Business Data Collection:
From questionnaires to data sets

Paulo Saraiva
paulo.saraiva@ine.pt
What are we going to talk about?

- Business Data Collection: Where we are?
- Discuss automated business data interchange.
- Sharing the experience of Portugal in this field.
- Pointing-out some challenges.
Business Data Collection: Toward Electronic Data Interchange. Experiences in Portugal, Canada, Sweden, and the Netherlands with EDI

Gert Buiten¹, Ger Snijkers², Paulo Saraiva³, Johan Erikson⁴, Anna-Greta Erikson⁴, and Alice Born⁵
Business Statistics: Production Cycle

Sources

SURVEYS

Administrative Data

COLLECT

PROCESS & ANALYSE

DISSEMINATE

REGISTERS
Let’s go ‘all-electronic’

- Statistical Producers move to web-based data collection.
- In 2005, Statistics Portugal introduced integrated web questionnaires as an option for business surveys.
- In Portugal, like the other countries, all business surveys use web questionnaires nowadays.
- 98% of the collected questionnaires (2017)
Registers and administrative data were being used.
Replacing surveys, supplementing data, validation or imputation.

Another secondary data source gets attention is big data (since 2013).

We are working towards both the extended use of secondary sources (including big data) as well as of web surveys.

Multi-source approaches combining primary and secondary data.
This process is still going on.
The way to Integration

- **Conceptual harmonisation:**
  - Eliminating differences in definitions, concepts, classifications, variables...
  - Sharing same metadata and taxonomies.
  - Both inside (horizontal) and outside (vertical) the organization.

- **Technical standardisation:**
  - Uniform data formats, file structures, transmission modes, software
Harmonisation: Metadata

- Metadata provide information on data, and describe them: they are the data definitions (UNECE 1995).
- Essential in an integrated statistical process.
- Metadata support the whole statistical production.
- Concepts, classifications, variables, data sources, and outputs.
- They are as important as data.
- A core tool to a Statistical Producer.
Portugal: Integrated Metadata System

- Built in 2003, it covers all Portuguese Statistical System
- Repository of harmonized concepts, classifications, variables, data collection instruments and methodological documentation.
- Its purpose is to support:
  - Survey design;
  - Data outputs.

http://smi.ine.pt
Portugal: Integrated Survey Management System

- Statistics Portugal, in 2005 a central data collection department was established.
- First SIGINQ covered the business surveys, and followed by the agriculture and household surveys.
Portugal: Integrated Survey Management System

- Integration in this system is based on four pillars:
  - processes, metadata, registers, and variable-oriented.
Following the principle of maximizing the reuse of components, the system combines generic functions with survey-specific features.

Some components are shared by all surveys (business or household), and others are specific for a single survey.

Sometimes, a specific component is adapted to become a generic one, if it has a certain potential.
Production System
# Reuse of components

<table>
<thead>
<tr>
<th>Survey</th>
<th>Unit</th>
<th>Occur</th>
<th>Collect</th>
<th>Report</th>
<th>Analysis</th>
<th>Update</th>
<th>Manag.</th>
<th>Help</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>SIGUA block prop</td>
<td>Supplement Launch</td>
<td>Errors</td>
<td>Specific Tables</td>
<td>Consult transfers</td>
<td>Batch update</td>
<td>Table Manag.</td>
<td>GPap</td>
</tr>
<tr>
<td>Specific</td>
<td>Manage</td>
<td>By mode</td>
<td>Status</td>
<td>Generic Tables</td>
<td>Transfer to analysis</td>
<td>Survey</td>
<td>Register</td>
<td>Sample</td>
</tr>
<tr>
<td>Generic</td>
<td>Cross</td>
<td>Consult</td>
<td>Validations</td>
<td>Specific Reports</td>
<td>Consult Analysis</td>
<td></td>
<td></td>
<td>Respondent</td>
</tr>
<tr>
<td>Specific</td>
<td>Open / Close</td>
<td></td>
<td>Primary Val</td>
<td>Generic Reports</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Upload</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Insert</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Manage entries</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Data Entry</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Generic**
- **Survey-specific**
In many countries, technical standards are not regulated by law. Portugal chose for another solution. In 2007: Simplified Business Information system (IES) – legal act. A mandatory nation-wide initiative to enable S2S financial reporting between businesses and government agencies.
Statistics Portugal’s Automated Data Transmission

- In addition to web questionnaires, businesses may also use another data collection mode: Automated Data Transmission (TAD).
- It is a S2S approach with the upload of XML files, or a dedicated Web service.
- TAD allows businesses respondents to prepare an XML envelope, with variables from one or more companies and surveys, avoiding the completion of e-questionnaires.
- After uploading the file, it integrates the variables accordingly, which can be accessed by the respondent or by data collection staff.
- Currently available for 14 surveys.
After requesting to use this service, a Globally Unique Identifier (GUID) code is generated to the respondent, which provides access to the TAD service.

