SIS2017 Statistical Conference

Statistics and Data Science: new challenges, new generations

Florence 28-30 June

Emerging challenges in official statistics: new sources, methods and skills

Giorgio Alleva | President of the

Italian National Institute of Statistics - Istat

Challenges in the new eco-system of statistical information

- Measuring a more complex and diverse society
- Wealth of information, new unstructured sources
- Availability of new methodological and technological tools
- Crisis of traditional data collection systems
- More flexible, agile and cost efficient NSIs
- New competitors on the market

The outside world is changing rapidly

Istat’s modernisation programme

Paradigm shift in methodology

Multi-sources environment

New competencies

> *“Official statistical offices need to move*
> *from the probability sample survey paradigm*
> *of the past 75 years*
> *to a mixed data source paradigm*
> *for the future”*

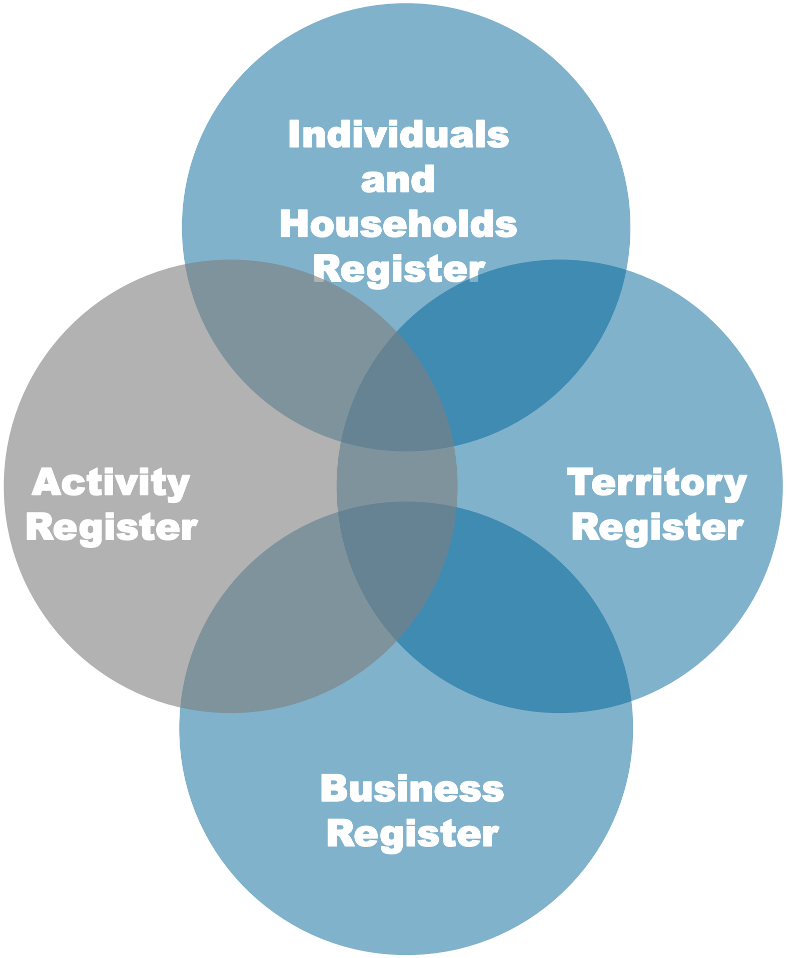
C. Citro (2014)

The Integrated System of Statistical Registers

Single logical data asset resulting from the **integration** of survey data, administrative data and new sources

Consistency in the **identification** and **estimation** of units and variables for the system as a whole

A "**system**", rather than a set, of registers, to connect people, businesses, places and their relations



