Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is
partly financed by the
European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

EUROPEAN CONFERENCE ON
QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

THE NEW ISTAT OPEN SOURCE AND STANDARDS BASED ARCHITECTURE FOR HIGH QUALITY WEB DISSEMINATION OF OFFICIAL STATISTICAL DATA

Mr Carlo Boselli, cboselli@istat.it

Mr Alessio Cardacino, alcardac@istat.it

ISTAT - Italian national institute of statistics, Italy

Instituto Nacional de Estatística
Statistics Portugal

eurostat

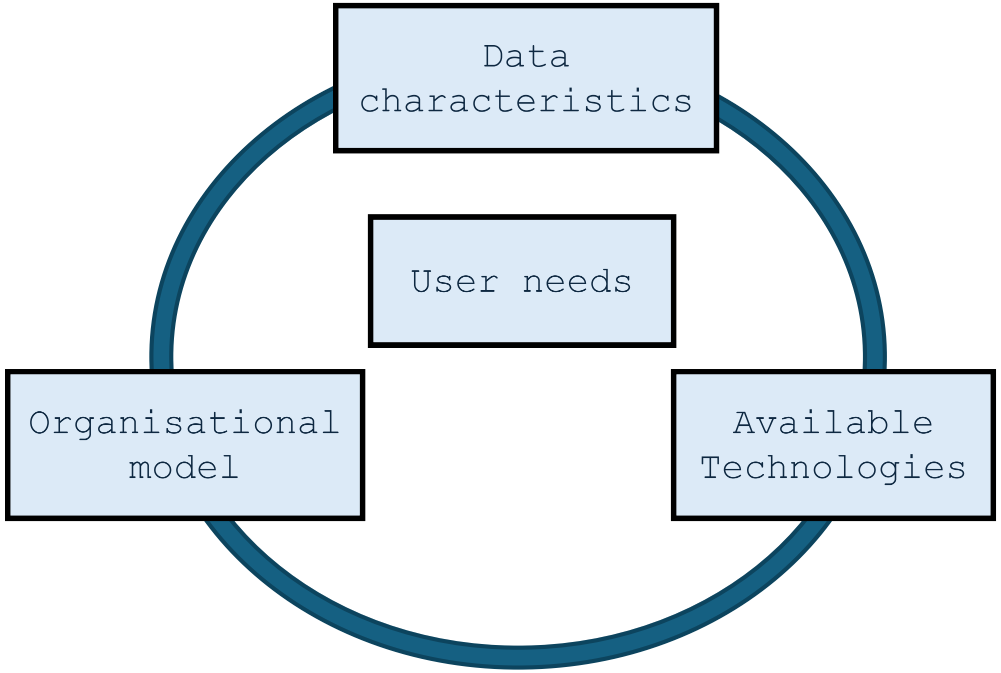
The conference is
partly financed by the
European Union

Index

- The choice of a dissemination platform
- High level needs: Stability and flexibility
- Internal and external user needs
- Techonological solution: Istat SDMX Toolkit - Data Browser, Meta and Data Manager
- Data Browser features
- Meta Data Manager features

The choice of a dissemination platform: Multidisciplinary approach

Different factors to consider before choosing a dissemination platform



Pre-existing complex system:

Different kinds of data: Admistrative data, survey data, micro and aggregate data

Different strategies and processes in the same organization

There is no better technology ever for data dissemination, but the one having the best balance between different aspects.

Technology must adapt to the objectives, the characteristics of the data and processes

Multidisciplinary approach:

Experts in business organization, statisticians, data modeling, IT

The choice of a dissemination platform: Multidisciplinary approach

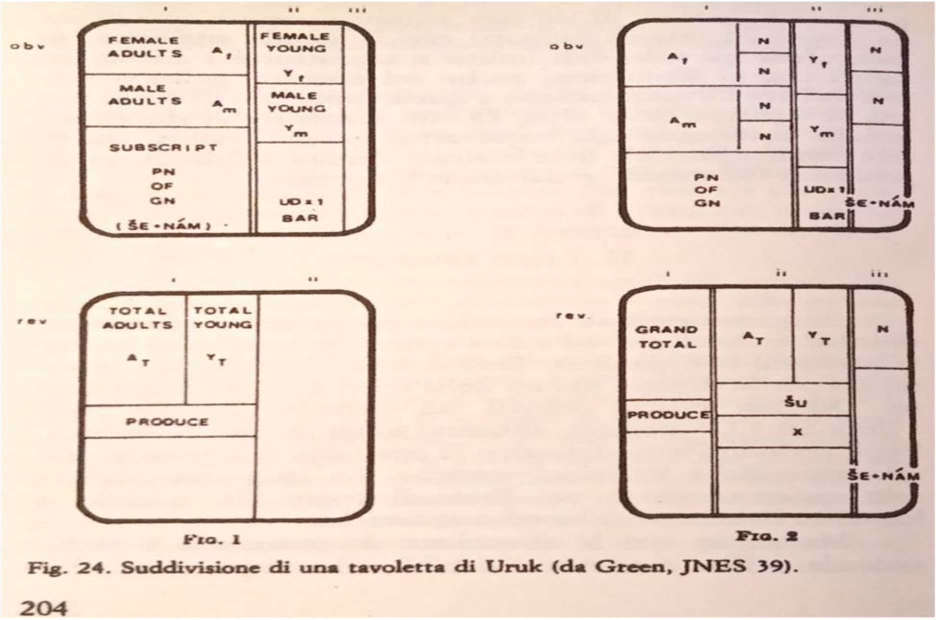


Fig. 24. Suddivisione di una tavoletta di Uruk (da Green, JNES 39).

204

Structure of Uruk tablets for the dissemination of agricultural data. (M.W. Green, Journal of Near Eastern Studies 39 - Animal Husbandry at Uruk in the Archaic Period), in «I Sumeri» - Giovanni Pettinato - Bompiani

URUK III 4500 years ago, tablets for the dissemination of agricultural data:

Subdivision of animals by sex, by age, the animals born from the precedents, the related production of butter and cheese, those responsible for registration and administrative information

A long form and a short form of the tablets structure

Metadatation, data modelling, data comparison and linkage

Standardization and stability of the dissemination process

Some topics of data dissemination are a kind of archetype, involving more the users needs, data characteristics, the organizational system than the available technology:

Consider context, user needs and technological solutions

High level needs

Stability in reporting:

National Statistical Institutes are required to respond to other organization in a reporting format. From this point of view, a protocol is recommended that stabilize data provisioning and reduce the entropy.

Need: Architecture based on a standard to guarantee data and metadata transfer via machine-to-machine.

Flexibility in dissemination:

Each National Statistical Institute has the mandate to collect data from institutional units and provide back to society aggregated data with a statistical value added, following a dissemination flexible approach. This process constantly redefine data structure, depending on new possibility to join data, new available technology, new needs from users.

Need: Tools useful to regain in terms of flexibility and guarantee the possibility, for example, to add new modalities in the codelists, update data structure definition (DSD), and in the future to modify the star schema and re-align all the depending artefacts reducing impact on external users.

SDMX – Statistical Data and Metadata eXchange

Migrate to a standard (SDMX) in order to:

- take advantage of technological modularity (changes or improvement in each component of the platform) without changes in the data modeling and preserving data dissemination processes in the future

- allow machine-to-machine access using API, and avoid misalignments between what is stored in DB and what is exposed via webservice

Internal user needs

Specific context in Istat for a large-scale dissemination of aggregate data:

Pre-existing processes:

Dissemination of aggregate data from a large number of surveys under Eurostat Regulation. A large number of processes already existing of data aggregation and dissemination.

Need: The new platform must be linkable with pre-existing processes and allow data migration form legacy systems without altering production flows or increase burden on the production sectors and reuse of existing procedures for populating and updating data and metadata

Administrative data:

Dissemination of aggregate data from a combination of administrative and survey data that is possible to integrate, for specific domains of analysis, to obtain a more granular data dissemination.