Technical guides are available.

Formats and instructions on how to integrate in external applications,

and the use of web services for an automatic and unattended solution.
Technical Guides

- GENERIC GUIDE
- SURVEY GUIDE
- INTEGRATION GUIDE

- VALIDATION RULES
- VARIABLES
- Examples

Test area
Basic XML Structure

```xml
<?xml version="1.0" encoding="utf-8" ?>
<ine xmlns="http://ine.pt/public/xsd/forms"
     xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  <header chave="...">
    ...
  </header>
  <respostas>
    <resposta>
      <referencia>
        ...
      </referencia>
      <dados tipo="..." ult_actualizacao="YYYY-MM-DDTHH:MI:SS">
        ...
      </dados>
      <info>
        ...
      </info>
    </resposta>
  </respostas>
</ine>
```
### Variables description for a Survey

<table>
<thead>
<tr>
<th>Variable ID</th>
<th>Description</th>
<th>Type</th>
<th>Format</th>
<th>Domain</th>
<th>Segment</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC101</td>
<td>Economic Activity</td>
<td>num</td>
<td>5</td>
<td>list:CAE</td>
<td>d</td>
</tr>
<tr>
<td>BC105</td>
<td>Activity Status</td>
<td>num</td>
<td>2</td>
<td>list:STA</td>
<td>d</td>
</tr>
<tr>
<td>BC151</td>
<td>Sazonality?</td>
<td>num</td>
<td>1</td>
<td>list:yes_no</td>
<td>d</td>
</tr>
<tr>
<td>BC153</td>
<td>Establishment Category</td>
<td>txt</td>
<td>2</td>
<td>list:categories</td>
<td>d</td>
</tr>
<tr>
<td>BC156</td>
<td>Beach</td>
<td>num</td>
<td>8</td>
<td>lista:beaches</td>
<td>d</td>
</tr>
<tr>
<td>V100</td>
<td>Number of available rooms</td>
<td>num</td>
<td>5</td>
<td>&gt;=0</td>
<td>c</td>
</tr>
<tr>
<td>V101</td>
<td>Total of single beds</td>
<td>num</td>
<td>5</td>
<td>&gt;=0</td>
<td>c</td>
</tr>
<tr>
<td>V102</td>
<td>Total of double beds</td>
<td>num</td>
<td>5</td>
<td>&gt;=0</td>
<td>c</td>
</tr>
<tr>
<td>V103</td>
<td>Total of occupied rooms</td>
<td>num</td>
<td>6</td>
<td>&gt;=0</td>
<td>a</td>
</tr>
</tbody>
</table>
Integration: Application Guide
TAD acceptance and the way forward

- 24% of the collected questionnaires (WebInq, 2017)
- There has been an increasing acceptance, indicating a significant reduction of the resources involved, as well as increased quality.
Multiple ways to create, view and modify

Questionnaire Level

Variable Level

CREATE, VIEW, MODIFY

CREATE, MODIFY

PDF

XML

webinq
Feeding back data to businesses in Portugal

- Statistics Portugal evaluated possibilities of including information flows from the users back to the suppliers.
- Providing feedback and benchmark data is an incentive for businesses to participate in surveys, even if mandatory.
- It may serve as a driver for businesses, their service providers and software developers to adopt S2S.
- Such ‘feedback data’ can be used by them to provide better and even new services to their customers.
Challenges
Challenges: Methodological Approaches

- Sampling has been an efficient way to collect data from a population of units.
- **Is sampling still efficient within an S2S approach?**
- Maybe a census is much more efficient, considering the required investments by businesses, software developers.
- This applies to the IES system, which includes all businesses in Portugal that by law need to report financial statements to the government.
Challenge: Source Integration

MICRODATA
SURVEY DATA
INTEGRATION
PARADATA
METADATA
REGISTERS
MICRODATA
Analysis, validation, edition
ADMINISTRATIVE DATA
Challenge: Administrative Sources – Processes

Administrative Data

Capture

Process & Analyse

Registers

Metadata

Data

Multi-surveys
Challenge: Data Integration

SOURCE 1

V1 V2
C1 × ×
C2 ×
C3 ×
C4

SOURCE 2

V1 V2
C1 × ×
C2 × ×
C3 × ×
C4 × ×

SURVEY

V1 V2
C1 × ×
C2 × ×
C3 × ×
C4 ? ×

Combination Rules
Challenge: Unified Unit Registers

INTEGRATION

UNIFIED UNIT REGISTERS

BUSINESS

AGRICULTURAL HOLDING

HOUSEHOLD

OTHERS
Challenge: Reengineering Surveys

Proliferation of questionnaires and surveys

One Data Set for several Surveys

REENGINEER SURVEYS

S2S
Business Data Collection: From questionnaires to data sets

Thank you!

Paulo Saraiva
paulo.saraiva@ine.pt