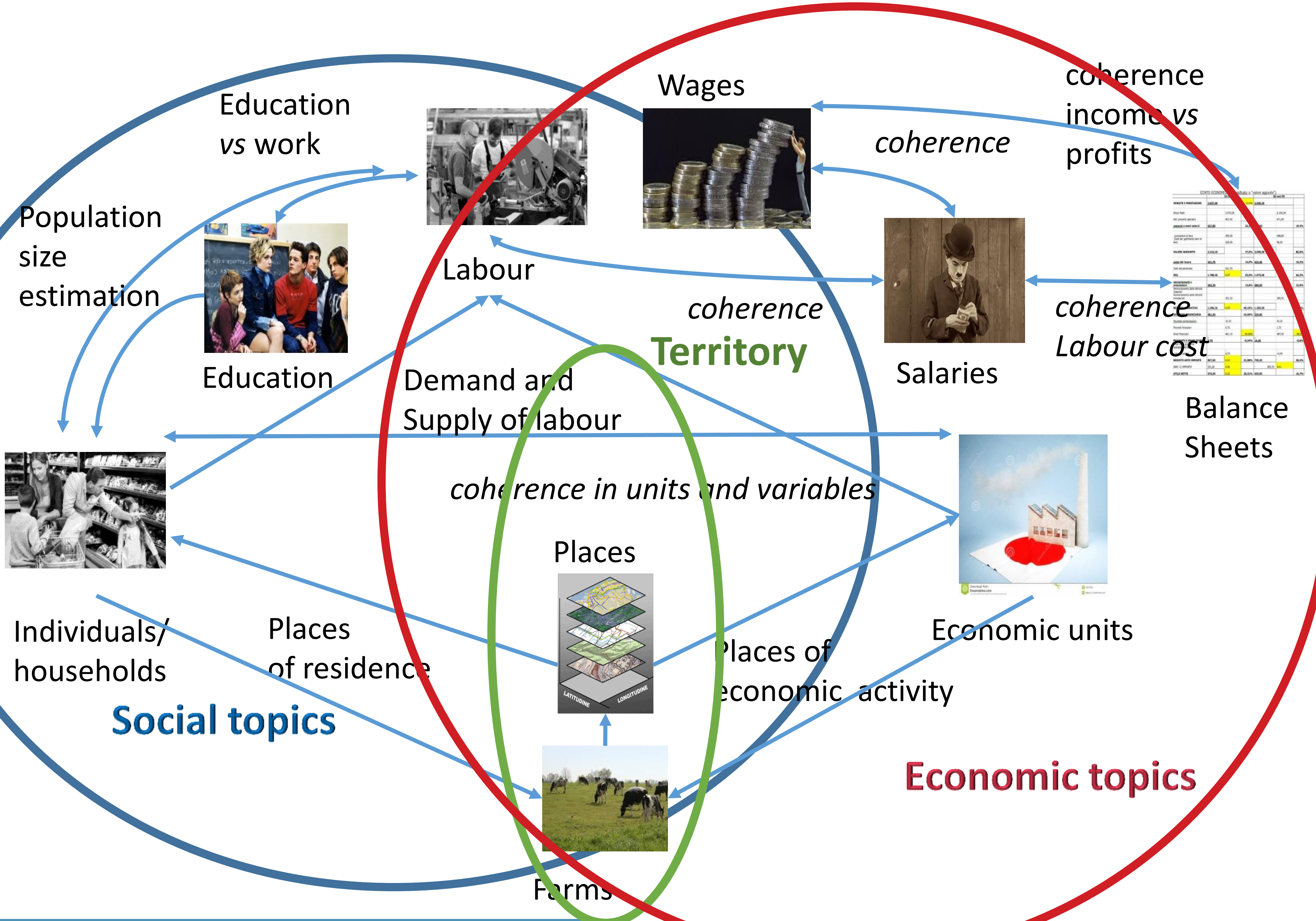
Methodological challenges

Some crucial methodological challenges to address

- data harmonisation (concepts, definitions, classifications)
- record linkage, statistical matching, micro-integration, modelling
- consistency of estimates from different sources
- how to deal with uncertainty

