Access to official statistics from R: part II

Statistics Netherlands

Olav ten Bosch, Edwin de Jonge

uRos2024, Greece, 27-29 November 2024

Contents

- What is the awesome list?
  - History, concept, working
  - What happened from uRos2023 -> uRos 2024?

- Category “access to official statistics”
  - Packages, data providers, standards, ***features***
  - A “one for all” package?

- Wrap-up

What is the awesome list?

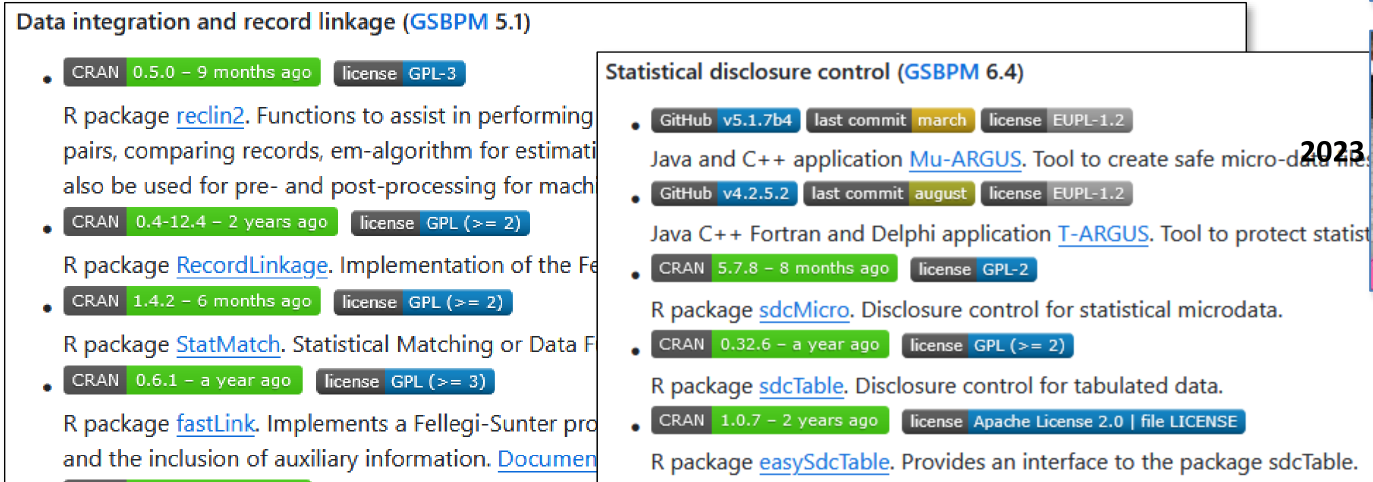
Awesome list of official statistics software

- Started at ***UNECE SDE conference*** 2017 (The Hague)
- a ***community approach*** to ***remember useful software***
- A ***public*** list, clear and simple ***criteria***
- Majority is ***R software***
- ***awesomeofficialstatistics.org***

2018

2019

2023



Data integration and record linkage (GSBPM 5.1)

- CRAN 0.5.0 – 9 months ago | license GPL-3

  R package reclin2. Functions to assist in performing
  pairs, comparing records, em-algorithm for estimati
  also be used for pre- and post-processing for mach
- CRAN 0.4-12.4 – 2 years ago | license GPL (>= 2)

  R package RecordLinkage. Implementation of the Fe
- CRAN 1.4.2 – 6 months ago | license GPL (>= 2)

  R package StatMatch. Statistical Matching or Data F
- CRAN 0.6.1 – a year ago | license GPL (>= 3)

  R package fastLink. Implements a Fellegi-Sunter pro
  and the inclusion of auxiliary information. Documen

Statistical disclosure control (GSBPM 6.4)

- GitHub v5.1.7b4 | last commit march | license EUPL-1.2

  Java and C++ application Mu-ARGUS. Tool to create safe micro-data files
- GitHub v4.2.5.2 | last commit august | license EUPL-1.2

  Java C++ Fortran and Delphi application T-ARGUS. Tool to protect statist
- CRAN 5.7.8 – 8 months ago | license GPL-2

  R package sdcMicro. Disclosure control for statistical microdata.
- CRAN 0.32.6 – a year ago | license GPL (>= 2)

  R package sdcTable. Disclosure control for tabulated data.
- CRAN 1.0.7 – 2 years ago | license Apache License 2.0 | file LICENSE

