Webinar on Building a Central Statistical Data Warehouse
Doha, 17 May 2022

حلقة عمل عن بعد بشأن بناء قاعدة بيانات إحصائية مركزية
الدوحة، 17 مايو 2022

Istat Istituto Nazionale di Statistica

Developing a statistical Data warehouse
17 May 2022

The Statistical Dissemination Data Warehouse in Istat

Domenico Fedele, Francesco Rizzo
Istat

Summary

- Standardisation activities in Istat
- Why supporting the dissemination business process with a Statistical DWH
- Dissemination DWH implementation steps and timeline
- Dissemination DWH architecture
- Distributed Dissemination DWH within the National Statistical System
- Dissemination DWH: the free and open source toolkit
- Design and implementation principles of the Toolkit
- Processing – Cubus Toolkit
- Transformation – Excel2CSV tool
- Meta and Data Manager - main features, architectures
- Data Browser - main features
- Lesson learnt

Standardisation activities in ISTAT

ISTAT has been running a modernisation program based on the guidelines of the UNECE “High-level Group for the Modernisation of Statistical Production and Services” in order to:

- ❑ Satisfy new demands for statistical information:
  - ❑ Produce more statistical indicators with more sectorial and territorial details
  - ❑ Improve quality (*Coherence, Timeliness, Comparability, Accessibility*)
  - ❑ Provide Governments needs to help the formulation of good policy, not only at national level
- ❑ Streamline the expenses due to a reduction of the financial allocation
  - ❑ Leveraging the ICT development that has allowed a cost reduction in producing statistics and its easily accessibility and dissemination
- ❑ Seizing the opportunities that Internet is offering in terms of new information sources and new ways of combining and using information

Modernisation = Standardisation + Industrialisation

Why supporting the dissemination business process with a Statistical DWH

❑ Facilitating data integration and process integration

❑ Providing a streamlined infrastructure for the standardisation and industrialisation of the data life cycle

❑ Driving the (data-centric) workflow through standardised information objects between different processes (metadata-drive approach)

❑ Increasing the quality of the statistics by reducing manual transformation steps with automated and stable processes

❑ Reducing the files dispersion on the local disks of statisticians, offering a dedicated collaborative infrastructure using databases

❑ Overcoming the issues related to dated technologies of the existing legacy systems by implementing a new services based IT platform in a loosely coupled architecture

❑ Offering the same platform to the Organisations part of the National Statistical System (implementing a distributed DWH) in order to improve quality in dissemination

Dissemination DWH implementation steps and timeline

❑ Analysis of different options: commercial solutions, tools available within the statistical community, evolution of a prototype developed in-house

⇩

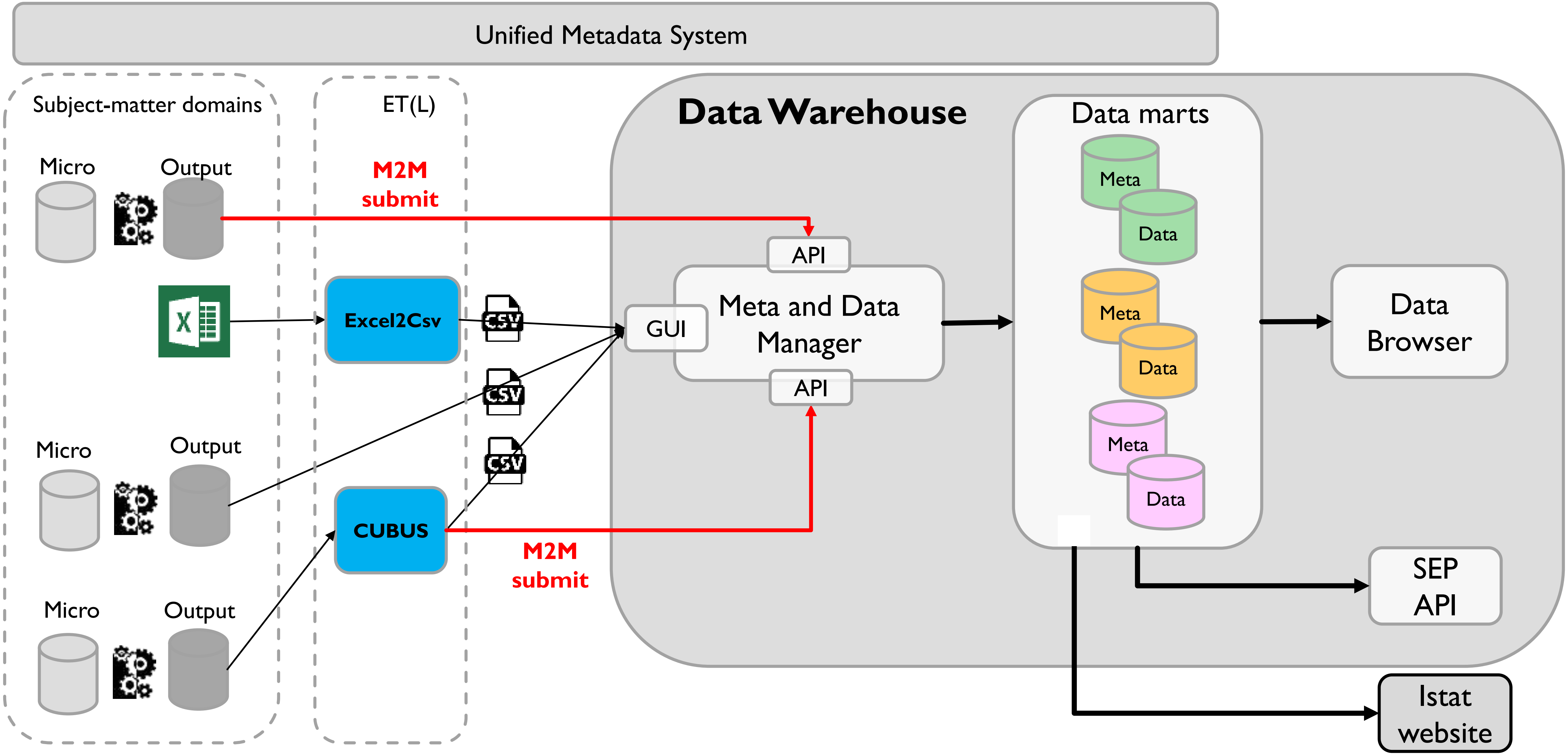
evolve the prototype and complement it with tools developed in other context

