Multiple Sources in Business Surveys

Paulo Saraiva
paulo.saraiva@ine.pt

Almiro Moreira
almiro.moreira@ine.pt

Sofia Rodrigues
sofia.rodrigues@ine.pt
Statistics Portugal and Administrative Sources

- Already used in several areas of official Portuguese statistical production:
  
  - +100 sources benefiting 80 statistical activities.
    
    - Business and Social statistics

  - Data Warehouse with 37 regular flows and 1.5 billion historical observations.
Tax and Customs Authority as one major Data Source

- Information from Registrations;
- Personal and Corporate Income Taxes;
- Value Added Tax (VAT) and VAT Information Exchange System;
- Custom and Excise duties;
- Local government taxes and revenues;
- Electronic invoices data.
Mixed-sources models

- The **Split Data Approach**:  
  - questionnaire only,
  - questionnaire supplemented by administrative register,
  - administrative data supplemented by questionnaire,
  - administrative data only

- **Pre-filled** questionnaires: respondents being asked to merely check and correct

- Using secondary sources for **non-responders** and for **estimation**
Integrated Metadata

- An integrated metadata system is a corner stone of a multi-source production system.
- Interconnecting metadata between internal production systems and processes is a must to produce integrated statistics.
- This is however a difficult harmonisation process.
Metadata System

http://smi.ine.pt

**Concepts**
Module that integrates and provides terms and definitions of concepts used for statistical purposes on the National Statistical System, which underlies the consistency and comparability of the data collected and disseminated. Read more...

<table>
<thead>
<tr>
<th>Code</th>
<th>Designation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2051</td>
<td>ECONOMIC ACTIVITY</td>
</tr>
</tbody>
</table>

**Variables**
Module that integrates and provides the observed variables and statistical indicators disseminated by the National Statistical System. Read more...

<table>
<thead>
<tr>
<th>Code</th>
<th>Designation</th>
</tr>
</thead>
<tbody>
<tr>
<td>10020</td>
<td>Compensation (P) of Enterprises by Ec...</td>
</tr>
<tr>
<td>1993</td>
<td>Live births (No.) by Place of residence c...</td>
</tr>
</tbody>
</table>

**Classifications**
Module that integrates and provides national, Community and international classifications used for statistical purposes on the National Statistical System. Read more...

<table>
<thead>
<tr>
<th>Code</th>
<th>Designation</th>
<th>Acronym</th>
</tr>
</thead>
<tbody>
<tr>
<td>V03004</td>
<td>Portuguese classification of economic activities, revision 3</td>
<td>CAE Rev 3</td>
</tr>
<tr>
<td>V03017</td>
<td>Code of the administrative division</td>
<td></td>
</tr>
<tr>
<td>V03503</td>
<td>Nomenclature of territorial units for statistics, 2013 version</td>
<td>NUTS 2013</td>
</tr>
</tbody>
</table>

**Methodological Documentation**
This module includes and provides access to documents (called Methodological Documents) that describe, at present time, each statistical operation undertaken under the National Statistical System. It respects a common structure with predefined topics. Read more...

<table>
<thead>
<tr>
<th>Code</th>
<th>Designation</th>
</tr>
</thead>
<tbody>
<tr>
<td>270</td>
<td>Survey on ICT usage in enterprises</td>
</tr>
</tbody>
</table>

**Data Collection Instruments**
This module includes and provides information concerning data collection instruments registered at INE, and access to their image. Whether they may be questionnaires (when connected to surveys) or files (in electronic format). Read more...

<table>
<thead>
<tr>
<th>Code</th>
<th>Designation</th>
</tr>
</thead>
<tbody>
<tr>
<td>10003</td>
<td>INQUIR-demais-1997-1999</td>
</tr>
</tbody>
</table>
Towards a data-centric perspective
Production chain: changing perspective

- Statistical organizations seek to produce high quality data that is supported by the metadata needed to make this data useful.
- It is often assumed that statistical production follows a certain chain of processes.
- It begins by collecting data, followed by processing, analysis and dissemination of statistical products.
- This is an inefficient approach to the present day as it consumes a lot of time and resources.
- Another perspective is needed.
Data-centric production

- We can look at statistical production from a data-centric point of view;
- Focus on the flow and the transformation of the data during the production process;
- Data Resources should be understood as an organizational asset, which must be managed and explored intensively;
- Reusing data is certainly a way to make statistical production more efficient, both for existing products and for responding to new needs, at a lower cost and much faster.
- Extending the use of mixed-sources is a way to maximize the reuse of data.
National Data Infrastructure: a data-centric approach