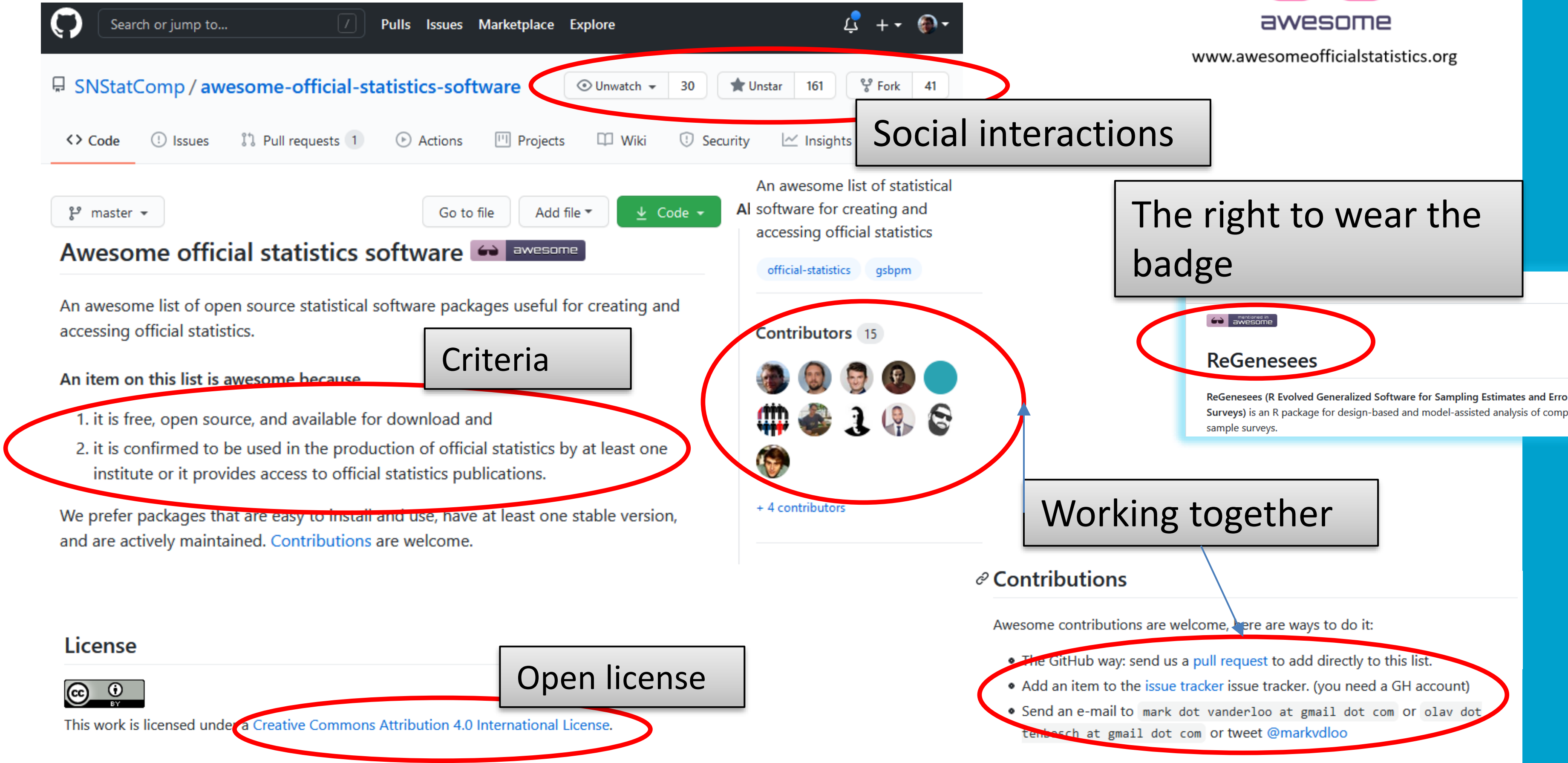
  R package easySdcTable. Provides an interface to the package sdcTable.

How does it work?

Curated list of software for official statistics

awesome

www.awesomeofficialstatistics.org



Social interactions

The right to wear the badge

Working together

Criteria

Open license

Awesome official statistics software

An awesome list of open source statistical software packages useful for creating and accessing official statistics.

An item on this list is awesome because

1. it is free, open source, and available for download and
2. it is confirmed to be used in the production of official statistics by at least one institute or it provides access to official statistics publications.

We prefer packages that are easy to install and use, have at least one stable version, and are actively maintained. Contributions are welcome.

An awesome list of statistical software for creating and accessing official statistics

official-statistics gsbpm

Contributors 15

+ 4 contributors

mentioned in awesome

ReGenesees

ReGenesees (R Evolved Generalized Software for Sampling Estimates and Erro Surveys) is an R package for design-based and model-assisted analysis of comp sample surveys.

License

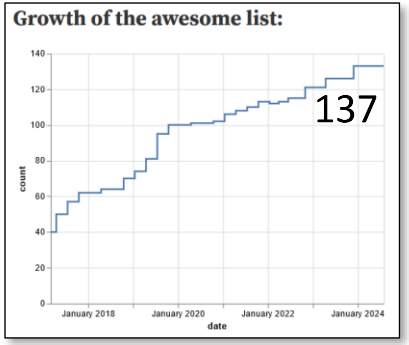
This work is licensed under a Creative Commons Attribution 4.0 International License.

Contributions

Awesome contributions are welcome, here are ways to do it:

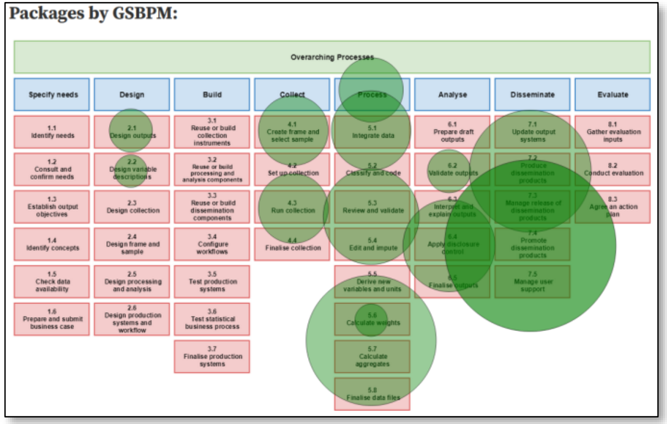
- The GitHub way: send us a pull request to add directly to this list.
- Add an item to the issue tracker issue tracker. (you need a GH account)
- Send an e-mail to `mark dot vanderloo at gmail dot com` or `olav dot tenbosch at gmail dot com` or tweet @markvdloo