❑ Design a suitable architecture

❑ organize the needed building blocks as a reusable Toolkit (free and open source)

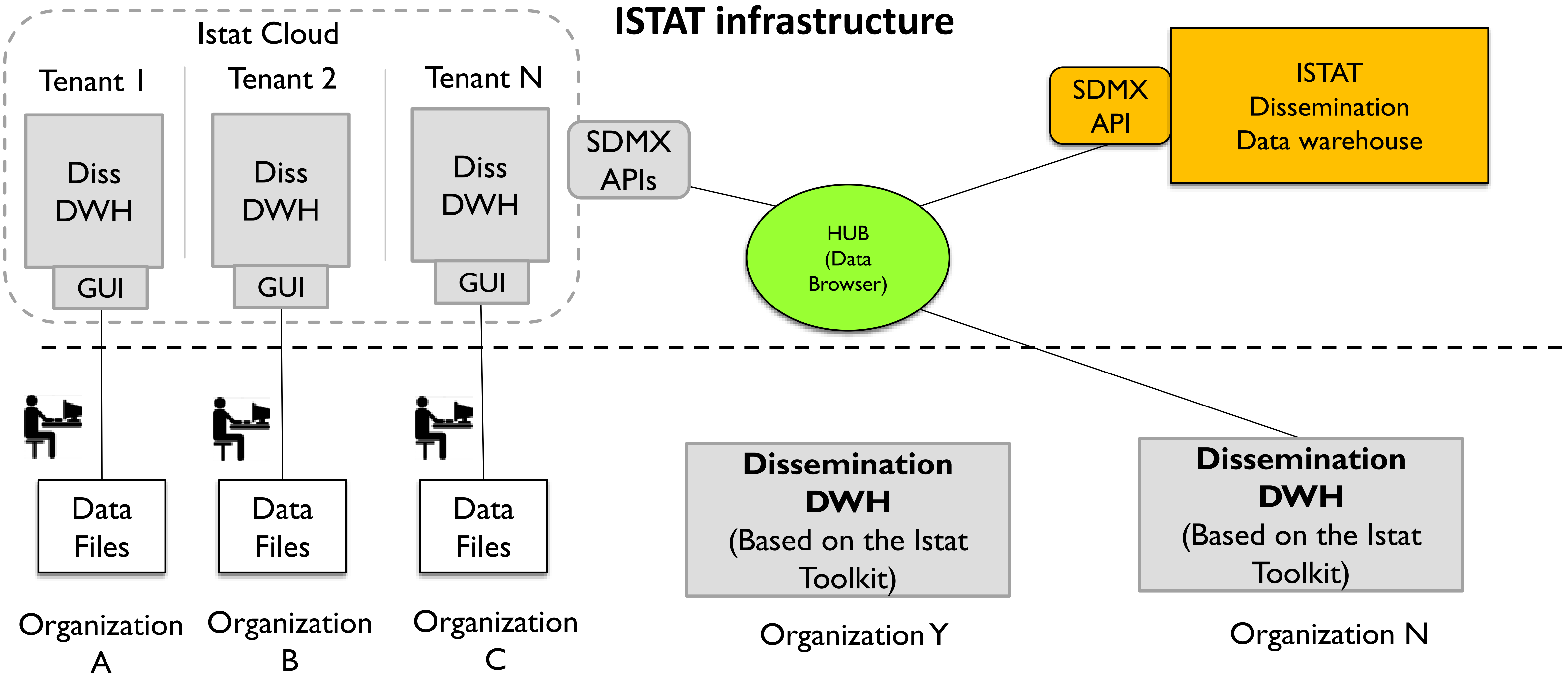
❑ Plan migration from legacy systems to the new DWH platform