MAIN GOALS

SECURITY AND QUALITY

DATA INTEGRATION

STATISTICAL PROCESSES

FLEXIBLE AND INNOVATIVE

ADRESSES NATIONAL DATABASE

SHARING ALONG THE STATISTICAL PRODUCTION CHAIN

SINGLE ENTRY POINT

TRUST

DEVELOPMENT OF STRATEGIES

PUBLIC WELL BEING: STATISTICAL DATA

PUBLIC DECISION-MAKING

REGISTRY INFORMATION SYSTEM
National Data Infrastructure: Challenges

- More intensive and integrated use of data
- Gain resources and time to intensify innovation
- Substantial enlargement of the domains covered
- Learn the multidimensionality of data
- New techniques of data extraction and analysis
- To profit from the production chain
- Intensification of appropriation of administrative data and other sources
- Increase feedback to society
National Data Infrastructure

**TECHNOLOGICAL INFRASTRUCTURE AND HUMAN RESOURCES**

- **HUMAN RESOURCES**
- **MULTIDISCIPLINARY TEAMS OF DATA SCIENTISTS**
- **THEMATIC SPECIALISTS**
- **DATA ANONYMIZATION SPECIALISTS**
- **STATISTICS PORTUGAL**

**NATIONAL DATA INFRASTRUCTURE**

**TECHNOLOGICAL INFRASTRUCTURE**
New perspective: DataLab and DataFactory

Potential Sources

- Data Exploration
- Data Knowledge
- Specify Needs
- Check data availability
- Data Catalogue

Existing Sources

Data Processes

DataLab

DataFactory
New perspective: DataLab and DataFactory
DataLab

Potential Sources

• Data Exploration
• Data Knowledge
• Specify Needs
• Check data availability
• Data Catalogue

DataLab
Data sources with **potential use** (or already in use)

- **Potential use** (Annual Activity Plan)
  - can be accessed in the **medium term**
  - can be accessed in the **long term**

- **Already in use**

- **Type of use**
  - Legal act needs to be amended
  - Survey replacement total or partial
  - Validation Calibration
  - Statistical units registers update

---

**Data sources**
Data access and data reuse barriers

- Statistical Units
- Access Stability
- Variables
- Classifications
- Timeless
- Inconsistencies between sources
- Missing Data
- Internal Resistance
- Concepts
Administrative data appropriation process

1. Identify **legal changes**
2. Clarify **requirements for administrative sources**
3. Assure an **inventory of sources**
4. Develop **Provision agreement models**
5. Participate in the **design of external data sources**

**Enhance the catalog of sources**
Data Catalogue

- Checking **current data sources** could meet user requirements, and the conditions to be available, including any restrictions on their use.

- An **assessment of possible alternatives**, including research into potential administrative or other non-statistical data sources, to determine whether they would be suitable for use for statistical purposes.

- When existing sources have been assessed, a strategy for filling any remaining gaps in the data requirement is prepared.

- Should include an assessment of the **legal framework** in which data would be collected and used, and identify proposals for changes to existing legislation or the introduction of a new legal framework.
Challenge: Administrative Data Catalogue

- **Source**
  - Provider
  - Technical Framework
  - Owner
  - Government Tutelage
  - Depositary Institution

- **Data Set**
  - Data Description
  - Quality Indicators
  - Exchange Channel

- **Flow**
  - Provision Agreement
  - Legal Framework
  - Confidentiality Rules
  - Contingences

- **Variable**
  - Variable Definition
  - Concepts
  - Statistical Units
  - Classification System

- **Agents**
  - Agent Role
  - Access Rules
  - Access Tracking

- **Existing Use**
  - Historical Coverage
  - Missing Data
  - Time Instances

- **Potential**
  - Transformation Level
  - Conversion Rules
DataFactory

Existing Sources

Data Processes

DataFactory
Administrative Sources – Processes

[Diagram showing the process flow from Administrative Data through Capture, Process & Analyse, and finally to Data, withMulti-surveys indicated as the output.]
Source Integration

SURVEY DATA

ADMINISTRATIVE DATA

INTEGRATION

PARADATA

METADATA

REGISTERS

MICRODATA

Analysis, checking errors, data edition and imputation
Workshop on Statistical Data Collection 2019
14-16 October 2019, Geneva

Multiple Sources in Business Surveys

Paulo Saraiva
paulo.saraiva@ine.pt

Almiro Moreira
almiro.moreira@ine.pt

Sofia Rodrigues
sofia.rodrigues@ine.pt