uRos2023 -> uRos2024

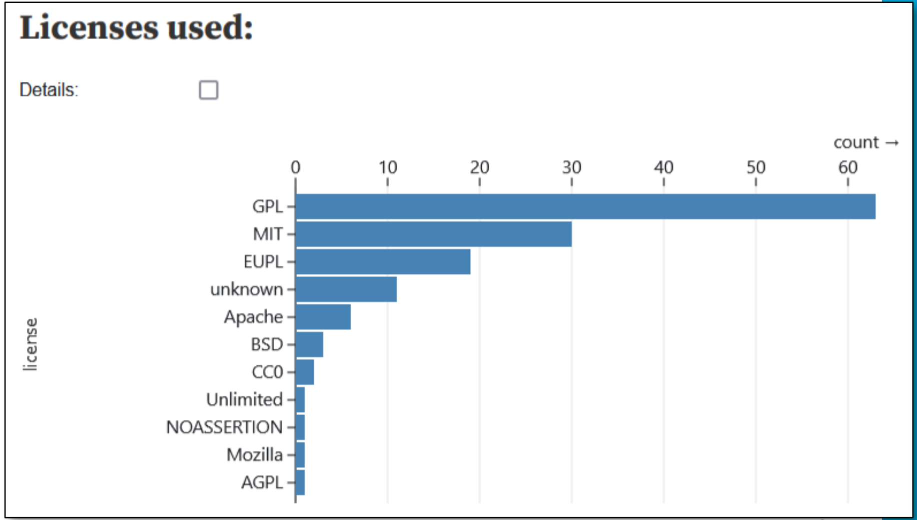

Growth of the awesome list:

137

- Quite stable: growing to 137 entries (~+10)
- UNECE HLG-MOS open source project:
  - ESS principles on OSS (derived from awesome list) adopted
  - Charter, awesome list case study
  - License stats


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			

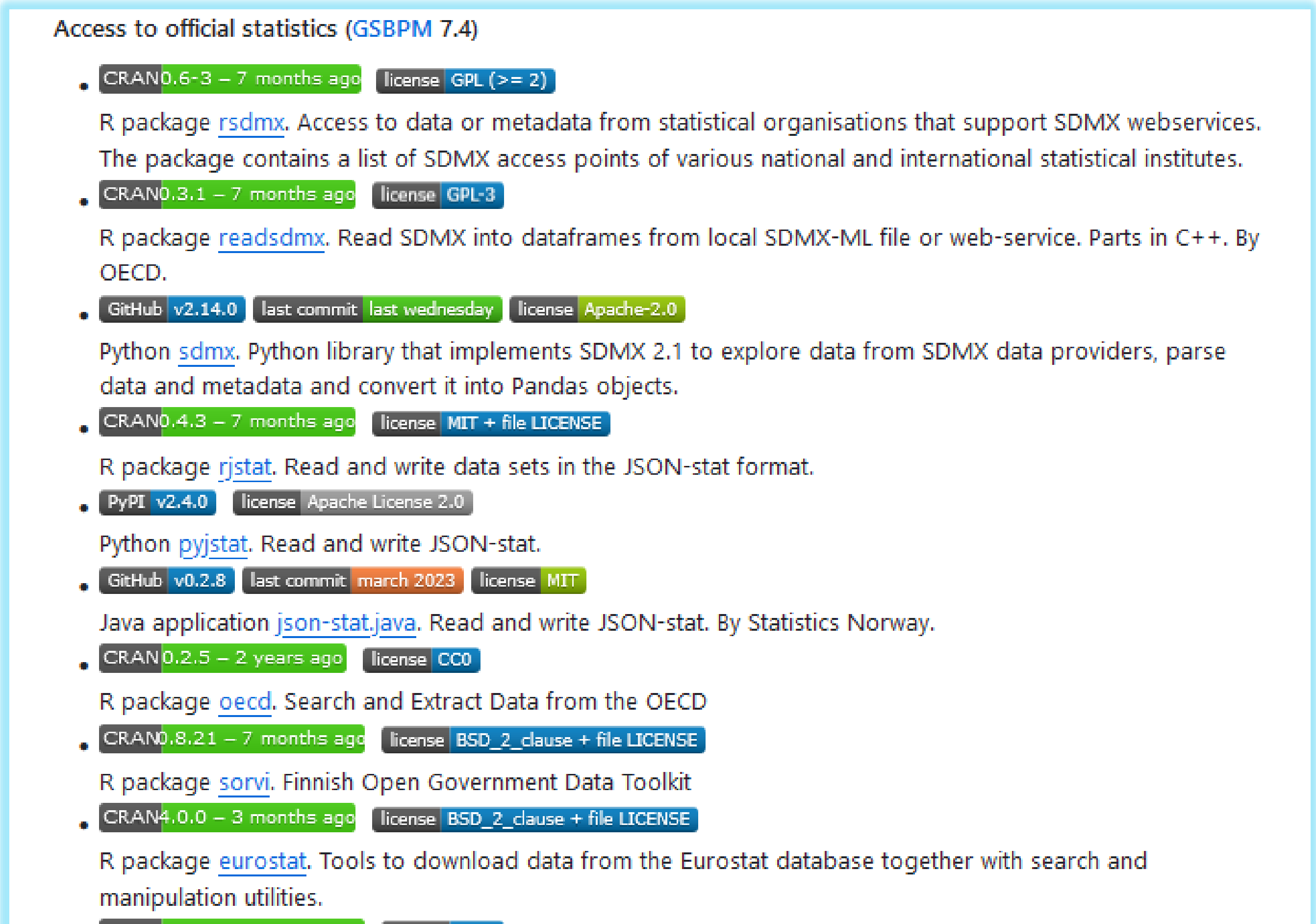

Licenses used:

Details: □

https://observablehq.com/@olavtenbosch/visualizing-awesomeofficialstatistics-org

Category “access to official statistics”

Over 30 software packages, giving access to > 80 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 SDMX webservices. The package contains a list of SDMX access points of various national and international statistical institutes.
- 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. Parts in C++. By 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 providers, parse 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

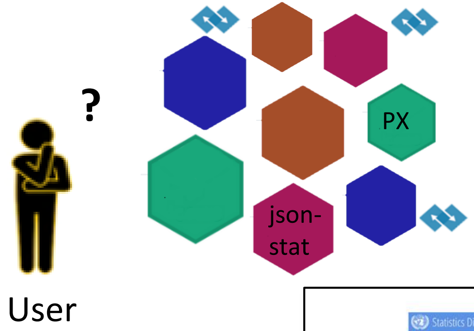
  R package oecd. Search and Extract Data from the OECD
- CRAN 0.8.21 – 7 months ago | license BSD_2_clause + file LICENSE

  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 and manipulation utilities.