Dissemination Data Warehouse architecture in Istat



Istat

Distributed Dissemination DWH within the National Statistical System



National Statistical System Organizations premises

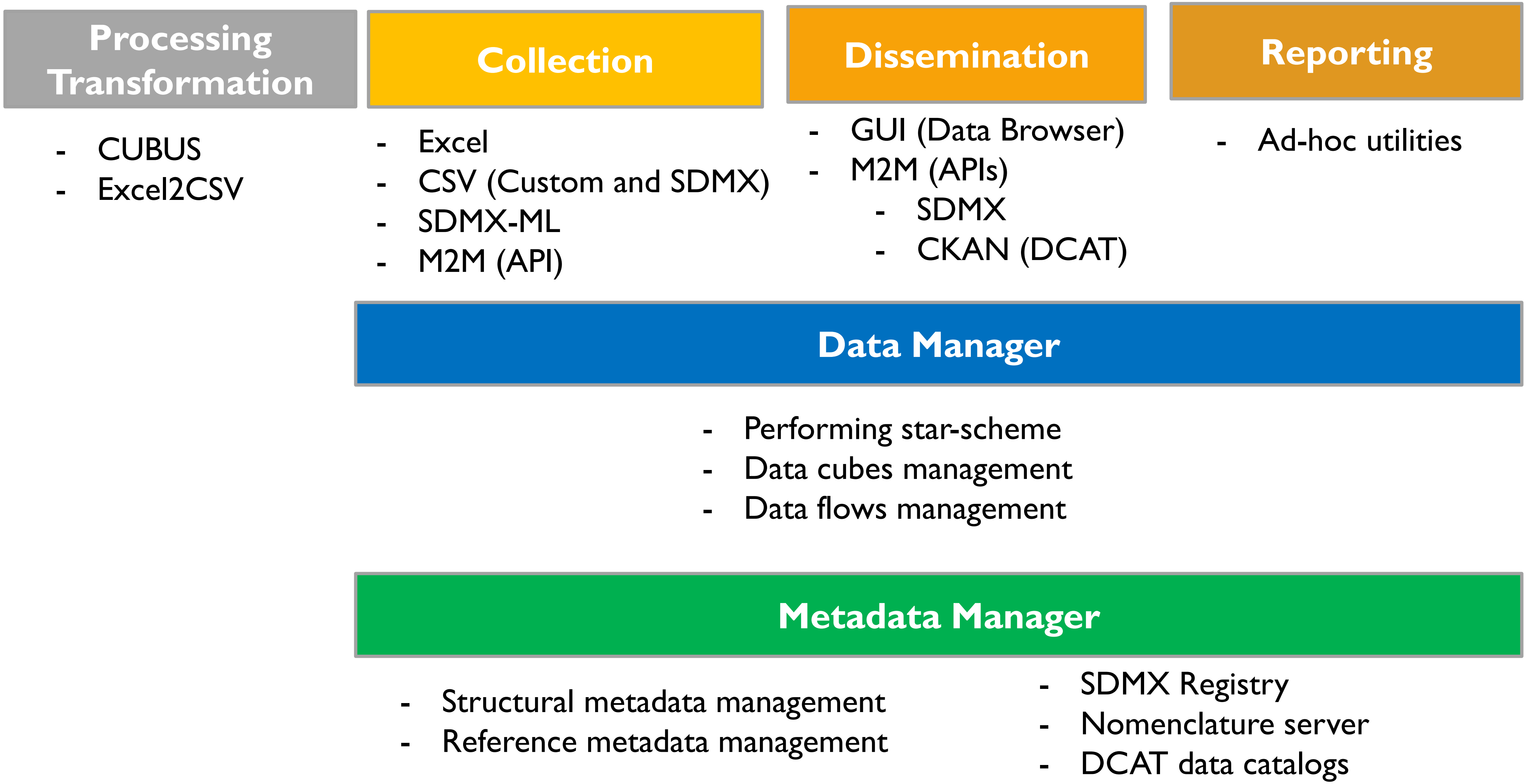
Dissemination DWH

The free and open source toolkit

Design and implementation principles of the Toolkit

- ❑ Reengineering by experience
  - Old version used as prototype
  - Feedbacks collected from +500 users (trainings and national and international projects)
- ❑ Standards-based
  - SDMX, DCAT
- ❑ Easy to install (Plug and Play)
  - Not more than 20 min
  - Technical skills easy to find in statistical organizations
- ❑ Easy to configure
  - Default configuration (Play)
  - Extended configuration through GUI
- ❑ Easy to change the brand and to localize in different languages
  - CSS files
  - resource files
- ❑ Adequate performance
  - Data rendering of a multidimensional table with 100.000 cells in less than 10 seconds
  - Benchmark with the legacy Istat dissemination I.Stat, and other tools available in the statistical community
- ❑ Easy to use (GUI-design driven by users)

The Toolkit and the statistical business processe



Processing Transformation

- CUBUS
- Excel2CSV

Collection

- Excel
- CSV (Custom and SDMX)
- SDMX-ML
- M2M (API)

Dissemination

- GUI (Data Browser)
- M2M (APIs)
  - SDMX
  - CKAN (DCAT)

Reporting

- Ad-hoc utilities

Data Manager

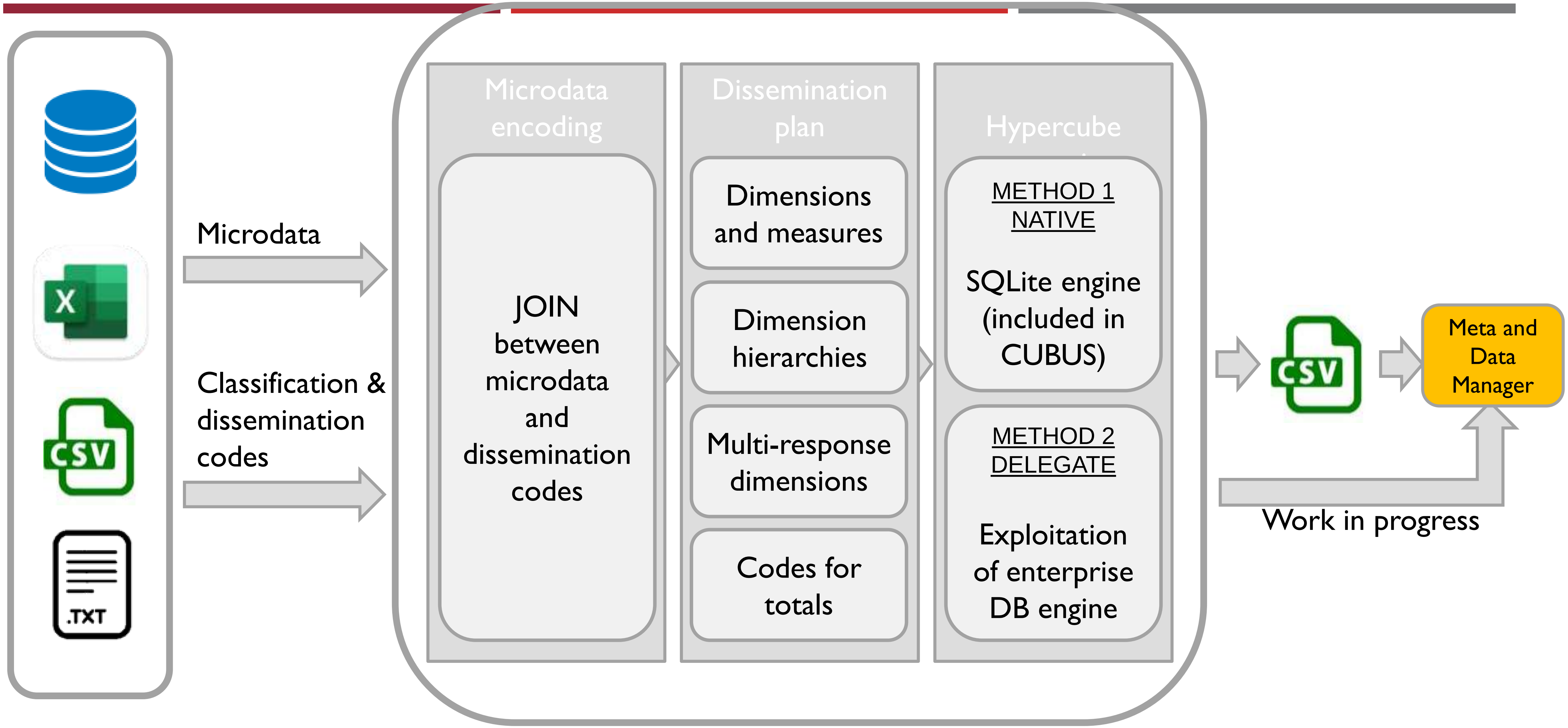
- Performing star-scheme
- Data cubes management
- Data flows management