Need: High data storage capacity per single data cube. In particular, for the population and foreign trade aggregate data, there are datasets with more than 60 million of data points, and it is necessary to store, query and browse them on the web as well

External user needs

- Homogeneous environment to reduce access costs
- Possibility to publish different nodes in the same instance
- Synthetic data representation, as dashboards
- Possibility to visualize tables with an high number of data points (at least one million)
- High performance in data retrieving and visualization using an innovative cache system
- Text search system that allows the use of artificial intelligence in the future
- An ease installation and configuration of the tools

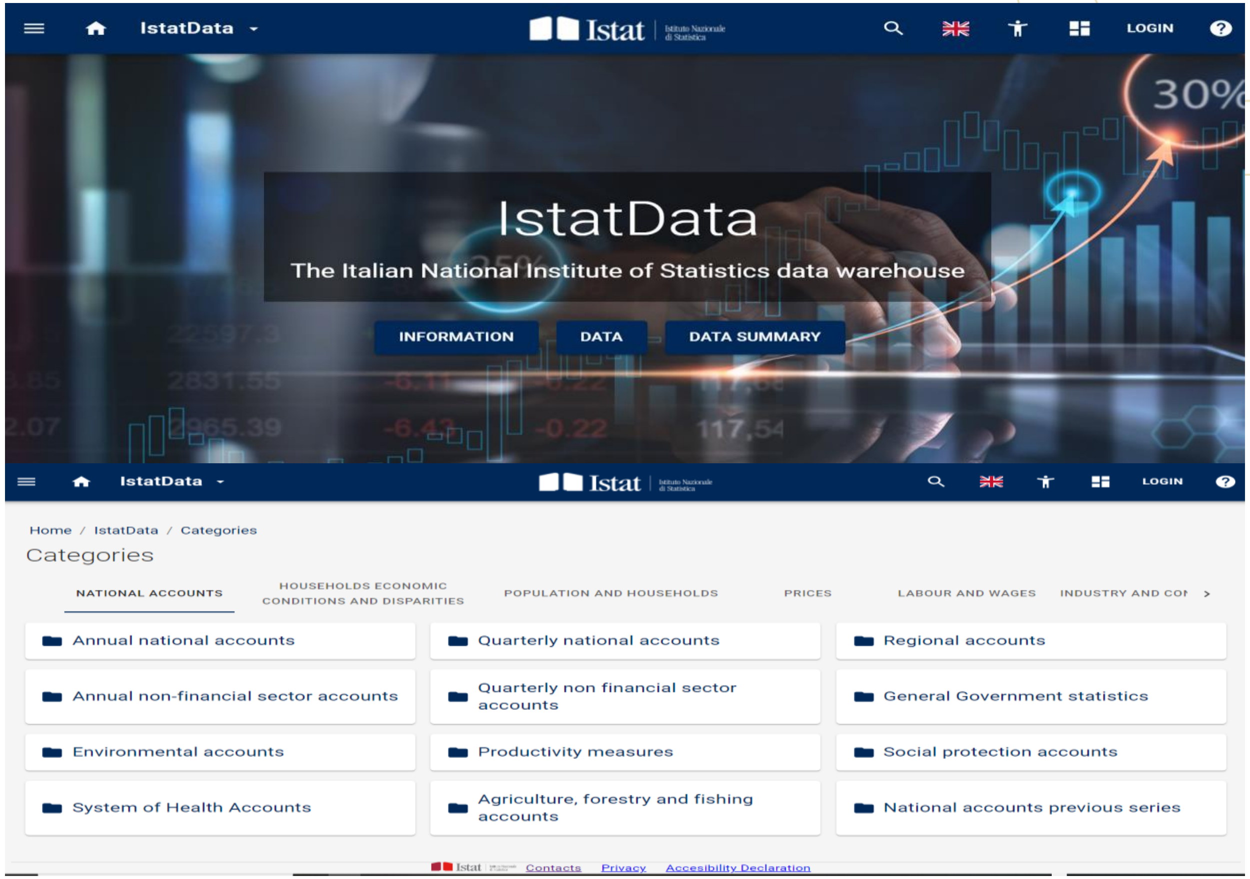
The techonological solution: SDMX Istat Toolkit - Data Browser, Meta and Data Manager

Istat SDMX Toolkit (https://sdmxistattoolkit.github.io) is an open source platform developed by Istat based on two modules:

- Data Browser (Front end): to allow external users to browse data
- Meta & Data Manager (Back end): to create SDMX data structure, data cubes, data mapping, data loading and data flows (for Istat internal users)

Front and back-end technologies are based on the SDMX standard.