Software layer surrounding Official Statistics

> 40 software packages
> 80 dataproviders
> majority R-packages
> Some use standards (SDMX, PX, JSON-STAT)



STATISTICS DENMARK | SEARCH

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

Download file as... Excel (*.xlsx) | Edit table: Pivot (clockwise)

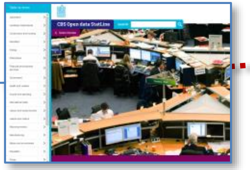
□ Codes in sep. columns
☑ Incl. Footnotes etc.

Calculate

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

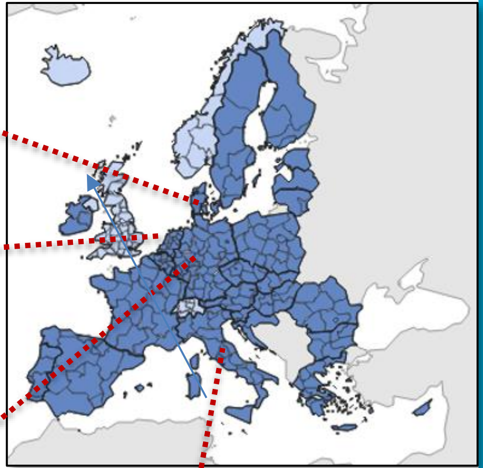
	Total	Men	Women
Economic activity rate			
2020Q4			
Age, total	79.4	82.2	76.6
15-24 years	60.1	59.6	60.6
25-34 years	84.2	88.4	79.8
35-44 years	88.2	92.1	84.2
45-54 years	88.4	90.1	86.7
55-64 years	75.3	80.1	70.5

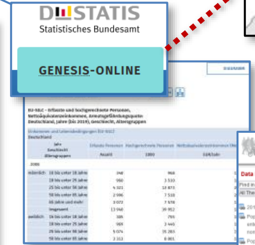
DK



CBS Open data StatLine

NL

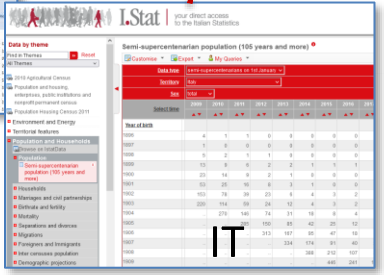




DESTATIS Statistisches Bundesamt

GENESIS-ONLINE

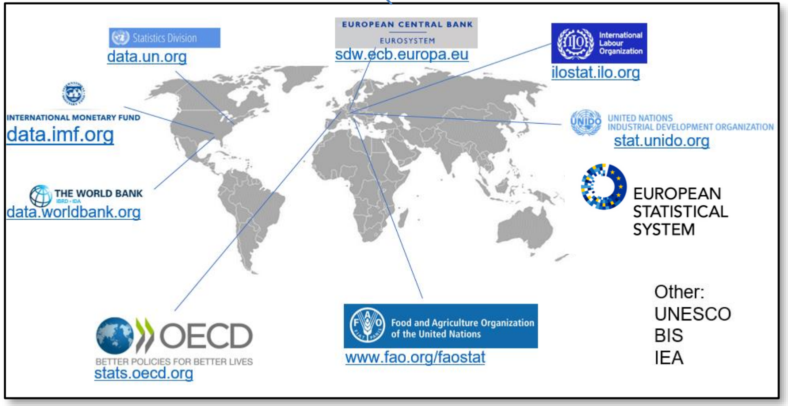
DE



I.Stat | your direct access to the Italian Statistics

Semi-supercentenarian population (105 years and more)

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
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

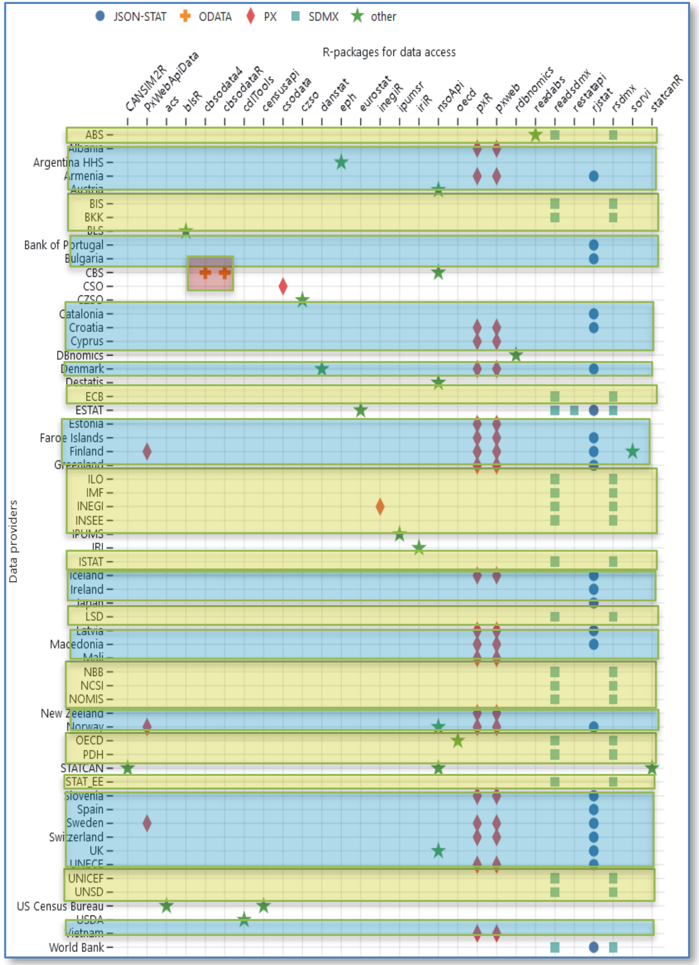
“access to official statistics” software landscape