Metadata Manager

- Structural metadata management
- Reference metadata management

- SDMX Registry
- Nomenclature server
- DCAT data catalogs

Processing – Cubes Tool



Microdata

Classification & dissemination codes

Microdata encoding

JOIN between microdata and dissemination codes

Dissemination plan

Dimensions and measures

Dimension hierarchies

Multi-response dimensions

Codes for totals

Hypercube

METHOD 1 NATIVE

SQLite engine (included in CUBUS)

METHOD 2 DELEGATE

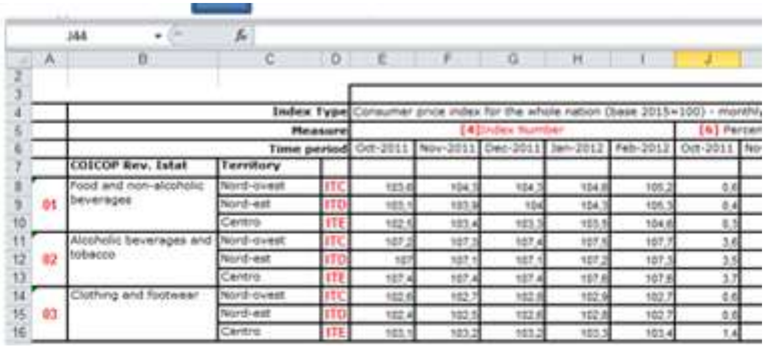
Exploitation of enterprise DB engine

Meta and Data Manager

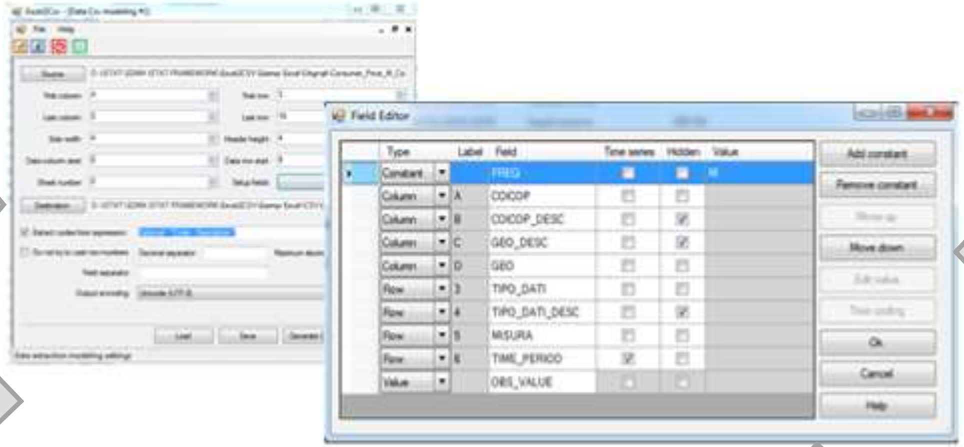
Work in progress

Cubus

Transformation – Excel2Csv Tool



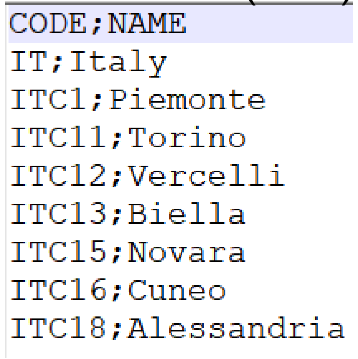
	A	B	C	D	E	F	G	H	I	J
2										
3										
4			Index Type		Consumer price index for the whole nation (base 2015=100) - monthly					
5			Measure		[4]Index Number					[6] Percen
6			Time period		Oct-2011	Nov-2011	Dec-2011	Jan-2012	Feb-2012	Oct-2011
7		COICOP Rev. Istat	Territory							
8	01	Food and non-alcoholic beverages	Nord-ovest	ITC	103,6	104,1	104,3	104,8	105,2	0,6
9			Nord-est	ITD	103,1	103,9	104	104,3	105,3	0,4
10			Centro	ITE	102,5	103,4	103,3	103,5	104,6	0,5
11	02	Alcoholic beverages and tobacco	Nord-ovest	ITC	107,2	107,3	107,4	107,5	107,7	3,6
12			Nord-est	ITD	107	107,1	107,1	107,2	107,3	3,5
13			Centro	ITE	107,4	107,4	107,4	107,6	107,8	3,7
14	03	Clothing and footwear	Nord-ovest	ITC	102,6	102,7	102,8	102,9	102,7	0,6
15			Nord-est	ITD	102,4	102,5	102,6	102,8	102,7	0,6
16			Centro	ITE	103,1	103,2	103,2	103,3	103,4	1,4



	Type	Label	Field	Time series	Hidden	Value
▸	Constant		FREQ	□	□	M
	Column	A	COICOP	□	□	
	Column	B	COICOP_DESC	□	☑	
	Column	C	GEO_DESC	□	☑	
	Column	D	GEO	□	□	
	Row	3	TIPO_DATI	□	□	
	Row	4	TIPO_DATI_DESC	□	☑	
	Row	5	MISURA	□	□	
	Row	6	TIME_PERIOD	☑	□	
	Value		OBS_VALUE	□	□	

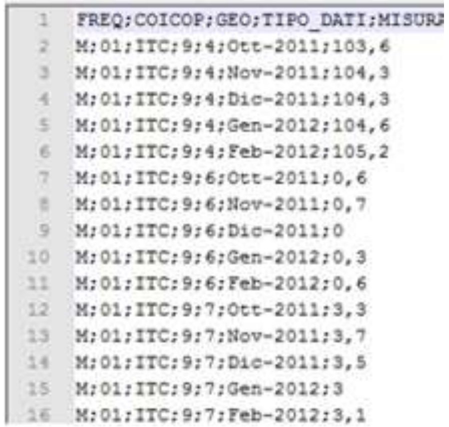
SDMX
Data
Structure Definition

Code List (CSV)



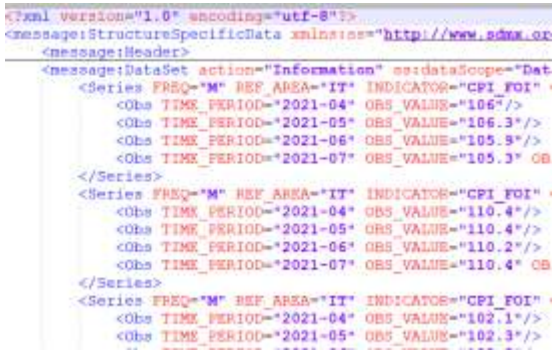
```
CODE;NAME
IT;Italy
ITC1;Piemonte
ITC11;Torino
ITC12;Vercelli
ITC13;Biella
ITC15;Novara
ITC16;Cuneo
ITC18;Alessandria
```