Starting from 2021, IstatData (https:\\esploradati.istat.it) is the new Istat corporate data warehouse for aggregate data dissemination



IstatData

The Italian National Institute of Statistics data warehouse

Categories

- Annual national accounts
- Quarterly national accounts
- Regional accounts
- Annual non-financial sector accounts
- Quarterly non financial sector accounts
- General Government statistics
- Environmental accounts
- Productivity measures
- Social protection accounts
- System of Health Accounts
- Agriculture, forestry and fishing accounts
- National accounts previous series

Different dissemination systems towards a new single platform

Istat Legacy systems:

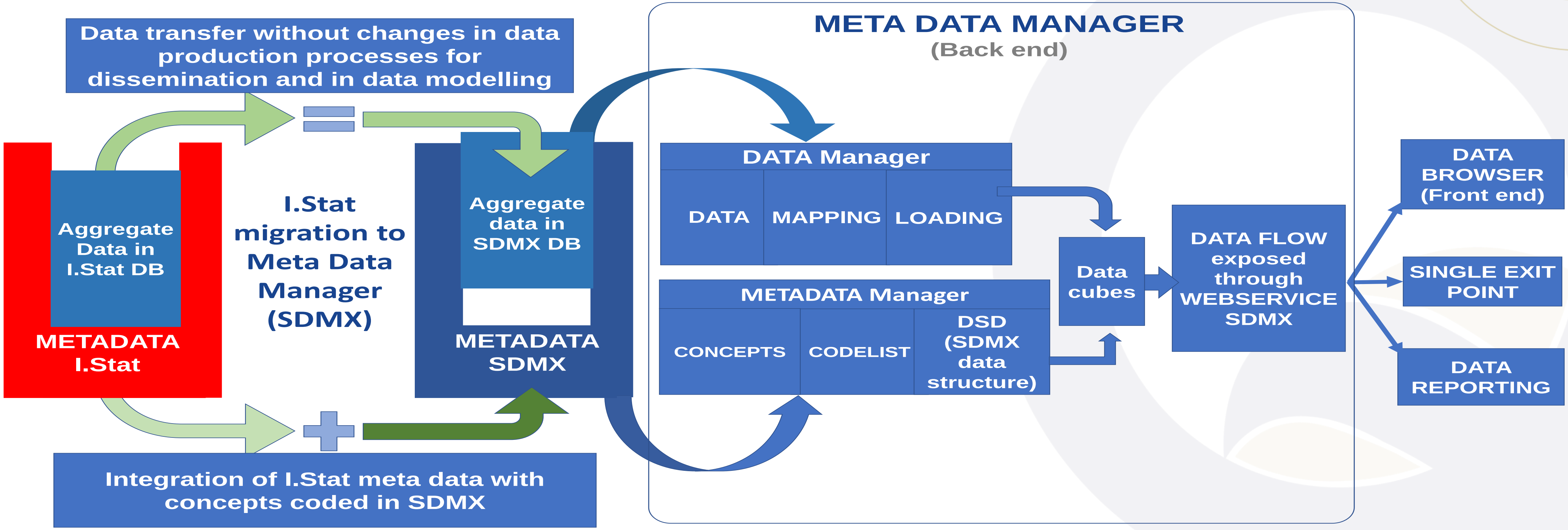
- I.Stat (800.000 users per year)
- Permanent census
- COEWEB (data of foreign trade)
- Other data bases storing aggregate data

Istat new aggregate data dissemination systems:

- IstatData (https:\\esploradati.istat.it) (2 MLD of data points, 450 datacubes, 3000 dataflows)
- Permanent census of population and housing (https://esploradati.censimentopopolazione.istat.it) (800 MLN records of data points)
- COEWEB (data of foreign trade) (https:\\esploradati.istat.it\coeweb) (10 MLD of data points, 40 datacubes)

Meta And Data Manager Tools

Set of tools useful in data modelling and data migration from legacy systems


Data transfer without changes in data production processes for dissemination and in data modelling
Aggregate Data in I.Stat DB
METADATA I.Stat
I.Stat migration to Meta Data Manager (SDMX)
Aggregate data in SDMX DB
METADATA SDMX
Integration of I.Stat meta data with concepts coded in SDMX
META DATA MANAGER
(Back end)
DATA Manager
DATA
MAPPING
LOADING
METADATA Manager
CONCEPTS
CODELIST
DSD (SDMX data structure)
Data cubes
DATA FLOW exposed through WEBSERVICE SDMX
DATA BROWSER (Front end)
SINGLE EXIT POINT
DATA REPORTING