- Matrix from docs, links to web pages and packages execution:

packages (40) x dataproviders (80) x standards (5)

- Standards: ● JSON-STAT ✚ ODATA ♦ PX ■ SDMX ★ other

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



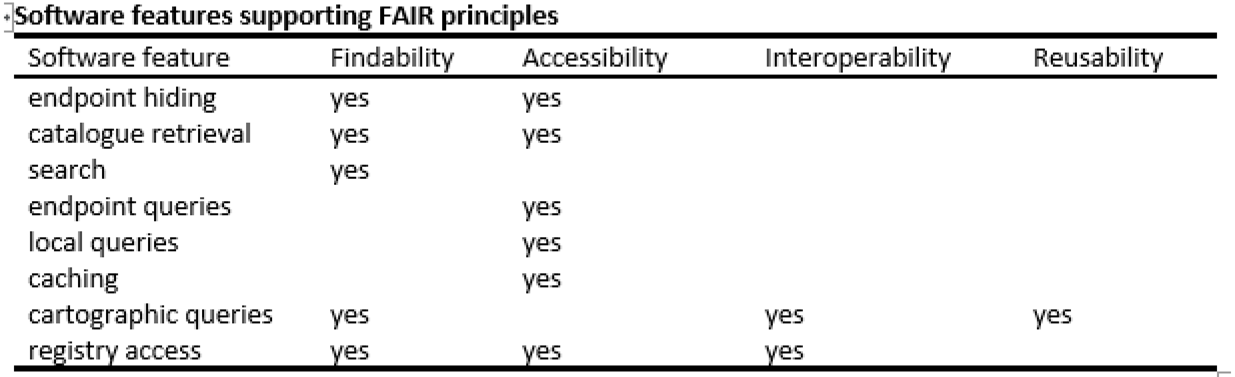
https://observablehq.com/@olavtenbosch/access_to_official_statistics

uRos2023 -> uRos2024

- New packages:
  - TEMPO (Romania), cancensus (STATCAN), insee (INSEE), nomisr (UK)
- New dataproviders per standard:
  - JSON-STAT: Kenya;
  - PX: Kosovo, Liechtenstein, North Macedonia, Jordan, Philippines, Ghana;
  - SDMX: Lithuania, Estonia, UNESCO, Chile, Cambodia, El Salvador, FAO, UAE, Luxembourg, Maldiven, Malta, Thailand, UNESCAP, Uruguay, Greece.

Features commonly offered

- ***to_df:*** getting (selections of) the data in a dataframe
- ***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