Data File (CSV)



```
FREQ;COICOP;GEO;TIPO_DATI;MISURA
M;01;ITC;9;4;Ott-2011;103,6
M;01;ITC;9;4;Nov-2011;104,3
M;01;ITC;9;4;Dic-2011;104,3
M;01;ITC;9;4;Gen-2012;104,6
M;01;ITC;9;4;Feb-2012;105,2
M;01;ITC;9;6;Ott-2011;0,6
M;01;ITC;9;6;Nov-2011;0,7
M;01;ITC;9;6;Dic-2011;0
M;01;ITC;9;6;Gen-2012;0,3
M;01;ITC;9;6;Feb-2012;0,6
M;01;ITC;9;7;Ott-2011;3,3
M;01;ITC;9;7;Nov-2011;3,7
M;01;ITC;9;7;Dic-2011;3,5
M;01;ITC;9;7;Gen-2012;3
M;01;ITC;9;7;Feb-2012;3,1
```

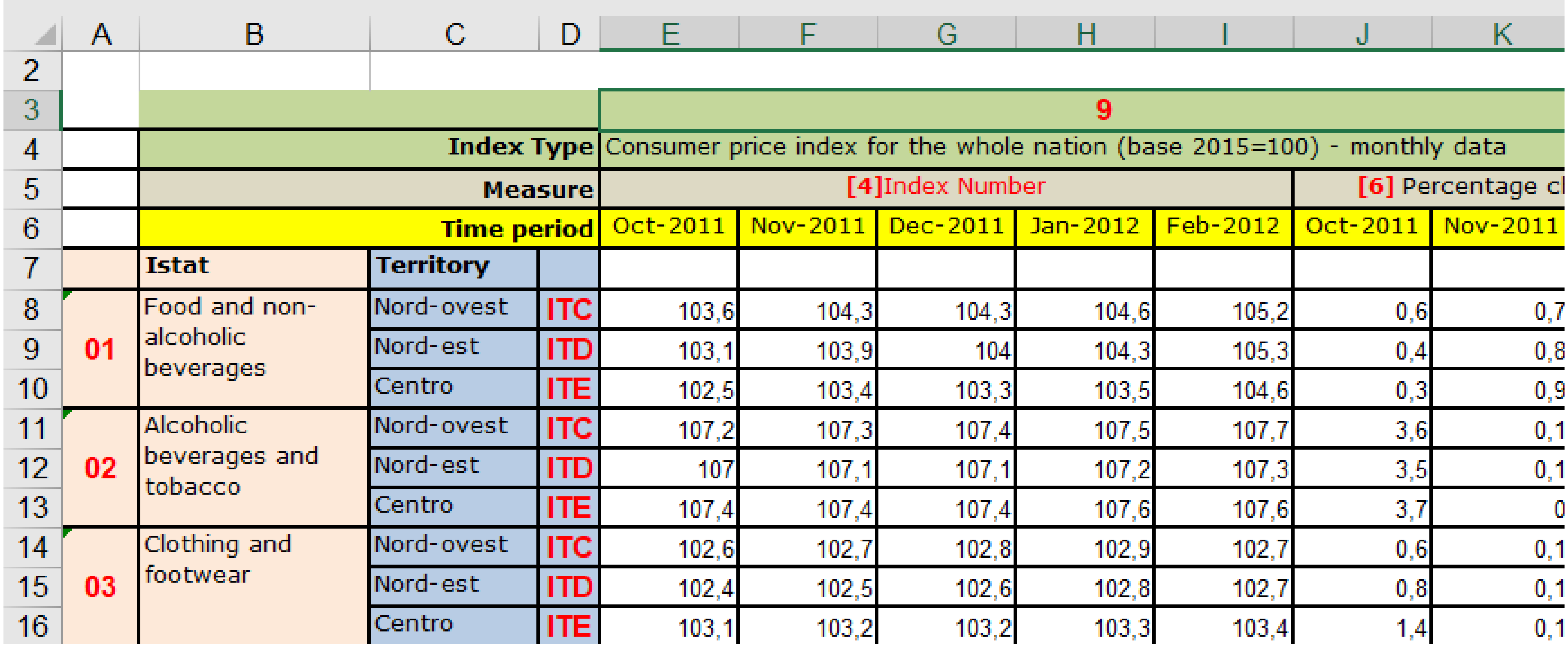
SDMX Data File



```
<?xml version="1.0" encoding="utf-8"?>
<message:StructureSpecificData xmlns:ss="http://www.sdmx.or
   <message:Header>
   <message:DataSet action="Information" ss:dataScope="Dat
      <Series FREQ="M" REF_AREA="IT" INDICATOR="CPI_FOI"
         <Obs TIME_PERIOD="2021-04" OBS_VALUE="106"/>
         <Obs TIME_PERIOD="2021-05" OBS_VALUE="106.3"/>
         <Obs TIME_PERIOD="2021-06" OBS_VALUE="105.9"/>
         <Obs TIME_PERIOD="2021-07" OBS_VALUE="105.3" OB
      </Series>
      <Series FREQ="M" REF_AREA="IT" INDICATOR="CPI_FOI"
         <Obs TIME_PERIOD="2021-04" OBS_VALUE="110.4"/>
         <Obs TIME_PERIOD="2021-05" OBS_VALUE="110.4"/>
         <Obs TIME_PERIOD="2021-06" OBS_VALUE="110.2"/>
         <Obs TIME_PERIOD="2021-07" OBS_VALUE="110.4" OB
      </Series>
      <Series FREQ="M" REF_AREA="IT" INDICATOR="CPI_FOI"
         <Obs TIME_PERIOD="2021-04" OBS_VALUE="102.1"/>
         <Obs TIME_PERIOD="2021-05" OBS_VALUE="102.3"/>
```