Meta And Data Manager main features

- Modelling of statistical tables through the definition of the relevant metadata, which specifies the role that each variable plays in the structure of the tables, as well as their representation through appropriate lists of codes or classifications. Creation and publication of descriptive metadata (conceptual, methodological, quality)
- Possibility to download and upload each artefact (as Data Structure Definition DSD, codelists, concept scheme) in different format, SDMX 21., SDMX 2.0 XML
- Creation and management of statistical databases organized as multidimensional cubes in which to insert and update data starting from CSV, Excel and XML files
- Display of multidimensional statistical tables through a specific API according to the SDMX standard
- Creation and publication of catalogs of the datasets that you intend to disseminate. These catalogs are published using "DCAT vocabulary" according to the RDF standard for Linked Open Data. In particular, as default, everything necessary to implement the DCAT-AP_IT application profile created by AgID (Italian agency for digitalization) is provided;
- Creation, management and publication of thematic glossaries
- Ability to manage data structure definition (DSD), codelists and concepts through an intuitive interface that allows the internal user to easily edit, modify, upload or clone for easy reuse of the artefacts that allows you to create new versions without repeating all the steps
- Ability to add modalities in codelists in a specific DSD and datacube, without modifying the versioning, guaranteeing external users a reuse of the pre-existing dataflows
- Ability to add new attributes, with new footnotes or multiple encodings associated for a specific dimension of analysis, directly updating a pre-existing DSD, without destroying data cubes, dataflows and other depending artefacts
- Using Meta-Data Handler tools, ability to make available data, already published according to a specific data structure, on the base of a different DSD defined by another international organization for reporting purposes
- Ability to upload pre-existing csv files, with different formats, using file-mapping tools. This functionality allows internal user to upload in the Meta and Data Manager, without changes in the format, files already defined from production units to populate legacy dissemination systems, reducing the burden during data migration

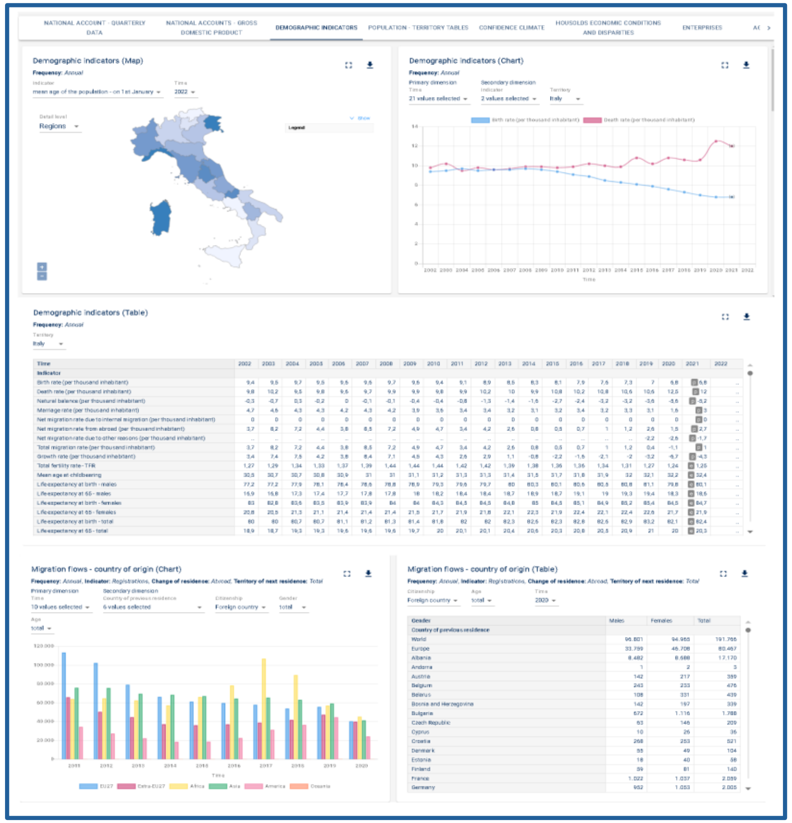
EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