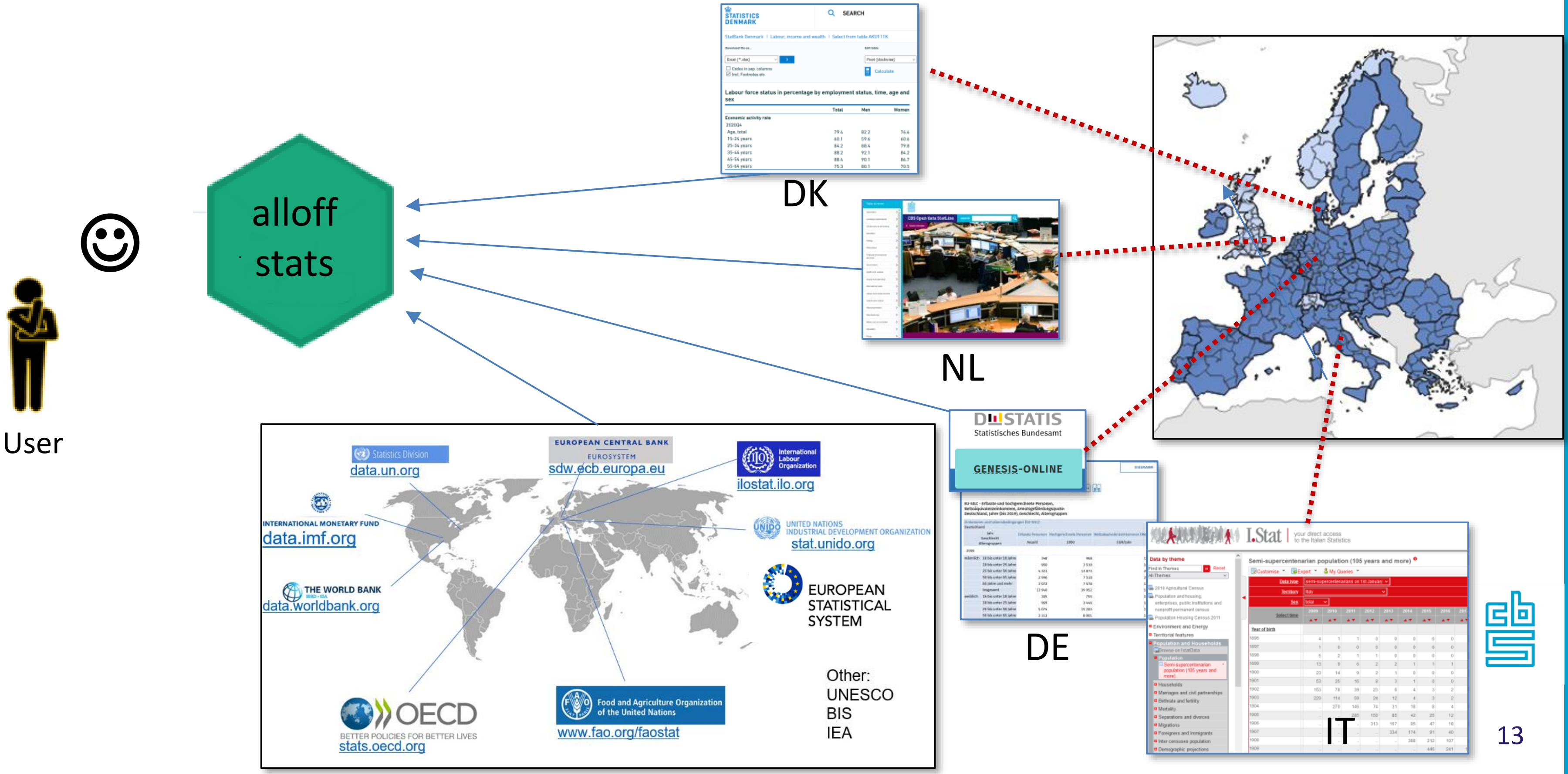


Software features supporting FAIR principles

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	

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

One FAIR R-package for all data providers?



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.

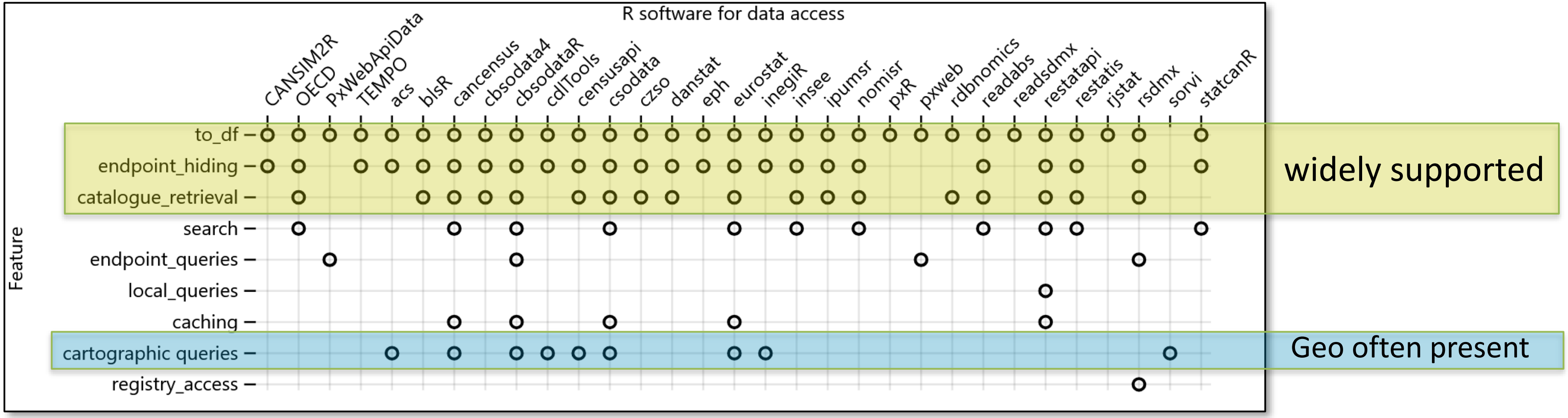
Edit table: Pivot (clockwise)
Calculate

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

☺
User
alloff stats
DK
CBS Open data StatLine
NL
DESTATIS Statistisches Bundesamt
GENESIS-ONLINE
DE
I.Stat your direct access to the Italian Statistics
Semi-supercentenarian population (105 years and more)
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