Shift in data collection: use already available sources

Same methods and generalized tools: harmonisation of processes



Education
vs work

Wages

coherence

coherence
income *vs*
profits

Population
size
estimation

Labour

Education

coherence

Territory

Salaries

coherence
Labour cost

Balance
Sheets

Demand and
Supply of labour

coherence in units and variables

Places

Individuals/
households

Places
of residence

Places of
economic activity

Economic units

Social topics

Economic topics

Farms

ISSR: the to be state

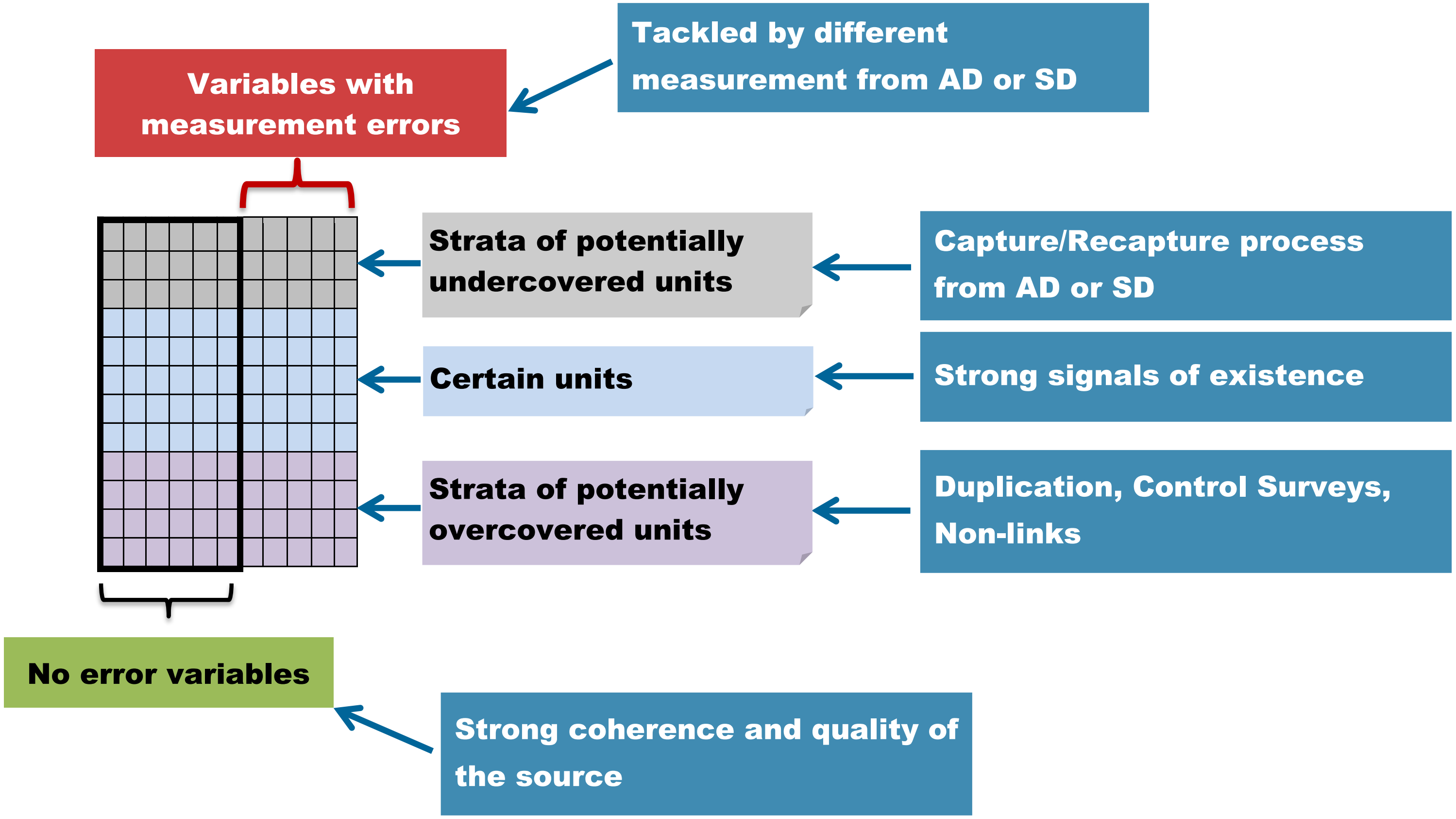
A new role for sample surveys in official statistics

