Statistical business registers: IT considerations

V. Todorov[1]

[1]United Nations Industrial Development Organization, Vienna

Regional Workshop on the Statistical Business registers for the Arab Countries
26-29 September 2016
Amman, Jordan

Outline

Outline

Introduction

IT infrastructure and programming requirements for the build phase of an SBR system

- When establishing any system, there are many possible technologies.
- Of course this is valid for SBR too.
- The choice should take into account
  - scalability,
  - cost and
  - maintenance.
- The technology should be flexible enough to evolve with new requirements.

Outline

General considerations

- There is no international standard or even commonly used practice amongst NSIs regarding the design of an SBR system per se.
- The main consideration is to develop an SBR system that
  - fits within the NSI IT architecture and
  - it is as compatible as possible with other systems like the administrative data acquisition systems and the business survey collection systems.

General considerations

- Project management methodology
- Software development methodology
- Solution architecture: functional and non-functional requirements; layered architecture
- Database: relational database management system (RDBMS)
- Frame, collection and respondent burden modules
- Documentation

General considerations:Layered architecture

Simple layered architecture

Presentation layer	Implements the user interface and manages user interaction with the system.
Service layer	Exposes interfaces and system functionality to other systems, and may also be the boundary between the presentation and business layers.
Business layer	Implements the core functionality and business logic.
Data access layer	Implements access to and interaction with data stores.

Outline

Database Management System (DBMS) options

- Database Management System (DBMS)
  - ORACLE or MS SQL Server
  - MySql
  - SQLite
  - MS ACCESS
  - Other more exotic ones: MaxDB

Outline

Programming requirements

- Graphical user interface (GUI)
- Programming language
  - .Net and C#
  - Java
  - Python
  - R, SAS

Programming requirements

- IT Environments: a full scale solution may involve five distinct environments
  1. Production environment
  2. Practice (training) environment
  3. User acceptance environment
  4. Development environment
  5. Analysis environment
- The first and fifth environments are essential; all forms of testing and analysis can take place in environment 5.

Programming requirements

Establishing a unique identifier for statistical units

- Essential for accurate maintenance of the SBR.
- Sequential assignment of unique identifiers, managed centrally.
- Reduces the risk of inadvertently disclosing confidential micro-data by use of easily identifiable and recognizable information.

Establishing a unique identifier for statistical units

Some of the key elements to consider when creating a unique identifier for the SBR

- Create an identification numbering system for each statistical unit, no matter what type
- Use a non-confidential identifier in order to facilitate the statistical processing.
- Ensure these unique identifiers have no meaning.
- Ensure that the unique identifiers cannot be reused.

Establishing a unique identifier for statistical units

Technical considerations in the generation of a unique identifier are as follows

- Include a check digit function (algorithm for calculating a check digit according to Modulo 11 is included in Annex E2).
- Use can be made of a key generator function.
- preferably use alpha characters combined with numbers (to avoid confusion with any other numeric data).

Outline

Software for record linkage

- According to the Guidelines: the results quoted for the performance of automated data matching tools and software tend to be overly optimistic.
- Deterministic matching or probabilistic record linkage often yields:
  - mismatches when the matching rules are too loose and
  - a high percentage of missed matches when the rules are too rigid.

Software for record linkage

- Commercial Software
  - Most of them are "black box" from the users' perspective (the source code of their linkage engines is not available for inspection).
  - Specialized to a certain domain, e.g. de-duplication of customer mailing lists
  - Affordable are only smaller systems limited in their ability to process different data types and; limited functionality; can process only small data sets.

Software for record linkage

- Open Source and Free Software
  - Allow access to the source code of their linkage engines.
  - Free of charge - although this not always means “no costs”.
  - Flexible and extendible.
  - Include large number of linkage techniques.
  - Allow the practitioner to experiment with traditional as well as advanced linkage techniques; the user is able to understand up to a certain degree, many technical details.

Software for record linkage

- RELAIS
  - ISTAT
  - Implemented in Java and R (both languages are open source and can be used on different platforms)
  - Graphical User Interface (GUI) available, written in Java
  - Input and output data in relational database—mySql—also open source product
  - Available for all major platforms
  - `https://joinup.ec.europa.eu/software/relais/description`

Software for record linkage

- FEBRL
  - *Freely Extensible Biomedical Record Linkage* (FEBRL) System
  - The Australian National University, Canbera
  - Implemented in Python (free object oriented programming language)
  - Graphical User Interface (GUI) available
  - Input from text files (CSV), SQL in the future
  - Available for all major platforms
  - Contains many recently developed record linkage techniques
  - `http://sourceforge.net/projects/febrl/`

Software for record linkage

- RecordLinkage R package
  - An R package available from CRAN
  - Machine learning methods are utilized
  - Decision trees (rpart), bootstrap aggregating (bagging), ada boost (ada), neural nets (nnet) and support vector machines (svm).
  - `http://cran.r-project.org/web/packages/RecordLinkage/index.html`
  - `http://journal.r-project.org/archive/2010-2/RJournal_2010-2_Sariyar+Borg.pdf`

Outline

Data retention

- SBR data retention strategy should be articulated in accordance with operational and analytical needs
- Should begin with the determination on how changes made to the SBR will be tracked and what historical information will need to be kept.
- Tracking changes
- Frequency and content of snapshots
- Administrative updates

Example from Statistics Canada

Example from Statistics Canada

A complete copy—snapshot—of the SBR database (live register) is taken just prior to the first day of every month. A generalized survey universe file (GSUF), i.e. frozen frame, containing every statistical unit, is created from the snapshot every month. Although frozen frames are primarily used for sampling, normally soon after their creation, they are retained for an extended period for analysis purposes.

Monthly frozen frames	**Retention period**
January	Indefinite
February to December	24 months

Inclusive and Sustainable Industrial Development

Creating shared prosperity | Safeguarding the environment

UNIDO
UNITED NATIONS
INDUSTRIAL DEVELOPMENT ORGANIZATION