Features x R-packages

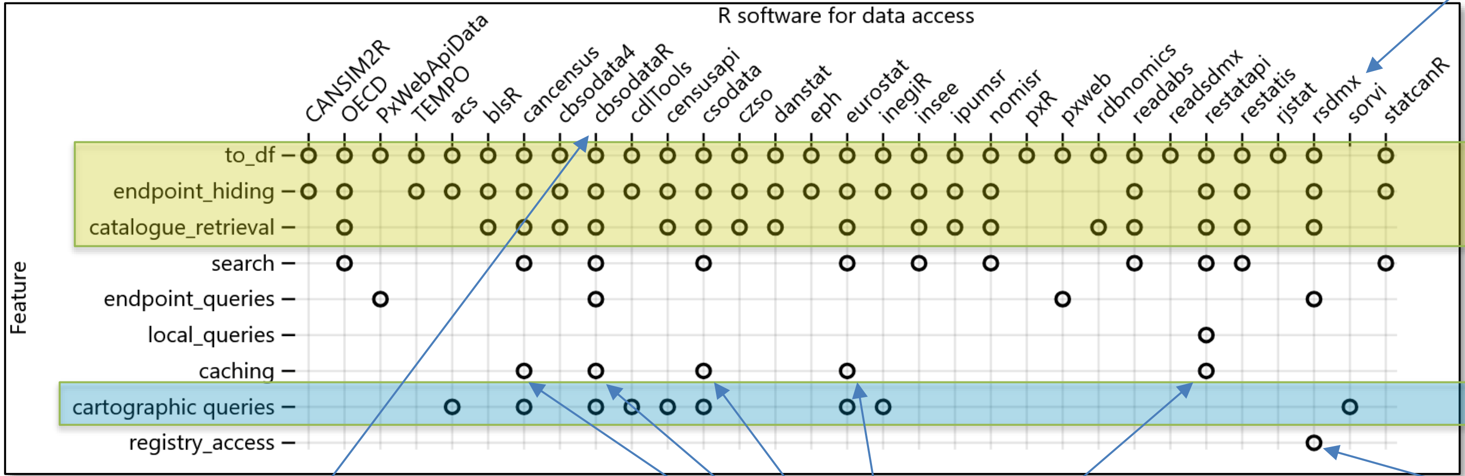


R software for data access

Feature	CANSIM2R	OECD	PxWebApiData	TEMPO	acs	blsR	cancensus	cbsodata4	cbsodataR	cdlTools	censusapi	csodata	czso	danstat	eph	eurostat	inegiR	insee	ipumsr	nomisr	pxR	pxweb	rdbnomics	readabs	readsdmx	restatapi	restatis	rjstat	rsdmx	sorvi	statcanR
to_df	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○		○
endpoint_hiding	○	○		○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○				○		○	○		○		○
catalogue_retrieval		○				○	○	○	○		○	○	○	○		○		○	○	○			○	○		○	○		○		
search		○					○		○			○				○		○		○				○		○	○				○
endpoint_queries			○						○													○							○		
local_queries																										○					
caching							○		○			○				○										○					
cartographic queries					○		○		○	○	○	○				○	○													○	
registry_access																													○		

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

Features x R-packages



R software for data access

Feature	CANSIM2R	OECD	PxWebApiData	TEMPO	acs	blsR	cancensus	cbsodata4	cbsodataR	cdlTools	censusapi	csodata	czso	danstat	eph	eurostat	inegiR	insee	ipumsr	nomisr	pxR	pxweb	rdbnomics	readabs	readsdmx	restatapi	restatis	rjstat	rsdmx	sorvi	statcanR
to_df	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○		○
endpoint_hiding	○	○		○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○				○		○	○		○		○
catalogue_retrieval		○				○	○	○	○		○	○	○	○		○		○	○	○			○	○		○	○		○		
search		○					○		○			○				○		○		○				○		○	○				○
endpoint_queries			○						○													○							○		
local_queries																										○					
caching							○		○			○				○										○					
cartographic queries					○		○		○	○	○	○				○	○													○	
registry_access																													○		

Standard-based, not feature-based

widely supported

Geo often present

NL: will be re-implemented for .STAT/SDMX

Caching surprisingly often supported

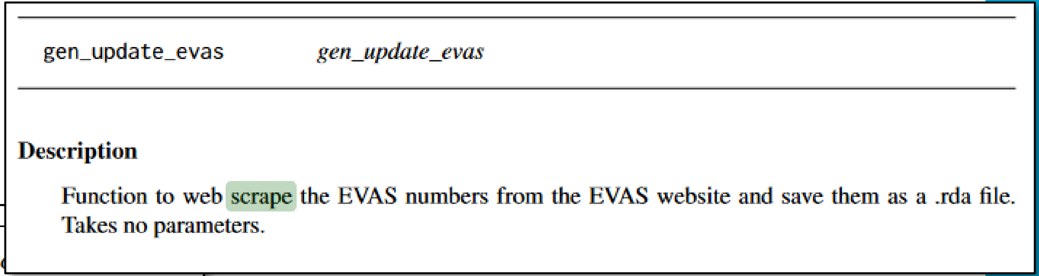
Registry support only in SDMX

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

Remarkable: scraping functions

- If not supported by API, then ***scraped*** and offered as a function in an R-package
- API developers: watch the R community for what's ***really needed*** ☺

Restatis

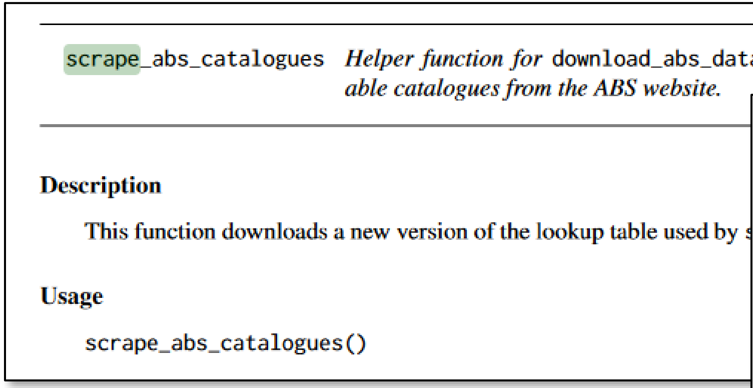


`gen_update_evas` *gen_update_evas*

Description

Function to web scrape the EVAS numbers from the EVAS website and save them as a .rda file. Takes no parameters.

readabs



`scrape_abs_catalogues` *Helper function for* `download_abs_data_cube` *to s* *able catalogues from the ABS website.*