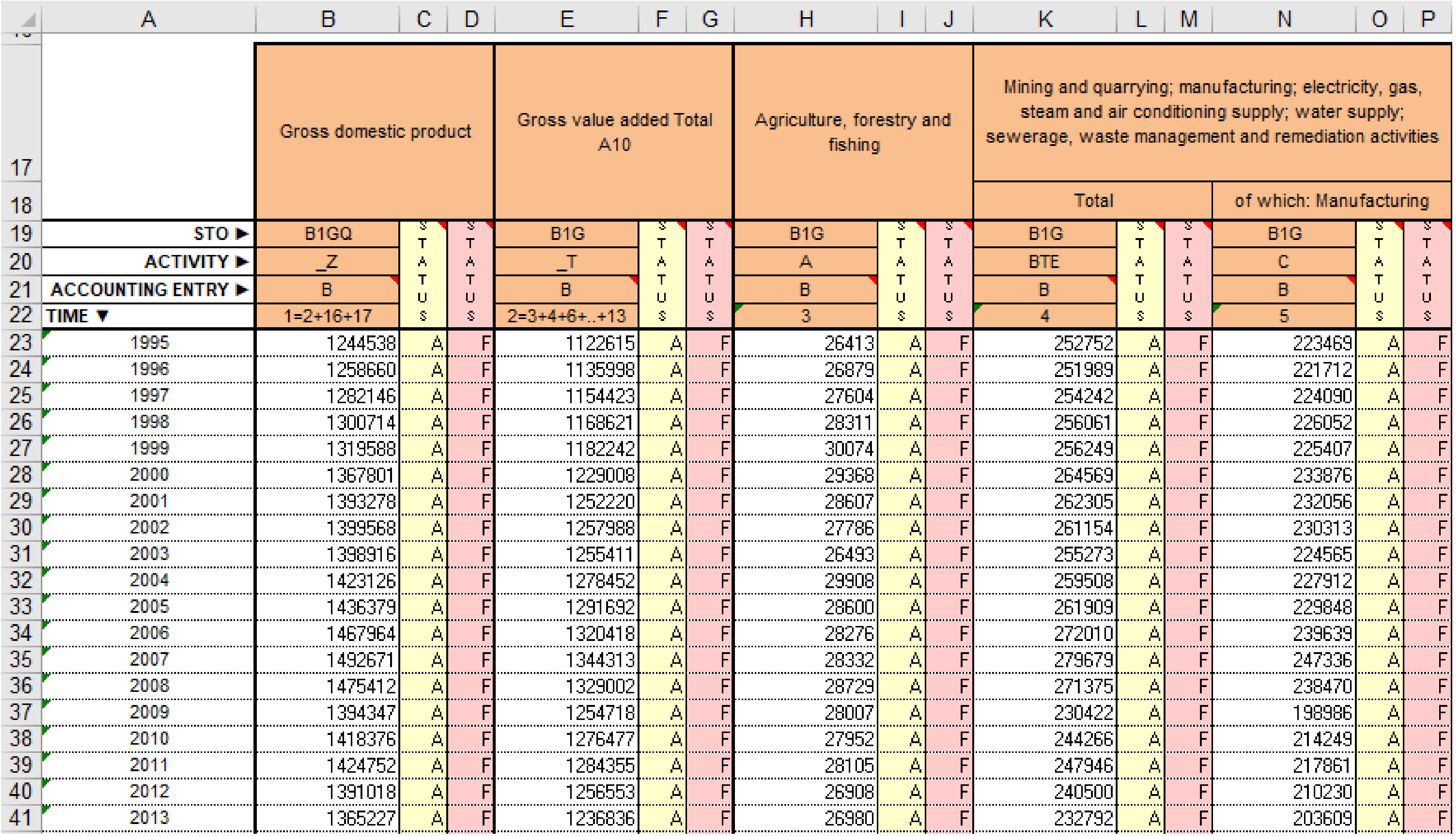
Istat

Excel2CSV tool – example of Excel multidimensional table



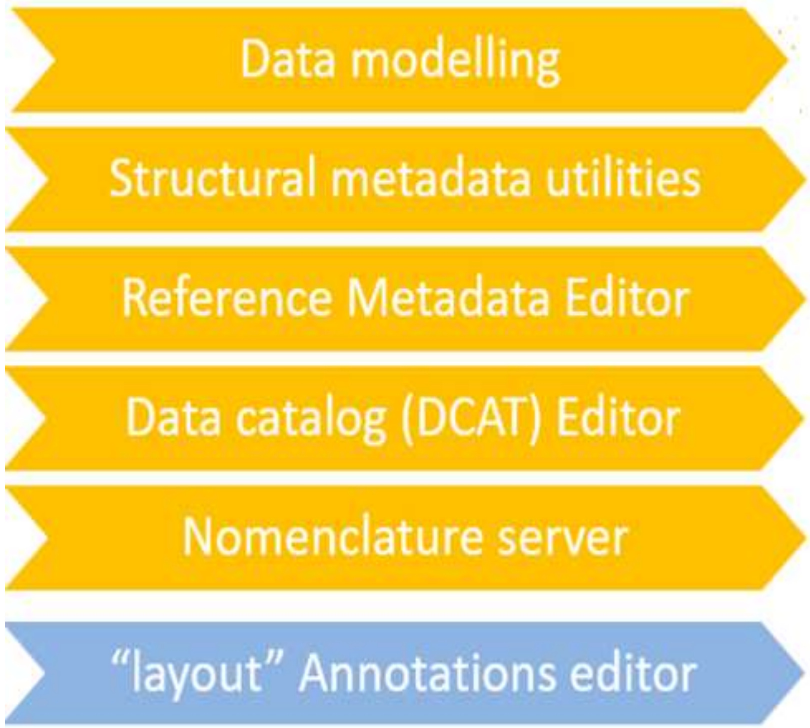
	A	B	C	D	E	F	G	H	I	J	K
2											
3					9						
4		Index Type			Consumer price index for the whole nation (base 2015=100) - monthly data						
5		Measure			[4]Index Number					[6] Percentage cl	
6		Time period			Oct-2011	Nov-2011	Dec-2011	Jan-2012	Feb-2012	Oct-2011	Nov-2011
7		Istat	Territory								
8	01	Food and non-alcoholic beverages	Nord-ovest	ITC	103,6	104,3	104,3	104,6	105,2	0,6	0,7
9			Nord-est	ITD	103,1	103,9	104	104,3	105,3	0,4	0,8
10			Centro	ITE	102,5	103,4	103,3	103,5	104,6	0,3	0,9
11	02	Alcoholic beverages and tobacco	Nord-ovest	ITC	107,2	107,3	107,4	107,5	107,7	3,6	0,1
12			Nord-est	ITD	107	107,1	107,1	107,2	107,3	3,5	0,1
13			Centro	ITE	107,4	107,4	107,4	107,6	107,6	3,7	0
14	03	Clothing and footwear	Nord-ovest	ITC	102,6	102,7	102,8	102,9	102,7	0,6	0,1
15			Nord-est	ITD	102,4	102,5	102,6	102,8	102,7	0,8	0,1
16			Centro	ITE	103,1	103,2	103,2	103,3	103,4	1,4	0,1