DATA BROWSER: Data comparison and synthetic prospects (Dashboard)



Demographic indicators (Map)

Frequency: *Annual*

Indicator: mean age of the population - on 1st January

Time: 2022

Detail level: Regions

Demographic indicators (Chart)

Frequency: *Annual*

Primary dimension — Time: 21 values selected

Secondary dimension — Indicator: 2 values selected

Territory: Italy

Demographic indicators (Table)

Frequency: *Annual*

Territory: Italy

Time	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Indicator																					
Birth rate (per thousand inhabitant)	9,4	9,5	9,7	9,5	9,5	9,5	9,7	9,5	9,4	9,1	8,9	8,5	8,3	8,1	7,9	7,6	7,3	7	6,8	6,8	..
Death rate (per thousand inhabitant)	9,8	10,2	9,5	9,8	9,5	9,7	9,9	9,9	9,8	9,9	10,2	10	9,9	10,8	10,2	10,8	10,6	10,6	12,5	12	..
Natural balance (per thousand inhabitant)	-0,3	-0,7	0,3	-0,2	0	-0,1	-0,1	-0,4	-0,4	-0,8	-1,3	-1,4	-1,6	-2,7	-2,4	-3,2	-3,2	-3,6	-5,6	-5,2	..
Marriage rate (per thousand inhabitant)	4,7	4,5	4,3	4,3	4,2	4,3	4,2	3,9	3,5	3,4	3,4	3,2	3,1	3,2	3,4	3,2	3,3	3,1	1,6	3	..
Net migration rate due to internal migration (per thousand inhabitant)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	..
Net migration rate from abroad (per thousand inhabitant)	3,7	8,2	7,2	4,4	3,8	8,5	7,2	4,9	4,7	3,4	4,2	2,6	0,8	0,5	0,7	1	1,2	2,6	1,5	2,7	..
Net migration rate due to other reasons (per thousand inhabitant)	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	-2,2	-2,6	-1,7	..
Total migration rate (per thousand inhabitant)	3,7	8,2	7,2	4,4	3,8	8,5	7,2	4,9	4,7	3,4	4,2	2,6	0,8	0,5	0,7	1	1,2	0,4	-1,1	1	..
Growth rate (per thousand inhabitant)	3,4	7,4	7,5	4,2	3,8	8,4	7,1	4,5	4,3	2,6	2,9	1,1	-0,8	-2,2	-1,6	-2,1	-2	-3,2	-6,7	-4,3	..
Total fertility rate - TFR	1,27	1,29	1,34	1,33	1,37	1,39	1,44	1,44	1,44	1,42	1,42	1,39	1,38	1,36	1,36	1,34	1,31	1,27	1,24	1,25	..
Mean age at childbearing	30,5	30,7	30,7	30,8	30,9	31	31	31,1	31,2	31,3	31,3	31,4	31,5	31,7	31,8	31,9	32	32,1	32,2	32,4	..
Life expectancy at birth - males	77,2	77,2	77,9	78,1	78,4	78,6	78,8	78,9	79,3	79,6	79,7	80	80,3	80,1	80,6	80,5	80,8	81,1	79,8	80,1	..
Life expectancy at 65 - males	16,9	16,8	17,3	17,4	17,7	17,8	17,8	18	18,2	18,4	18,4	18,7	18,9	18,7	19,1	19	19,3	19,4	18,3	18,6	..
Life expectancy at birth - females	83	82,8	83,6	83,5	83,9	83,9	84	84	84,3	84,5	84,5	84,8	85	84,5	85,1	84,9	85,2	85,4	84,5	84,7	..
Life expectancy at 65 - females	20,8	20,5	21,3	21,1	21,4	21,4	21,4	21,5	21,7	21,9	21,8	22,1	22,3	21,9	22,4	22,1	22,4	22,6	21,7	21,9	..
Life expectancy at birth - total	80	80	80,7	80,7	81,1	81,2	81,3	81,4	81,8	82	82	82,3	82,6	82,3	82,8	82,6	82,9	83,2	82,1	82,4	..
Life expectancy at 65 - total	18,9	18,7	19,3	19,3	19,6	19,6	19,6	19,7	20	20,1	20,1	20,4	20,6	20,3	20,8	20,5	20,9	21	20	20,3	..

