possible objects for RDF triples).

- 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

And what about linked data? (2)

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



Linking within ESS

Linking outside ESS

Geographical area Netherlands and country schemes

- 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)

There is no term for this concept in this language.

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

ShowVoc

ESTAT_Nomenclature_of_Territorial_Units_for_Statistics

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

WIKIDATA

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

Configure

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	edit
	start time	1 January 1970
	▸ 2 references	

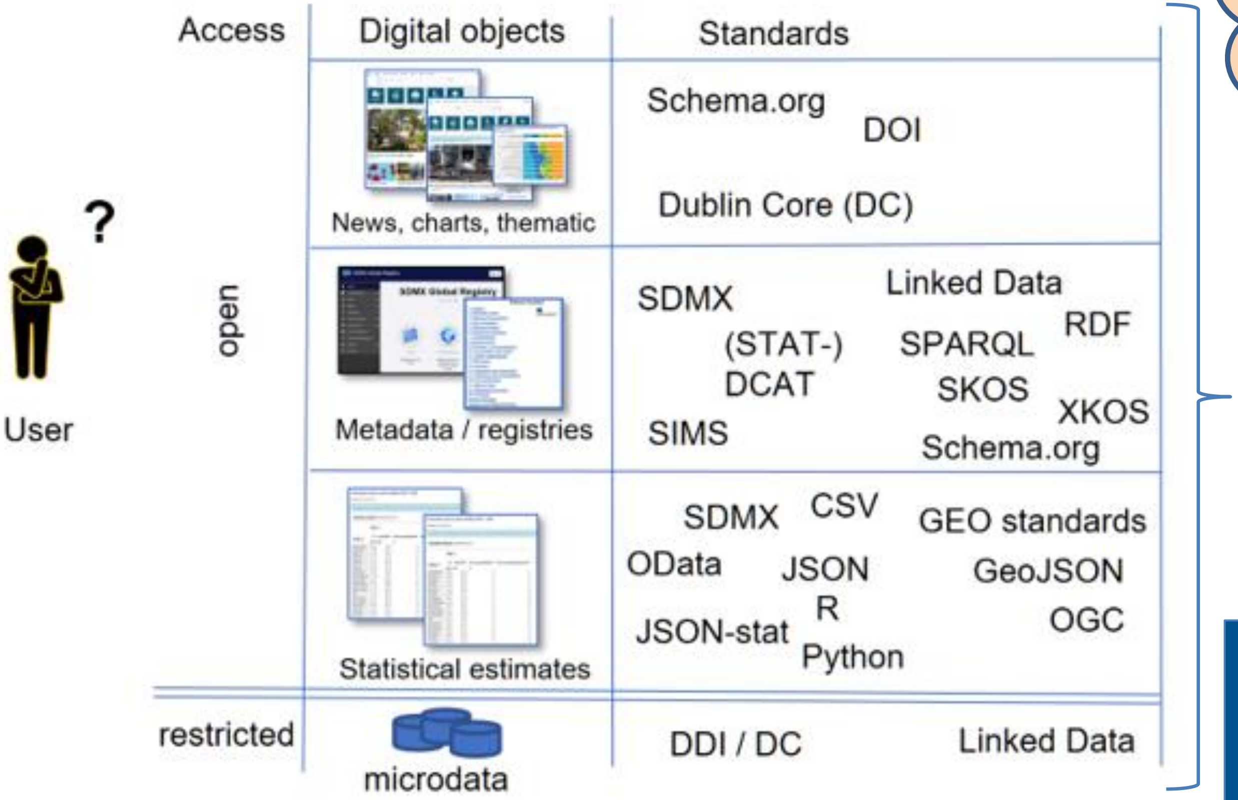
vocabs.cbs.nl/cbs_geo/en

Reflections on linked data (LD) in offstats

- LD is well suited to ***link across ESS*** and to ***non-ESS communities***
- LD suited to model ***complex statistical metadata dependencies***
  - Example: 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 implementations among ESS => need for ***harmonization***

- Food for thought: with the current ***AI training data hunger***: could statistical LD content more easily picked up than ESS-specific standards like SDMX?

Becoming (more) FAIR…

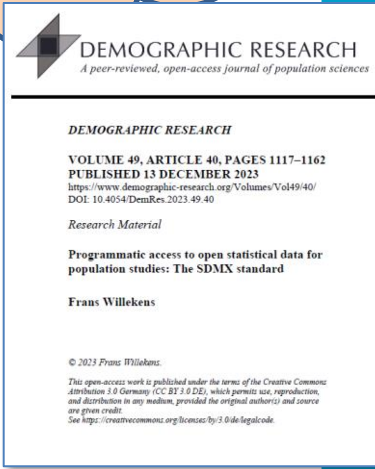


User

?

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

- ***# standards***
- ***Identifiability***
- ***Note the software layer!***
- ***Work with communities***



DEMOGRAPHIC RESEARCH

A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 49, ARTICLE 40, PAGES 1117–1162
PUBLISHED 13 DECEMBER 2023
https://www.demographic-research.org/Volumes/Vol49/40/
DOI: 10.4054/DemRes.2023.49.40

Research Material

Programmatic access to open statistical data for population studies: The SDMX standard

Frans Willekens

© 2023 Frans Willekens.

This open-access work is published under the terms of the Creative Commons Attribution 3.0 Germany (CC BY 3.0 DE), which permits use, reproduction, and distribution in any medium, provided the original author(s) and source are given credit.
See https://creativecommons.org/licenses/by/3.0/de/legalcode.

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

Wrap-up

- Official Statistics Landscape grows towards ***standardisation: SDMX, JSON-stat/PX***
- Data user can choose from over ***30 open source software packages***, each offering specific functionality to over ***60 data sources***
- Many NSIs offer ***targeted (R-)packages*** for data and metadata access
- ***Common features*** in software layer identifie
- ***SDMX*** scores well on 3 of the features
- There is ***no software package*** that provides access to ***all*** official statistics
- Dream:
  - ***one generic (R-)package*** to access all official statistics content from all dataproviders supporting maximum FAIRness*
  - offstats community and the open source / data science community have to work together to reach that goal

* FAIR: Findable, Accesible, Interoperable, Reusable

Feedback: Olav ten Bosch o.tenbosch@cbs.nl pubs

awesomeofficialstatistics.org Please