Moving to a register-based system will generate a renewed role for sample surveys

In addition to their traditional role, surveys will be the key instruments for **specific purposes**:

- Observation of elusive or hard-to-reach populations not captured in the ISSRs
- Integrated survey framework
- Enhancing quality and contents of the ISSRs
- Evaluating the quality of new data sources

New role for the ISSRs



Tackled by different measurement from AD or SD

Variables with measurement errors

Strata of potentially undercovered units

Capture/Recapture process from AD or SD

Certain units

Strong signals of existence

Strata of potentially overcovered units

Duplication, Control Surveys, Non-links

No error variables

Strong coherence and quality of the source

AD = Administrative Data
SD = Survey Data

The Census and Social Surveys Integrated System (CSSIS)

The ISSRs will be the pillar for the permanent census, exploiting and integrating information from registers with data from a set of balanced and coordinated sample surveys (Master sample, MS)

The first phase of the MS design

Planned to be held, yearly, in Autumn (starting from 2018), it aims at:

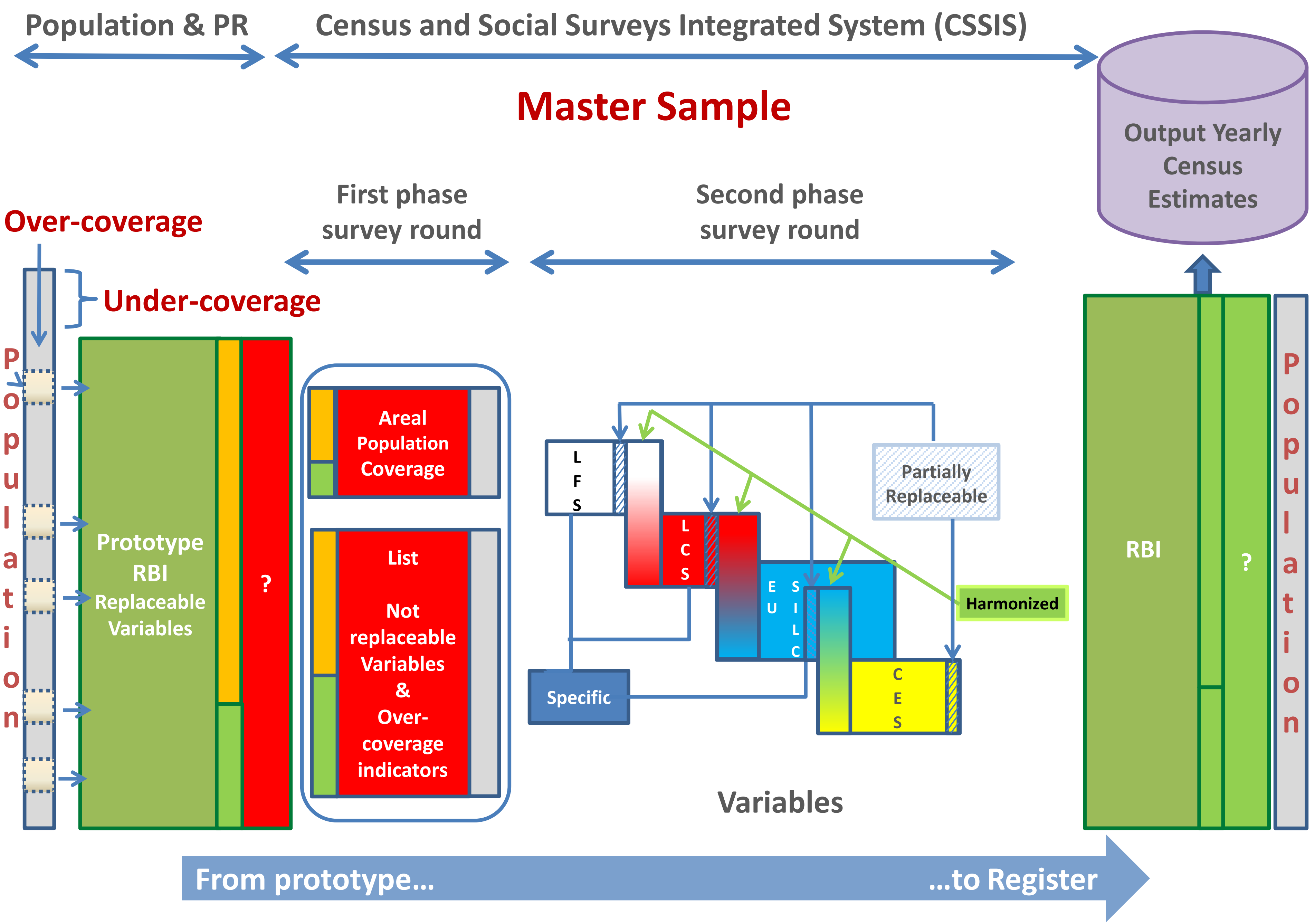
- correcting for under and over coverage the Base Register of individuals improving the quality of the population totals produced;
- collecting the information for not replaceable variables by means of an ad hoc sample survey (Master Sample)