Migration flows - country of origin (Chart)

Frequency: *Annual*, **Indicator:** *Registrations*, **Change of residence:** *Abroad*, **Territory of next residence:** *Total*

Primary dimension — Time: 10 values selected

Secondary dimension — Country of previous residence: 6 values selected

Citizenship: Foreign country

Gender: total

Age: total

Migration flows - country of origin (Table)

Frequency: *Annual*, **Indicator:** *Registrations*, **Change of residence:** *Abroad*, **Territory of next residence:** *Total*

Citizenship: Foreign country

Age: total

Time: 2020

Gender	Males	Females	Total
Country of previous residence			
World	96.801	94.965	191.766
Europe	33.759	46.708	80.467
Albania	8.482	8.688	17.170
Andorra	1	2	3
Austria	142	217	359
Belgium	243	233	476
Belarus	108	331	439
Bosnia and Herzegovina	142	197	339
Bulgaria	672	1.116	1.788
Czech Republic	63	146	209
Cyprus	10	26	36
Croatia	268	253	521
Denmark	55	49	104
Estonia	18	40	58
Finland	59	81	140
France	1.022	1.037	2.059
Germany	952	1.053	2.005

Combination of maps, charts tables and explanatory text

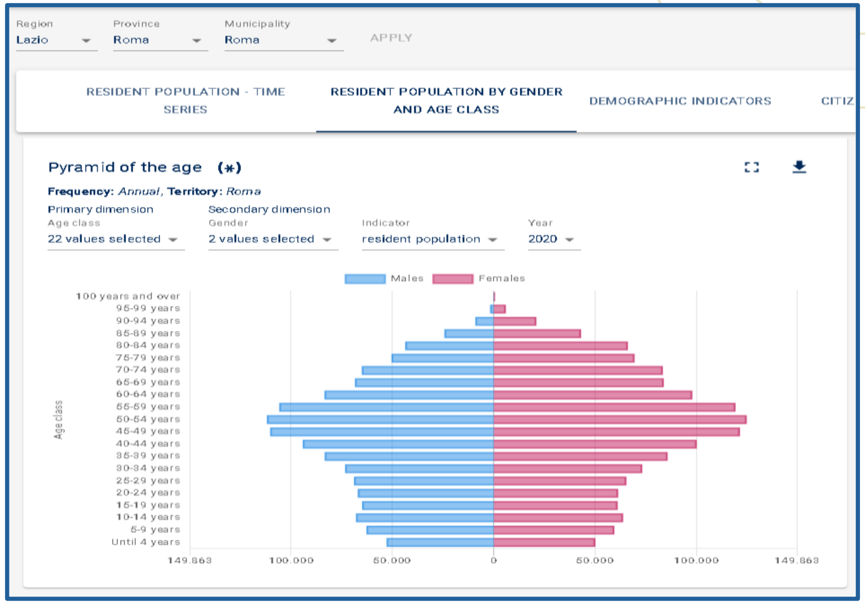
Objects created by combining different dataflow from different datasets and different nodes

Interactive selection

Full screen enlargement for each object

Data export in different format

Image export for charts and maps



Region: Lazio

Province: Roma

Municipality: Roma

Pyramid of the age (*)

Frequency: *Annual*, **Territory:** *Roma*

Primary dimension — Age class: 22 values selected

Secondary dimension — Gender: 2 values selected

Indicator: resident population

Year: 2020

Territorial data selection system at regional, provincial and municipality level: all the dashboard data automatically updated for a specific territory