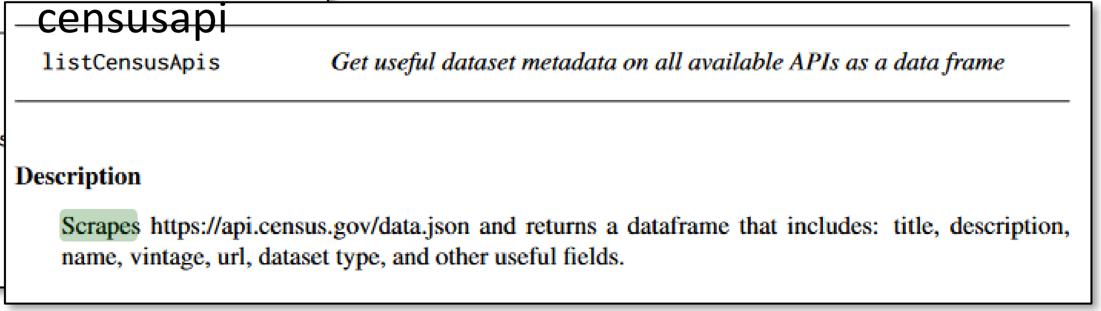
Description

This function downloads a new version of the lookup table used by s

Usage

```
scrape_abs_catalogues()
```

censusapi



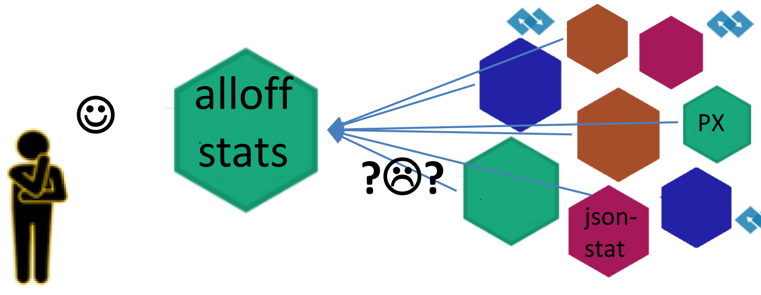
`listCensusApis` *Get useful dataset metadata on all available APIs as a data frame*

Description

Scrapes https://api.census.gov/data.json and returns a dataframe that includes: title, description, name, vintage, url, dataset type, and other useful fields.

One R-package for all data providers?

- Only for basic features:
  to_df, endpoint_hiding, catalogue_retrieval
- does it make sense?
- Would the 'user' use it?



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

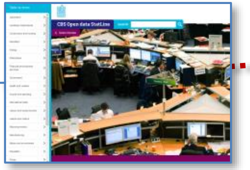
Pivot (clockwise)

Calculate

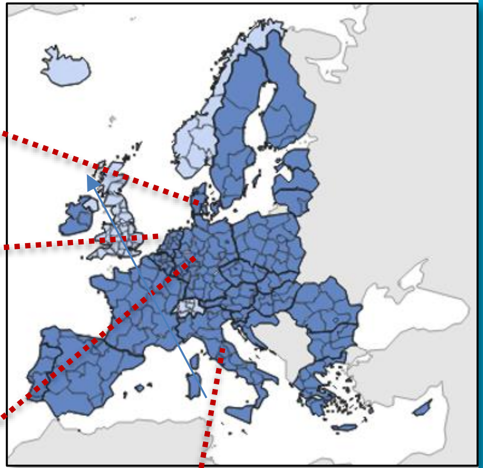
Labour force status in percentage by employment status, time, age and sex

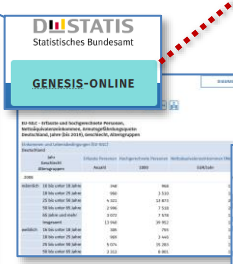
	Total	Men	Women
Economic activity rate			
2020Q4			
Age, total	79.4	82.2	76.6
15-24 years	60.1	59.6	60.6
25-34 years	84.2	88.4	79.8
35-44 years	88.2	92.1	84.2
45-54 years	88.4	90.1	86.7
55-64 years	75.3	80.1	70.5

DK



NL

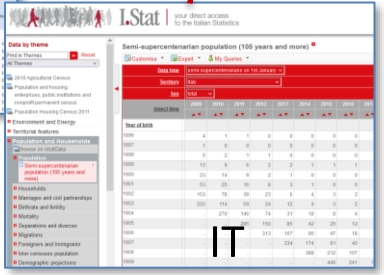




DESTATIS
Statistisches Bundesamt

GENESIS-ONLINE

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

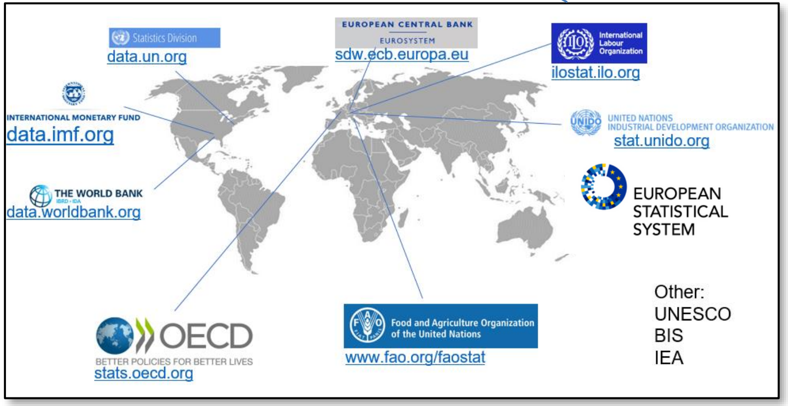
Data type: semi-supercentenarians on 1st January

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	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

Observations

- Currently ***> 30 R-packages*** for access to ***> 80 dataproviders***
- Further standardization on***: SDMX, JSON-stat, PX***
- Large ***variety*** but ***common*** features ***widely*** supported
- ***GEO features*** often supported, ***caching*** useful
- Some R-packages ***scrape*** metadata not covered in API -> added value
- Statistics Netherlands will redesign ***cbsOdataR*** for .Stat (SDMX)

- ***one*** package for all official statistics with ***all*** features -> too hard
- ***'alloffstats'*** package for all official statistics with ***common*** features -> ☺?

Common features: to_df, endpoint_hiding, catalogue_retrieval

Wrap-up

Wrap-up

- www.awesomeofficialstatistics.org awesome Please

  - Spread the word and help maintain!

- Study on category “access to offstats”:
  - Features R-packages reflect data science needs ***in practice***
  - ***Common features*** broadly supported, but different patterns
  - ‘***alloffstats***’ package with common features new goal???

Paper Cosmos 2024 conference :
https://olavtenbosch.github.io/pdf/2024_COSMOS2024_ten_Bosch_PuhlishedRef.pdf

Olav ten Bosch o.tenbosch@cbs.nl
Edwin de Jonge e.dejonge@cbs.nl