Two different schemes: one based on an **areal sample** (A) and one based on a **list sample** (L).

The second phase of Ms

The year following the first phase (i.e. from January 2019), sample households are selected as a sub-sample of those already involved in the first phase sample

The Census and Social Surveys Integrated System



Population & PR

Census and Social Surveys Integrated System (CSSIS)

Master Sample

Output Yearly Census Estimates

Over-coverage

Under-coverage

First phase survey round

Second phase survey round

Population

Prototype RBI Replaceable Variables

?

Areal Population Coverage

List

Not replaceable Variables & Over-coverage indicators

LFS

LCS

EU SILC

CES

Partially Replaceable

Harmonized

Specific

Variables

RBI

?

Population

From prototype…

…to Register

New sources: Big Data

Opportunity to produce timely high-quality statistics with greater detail

Big Data: open issues

① data access

② quality

③ methodology

④ legal framework

⑤ skills and competences

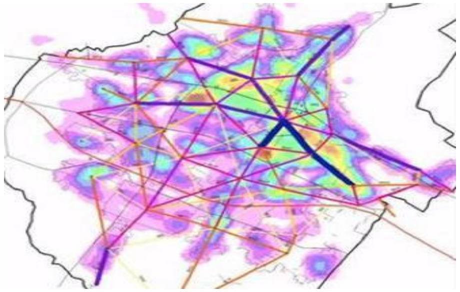
https://databigandsmalldotcom.files.wordpress.com/2015/02/bigdata.jpg

Big Data use: examples

Scanner data & web scraping

8 032089 000017
PREFISSO AZIENDALE GS1
CODICE PRODOTTO
CIFRA DI CONTROLLO

Mobile phone data



Web scraping & Text mining

Sensors

Methodological research at Istat

Deep investment on methodological and thematic research

Balancing the **independence** of research and its **relevance** for responding to the effective needs of production is crucial for NSIs (Fellegi, 2010)

Istat has recently set up some **infrastructures** for managing research proposals

- □ **Three-year plan for thematic and methodological research**
- □ **Innovation Lab**

Launch of a **Call for ideas**

Methodological research questions

① How to **integrate new and old sources** and to **increase the effectiveness of direct surveys?**

② How to ensure the necessary **cross-cutting and longitudinal consistency** of register-based estimates?

③ How to measure the **quality of a big data founded statistics,** or, in other words, how must the classic inferential statistical approaches (design, model, or Bayesian) be modified to make the construction of robust inferences possible from the new databases?