Excel2CSV tool – example of Excel multidimensional table



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
17		Gross domestic product			Gross value added Total A10			Agriculture, forestry and fishing			Mining and quarrying; manufacturing; electricity, gas, steam and air conditioning supply; water supply; sewerage, waste management and remediation activities					
18											Total			of which: Manufacturing		
19	STO ▶	B1GQ	STATUS	STATUS	B1G	STATUS	STATUS	B1G	STATUS	STATUS	B1G	STATUS	STATUS	B1G	STATUS	STATUS
20	ACTIVITY ▶	_Z			_T			A			BTE			C		
21	ACCOUNTING ENTRY ▶	B			B			B			B			B		
22	TIME ▼	1=2+16+17			2=3+4+6+..+13			3			4			5		
23	1995	1244538	A	F	1122615	A	F	26413	A	F	252752	A	F	223469	A	F
24	1996	1258660	A	F	1135998	A	F	26879	A	F	251989	A	F	221712	A	F
25	1997	1282146	A	F	1154423	A	F	27604	A	F	254242	A	F	224090	A	F
26	1998	1300714	A	F	1168621	A	F	28311	A	F	256061	A	F	226052	A	F
27	1999	1319588	A	F	1182242	A	F	30074	A	F	256249	A	F	225407	A	F
28	2000	1367801	A	F	1229008	A	F	29368	A	F	264569	A	F	233876	A	F
29	2001	1393278	A	F	1252220	A	F	28607	A	F	262305	A	F	232056	A	F
30	2002	1399568	A	F	1257988	A	F	27786	A	F	261154	A	F	230313	A	F
31	2003	1398916	A	F	1255411	A	F	26493	A	F	255273	A	F	224565	A	F
32	2004	1423126	A	F	1278452	A	F	29908	A	F	259508	A	F	227912	A	F
33	2005	1436379	A	F	1291692	A	F	28600	A	F	261909	A	F	229848	A	F
34	2006	1467964	A	F	1320418	A	F	28276	A	F	272010	A	F	239639	A	F
35	2007	1492671	A	F	1344313	A	F	28332	A	F	279679	A	F	247336	A	F
36	2008	1475412	A	F	1329002	A	F	28729	A	F	271375	A	F	238470	A	F
37	2009	1394347	A	F	1254718	A	F	28007	A	F	230422	A	F	198986	A	F
38	2010	1418376	A	F	1276477	A	F	27952	A	F	244266	A	F	214249	A	F
39	2011	1424752	A	F	1284355	A	F	28105	A	F	247946	A	F	217861	A	F
40	2012	1391018	A	F	1256553	A	F	26908	A	F	240500	A	F	210230	A	F
41	2013	1365227	A	F	1236836	A	F	26980	A	F	232792	A	F	203609	A	F

Meta Manager – main features



Data modelling

Structural metadata utilities

Reference Metadata Editor

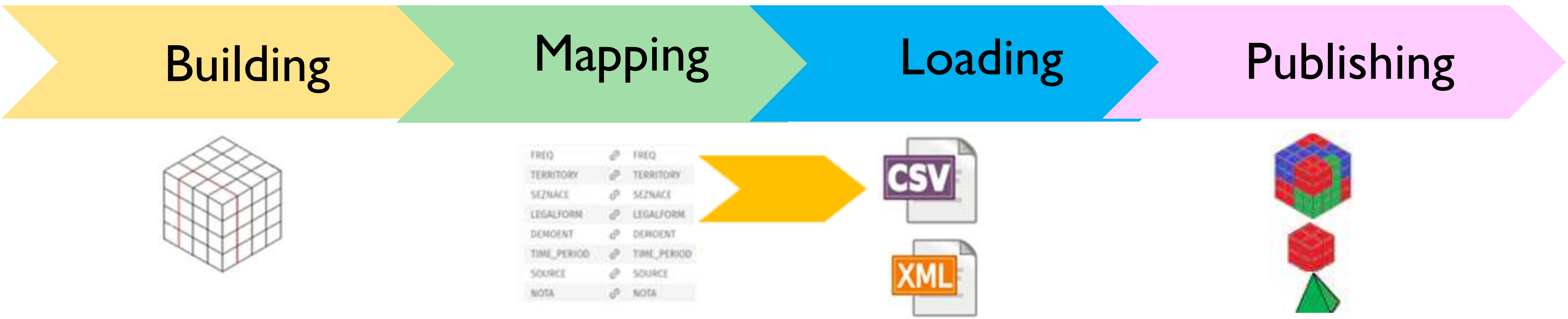
Data catalog (DCAT) Editor

Nomenclature server

“layout” Annotations editor

- □ **SDMX artefacts CRUD**: Concept Schemes, Code Lists, Concept Schemes, Category Schemes, Agency Schemes, Data Structure Definitions, Data Flows, Categorisations
  - ▪ Hierarchical CL, Metadata Structure Definitions, Metadata Flows can be used but not created
- □ **Structural metadata utilities**:
  - ▪ Derive item schemes (Code Lists, Concept Schemes, Category Schemes)
    - ✓ Add items preserving hierarchy
    - ✓ Include “Parents”, “Children”, “Descendants”
  - ▪ Merge item schemes (Code Lists, Concept Schemes, Category Schemes)
  - ▪ Compare item schemes (Code Lists, Concept Schemes, Category Schemes)
  - ▪ Compare DSDs
  - ▪ Upgrade DSD
- □ **Overcome SDMX 2.1 deficiencies**: add items to final item schemes, modify/add names, descriptions and annotations to final artefacts
- □ **Manage annotations** for presentation purpose (e.g. Order, Layout, etc.)
- □ **Reference metadata** editor
- □ **DCAT Data catalog** editor