Data Browser main features

- This application can be used within a single organization in order to disseminate datasets stored in one or more databases (e.g. https://esploradati.istat.it), or within a distributed architecture on web servers reachable via the http protocol on the Internet (e.g. https://idp.sister.it).Possibility to download and upload each artefact (as Data Structure Definition DSD, codelists, concept scheme) in different format, SDMX 21., SDMX 2.0 XML
- It is possible to browse datasets in the format of multidimensional tables, associated with various graph modes, thematic maps, and customize the layout of all forms of visualization. It is also possible to download data in various formats, including open data such as XML (SDMX-ML), SDMX-CSV, SDMX-JSON, CSV (custom), Excel, image and pdf (for graphs and maps).
- For the aspects related to performances, the module allows browsing a dataset with more than 300 million of data points, visualizing on the web a table with more than one million of data points per time and providing search criteria to change the selection to browse all stored data.
- The system is also equipped with a sophisticated cache, in order to make queries and data visualizations highly efficient even in the presence of particularly large datasets. This solution allows storing and reusing, for subsequent queries, the results of all the queries carried out by users. This system is consistent with updating criteria of the datasets themselves, which invalidate the caches previously created.
- Other features are the possibility to browse and search datasets within each single data source by typing keywords or directly selecting the dataset of interest with the help of specific hierarchical "theme trees", and the possibility for registered users to save their data views within a portfolio that can be consulted once logged into the system.

Webservice and single exit point

Architecture natively based on an internationally recognized data transfer standard (SDMX)

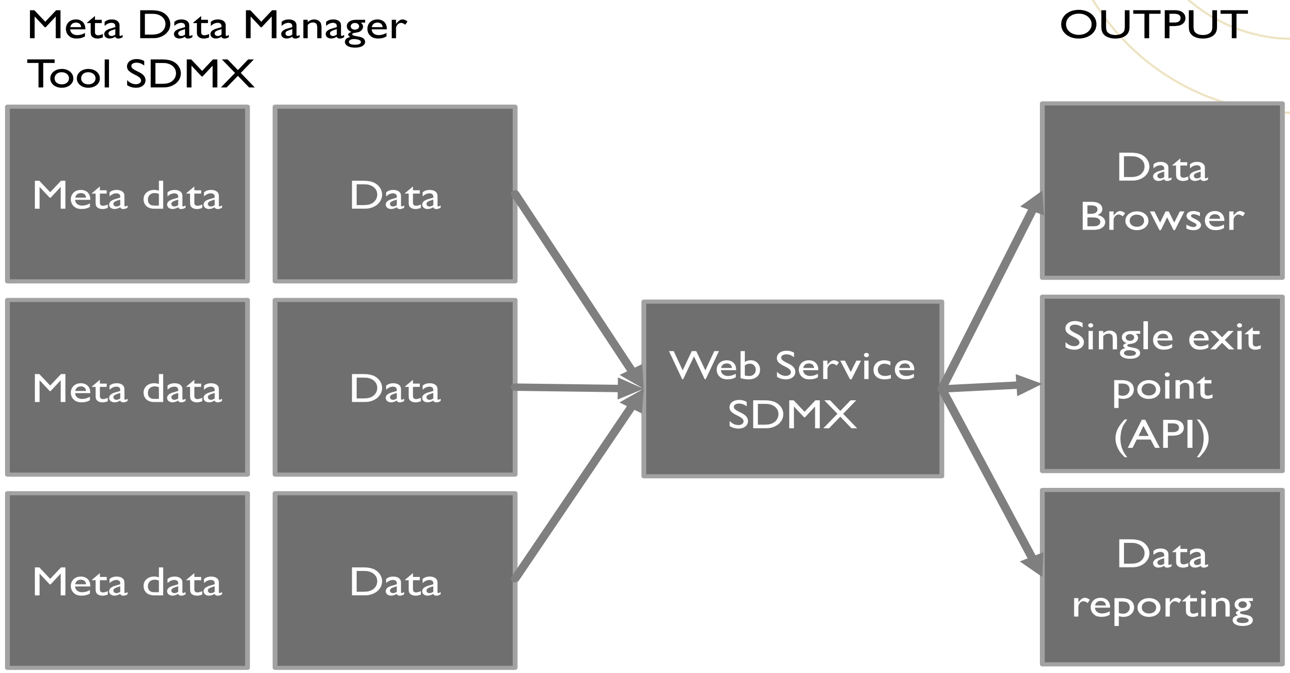
- High data store capacity per single datacube
- Modular infrastructure with access via webservice
- New technological framework is developed by ISTAT and completely Open Source

IT reference colleagues for technological aspects:

Francesco Rizzo (rizzo@istat.it)

Alessio Cardacino (alcardac@istat.it)

Simone Coccia (sicoccia@istat.it)



https://github.com/SDMXISTATTOOLKIT

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is
partly financed by the
European Union