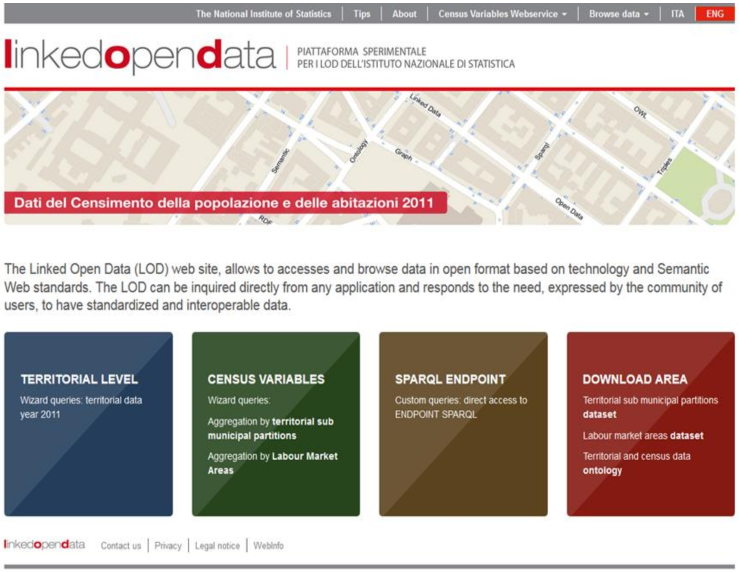
④ What is the necessary **technological/information architecture** to make the best of the new data bases?

⑤ How to modify the traditional approach to **quality evaluation in a multi-source environment**?

Istat’s Linked Open Data portal

Open data is a key enabler of data-driven innovation

The portal is the **single access point** to Istat’s open data and part of the Italian national data cloud



The National Institute of Statistics | Tips | About | Census Variables Webservice | Browse data | ITA | ENG
linkedopendata
PIATTAFORMA SPERIMENTALE
PER I LOD DELL’ISTITUTO NAZIONALE DI STATISTICA
Linked Data
OWL
Ontology
Graph
Semantic
Sparql
Triples
Open Data
RDF
linkedopendata
Contact us | Privacy | Legal notice | WebInfo

Dati del Censimento della popolazione e delle abitazioni 2011

The Linked Open Data (LOD) web site, allows to accesses and browse data in open format based on technology and Semantic Web standards. The LOD can be inquired directly from any application and responds to the need, expressed by the community of users, to have standardized and interoperable data.

TERRITORIAL LEVEL
Wizard queries: territorial data year 2011

CENSUS VARIABLES
Wizard queries:
Aggregation by **territorial sub municipal partitions**
Aggregation by **Labour Market Areas**

SPARQL ENDPOINT
Custom queries: direct access to ENDPOINT SPARQL

DOWNLOAD AREA
Territorial sub municipal partitions **dataset**
Labour market areas **dataset**
Territorial and census data **ontology**

Main features

- machine-to-machine data
- access at the finest granularity level
- flexible querying
- advanced navigation mechanisms
- direct access to data via Web Services

Conclusions. Key concepts

Statistics as a valuable public good

Modernisation to produce high quality data

Multiple use of data sources: integration

Improve data release for all users

Innovation and Research

Skills and competencies

Change driven culture

Emerging challenges in official statistics: new sources, methods and skills

Giorgio Alleva | President, Istat

How to deal with uncertainty

A strategic issue for NSIs (responsibility and transparency)

1. **Simply ignore** (traditional solution): simple but risk of severe bias.
2. **Evaluate** the sources of errors in order to inform the users (Eg. PES): lack of consistency of different production lines, 2 lines of production.
3. **Identify the improvements** in the process for building the registers: continuous improvement/ the identified bias is still present.
4. **Correct the bias** in the main estimates (External Benchmarks) without modifying the register: lack of consistency of different production lines, 2 lines of production.
5. **Modify units and variables in the register to correct the bias** in the main estimates: consistency of different outputs, relevant computable efforts, some outputs may be inaccurate (transfer the uncertainty to the microdata level).