Data Manager – main features



- **Building:** Multidimensional data cubes based on SDMX DSDs (advance database star-schema synchronized with the structural metadata repository)
- **Mapping:**
  - “Custom” CSV files
  - Excel files (multidimensional statistical tables)
- **Loading:** Excel, CSV, SDMX-ML data files
- **Publishing**: Data flows
  - As sub cube of a data cube
  - As sub cube with less dimensions (a new DSD is generated automatically)
- **Download:** SDMX-ML (2.0 & 2.1), SDMX-CSV, SDMX-JSON, Custom CSV, RDF/XML, DCAT-JSON

Data Browser – main features

A
D
M
I
N
I
S
T
R
A
T
I
O
N

NODES’ SWITCH

SEARCH

BROWSING

PRESENTATION

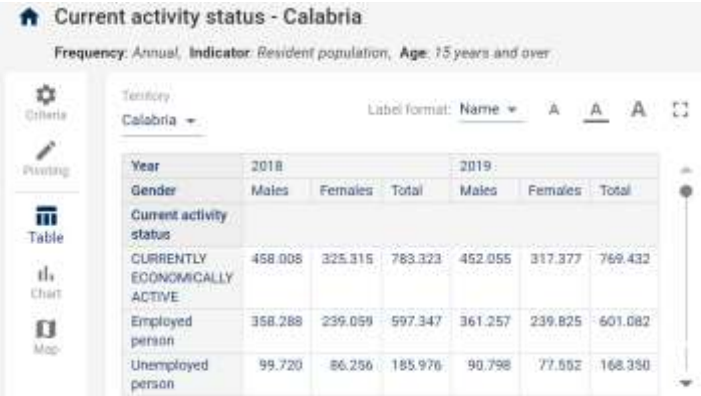
- TABLE
- CHART
- MAP

Criteria
Pivoting
Download

DASHBOARD

SHARING

LOGGED-IN USERS’ PORTFOLIO



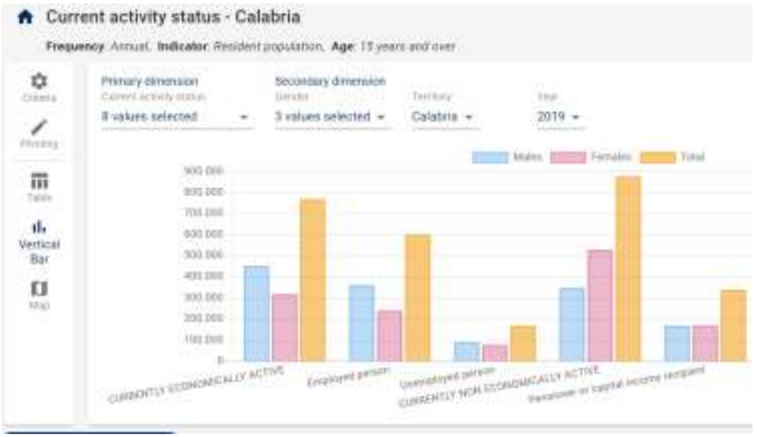
Current activity status - Calabria

Frequency: *Annual*, **Indicator**: *Resident population*, **Age**: *15 years and over*

Territory: Calabria

Label format: Name

Year	2018			2019		
Gender	Males	Females	Total	Males	Females	Total
Current activity status						
CURRENTLY ECONOMICALLY ACTIVE	458.008	325.315	783.323	452.055	317.377	769.432
Employed person	358.288	239.059	597.347	361.257	239.825	601.082
Unemployed person	99.720	86.256	185.976	90.798	77.552	168.350



Current activity status - Calabria

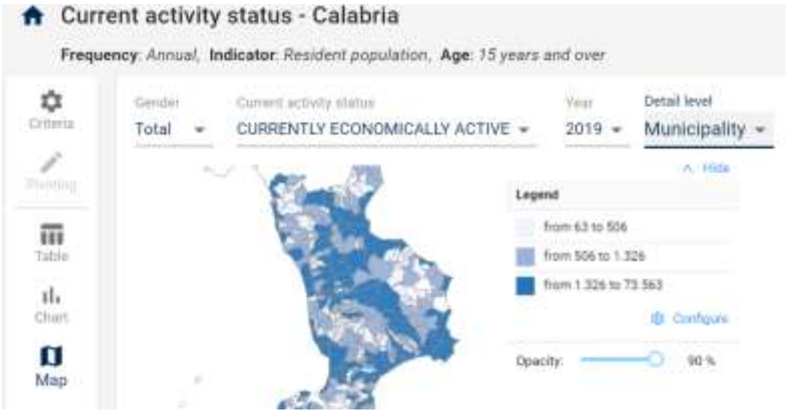
Frequency: *Annual*, **Indicator**: *Resident population*, **Age**: *15 years and over*

Primary dimension: Current activity status: 8 values selected

Secondary dimension: Gender: 3 values selected

Territory: Calabria

Year: 2019



Current activity status - Calabria

Frequency: *Annual*, **Indicator**: *Resident population*, **Age**: *15 years and over*

Gender: Total

Current activity status: CURRENTLY ECONOMICALLY ACTIVE

Year: 2019

Detail level: Municipality

Legend

- from 63 to 506
- from 506 to 1.326
- from 1.326 to 73.563

Opacity: 90 %

Lesson learnt

- Top management has to foster the initiative and must be involved regularly
  - The state of progress and bottlenecks must be reported in order to smooth the implementation (monitoring can be achieved by a cross-cutting working group)
- A suitable time must be devoted in analysing the experiences performed by other organisations
- Statistical standards must be the main ingredients of the innovation
  - No standards no industrialization
- “Wheel must not reinvented”
  - Reusing software available in the statistical community means save money and time
- A winning design is a good balance between innovation and integration
  - Step-by-step approach
- Capacity building actions facilitate the knowledge of the new methodologies and technologies

Contacts and links

- Meta and Data Manager Tool
- Data Browser Tool
- Excel2Csv Tool

Alessio Cardacino – alcardac@istat.it
Francesco Rizzo – rizzo@istat.it

- Cubus Tool

Luca Ramadori - luca.ramadori@istat.it
Guido Drovandi – drovandi@istat.it

Useful links:

SDMX Istat toolkit download: https://sdmxistattoolkit.github.io/index.html

“Permanent census of population and housing”: https://esploradati.censimentopopolazione.istat.it/databrowser/#/en

“Hub of the Public Statistics”: https://sistanhub.istat.it/databrowser/#/en

Thanks

Massimo FEDELI | fedeli@istat.it

Francesco RIZZO | rizzo@istat.it