Statistics Canada Statistique Canada

Canada

Statistics Canada
www.statcan.gc.ca

Toward the development of standards for record linkage at Statistics Canada

Claudia Sanmartin
Co-Chair, Working Group on Record Linkage
Statistics Canada

Workshop on Implementing Standards for Statistical Modernisation 21-23 September 2016
Geneva

Outline

- Why map the record linkage process
- Background
- Proposed Record Linkage Project Process Model
- Benefits of a model
- Seeking comments and feedback

Need for a Record Linkage Project Process Model

- Increasing importance of a standardised approach to record linkage at Statistics Canada
  - ✓ Record linkage activity has increased fivefold in past decade
  - ✓ Processes used are not consistent across Statistics Canada
- Analysis Coordinating Committee (ACC) established a Working Group on Record Linkage in June 2015
- Mandate
  - ✓ Map a generic process of a record linkage project
  - ✓ Identify issues and challenges
  - ✓ Consult with relevant divisions and committees
  - ✓ Review relevant policies and directives
  - ✓ Propose solutions
- Scope:
  - ✓ Record linkage for analytical, operational or official statistics in the social and economic domains

Why map the RL process at STC?

- Common understanding of the RL process –
  - ✓ Lots of groups involved in RL at STC use varying processes and methods;
  - ✓ Way to put everyone on the “same page”.
- Guide for those who are new to RL
  - ✓ Better understanding of the processes involved.
- Potential “generic” process to serve other statistical organisations involved in RL
  - ✓ Adaptation of the Generic Statistical Business Process Model (GSPBM)?

Sources

- Generic Statistical Business Process Model v5.0
  - ✓ Joint UNECE / Eurostat / OECD Work Session on Statistical Metadata (METIS), 2014
- International record linkage frameworks and process guides
  - ✓ Australian Bureau of Statistics
  - ✓ Agency for Healthcare Research and Quality (US)
- Record linkage methodology references
- Internal Statistics Canada policies and directives (some examples)
  - ✓ Directive on Record Linkage
  - ✓ Quality Framework – Directive on Validation
  - ✓ Departmental Project Management Framework

Generic Statistical Business Process Model (GSBPM) v5.0

Joint UNECE / Eurostat / OECD Work Session on Statistical Metadata (METIS), 2014

Levels 1 and 2 of the Generic Statistical Business Process Model

Quality management / Metadata management							
1 Specify needs	**2 Design**	**3 Build**	**4 Collect**	**5 Process**	**6 Analyse**	**7 Disseminate**	**8 Evaluate**
1.1 Identify needs	**2.1** Design outputs	**3.1** Build collection instrument	**4.1** Create frame and select sample	**5.1** Integrate data	**6.1** Prepare draft outputs	**7.1** Update output systems	**8.1** Gather evaluation inputs
1.2 Consult and confirm needs	**2.2** Design variable descriptions	**3.2** Build or enhance process components	**4.2** Set up collection	**5.2** Classify and code	**6.2** Validate outputs	**7.2** Produce dissemination products	**8.2** Conduct evaluation
1.3 Establish output objectives	**2.3** Design collection	**3.3** Build or enhance dissemination components	**4.3** Run collection	**5.3** Review and validate	**6.3** Interpret and explain outputs	**7.3** Manage release of dissemination products	**8.3** Agree on action plans
1.4 Identify concepts	**2.4** Design frame and sample	**3.4** Configure workflows	**4.4** Finalise collection	**5.4** Edit and impute	**6.4** Apply disclosure control	**7.4** Promote dissemination products	
1.5 Check data availability	**2.5** Design processing and analysis	**3.5** Test production system		**5.5** Derive new variables and units	**6.5** Finalise outputs	**7.5** Manage user support	
1.6 Prepare business case	**2.6** Design production systems and workflow	**3.6** Test statistical business process		**5.6** Calculate weights			
		3.7 Finalise production systems		**5.7** Calculate aggregates			
				5.8 Finalise data files			

Record Linkage Project Process Model

(Draft) (Updated July 2016)

Phase 1: Project Planning			Phase 2: Record Linkage			Phase 3: Post linkage activities		
1 Specify needs	2 Design	3 Approve	4 Prepare data	5 Link data	6 Assess quality	7 Integrate and Analyze	8 Access and Disseminate	9 Evaluate
1.1 Identify needs	2.1 Design linkage strategy	3.1 Consult and confirm approval process	4.1 Standardize linkage variables	5.1 Indexing (Blocking)	6.1 Internal validation	7.1 Integrate data, review and validate	8.1 Establish access process	9.1 Gather evaluation inputs
1.2 Consult and confirm needs	2.2 Design quality assessment strategy	3.2 Prepare approval documents	4.2 Assess linkage variables	5.2 Field and record comparison	6.2 External validation	7.2 Apply quality adjustments	8.2 Establish disclosure control protocols	9.2 Conduct evaluation
1.3 Check data availability	2.3 Plan for quality adjustments	3.3 Submit for approval	4.3 Identify in scope records for linkage	5.3 Linkage rules	6.3 Adjust record linkage strategy	7.3 Derive variables	8.3 Store and manage access	9.3 Agree on action plans
1.4 Determine feasibility of record linkage	2.4 Identify access needs	3.4 Archive approval	4.4 Evaluate results of data preparation	5.4 Finalize record linkage strategy	6.4 Produce linkage keys	7.4 Finalize linked data set and document	8.4 Destruction of files	9.4 Add to linkage tool box
1.5 Identify sponsor and custodian	2.5 Estimate cost		4.5 Initiate record linkage report	5.5 Document record linkage strategy	6.5 Finalize record linkage report	7.5 Analyze, validate and feedback		
OUTCOMES/OUTPUTS								
Decision to proceed with record linkage project	Project plan and budget	Approved record linkage project	Linkage ready data sets	Preliminary linkage keys	Final linkage keys; Record linkage report	Linked set; Documentation; Analytical products	Disclosure and Access protocols	Evaluation Report; Tool box

1. Specify Needs

1.1 Identify needs	1.2 Consult and confirm needs	1.3 Check data availability	1.4 Determine feasibility of record linkage	1.5 Identify sponsor and custodian

- Consult with client/stakeholder to confirm data needs – understanding objectives of analysis/research;
- Determine if existing data or linked data can meet the needs
- Explore feasibility of record linkage to meet needs –
  - Identify source (input) data files;
  - Map common identifying variables to support linkage (linkage variables)
- Identify sponsor (champion the record linkage project) and data custodian (assume responsibility for linked data)

- Outcome: Decision to proceed with record linkage project

2. Design

2.1 Design linkage strategy	2.2 Design the quality assessment strategy	2.3 Plan for quality adjustments	2.4 Identify access needs	2.5 Estimate cost

- Selection of record linkage methodology – deterministic, probabilistic – based on available data and linkage variables;
- Design quality assessment – measures of internal and external validation – consider cost implications;
- Consider the potential need for quality adjustments (e.g. weighting) based on use of data and level of precision required;
- Identify who will need access to the linked data – restricted to client/stakeholder or access for other researchers (e.g. Research Data Centres)
- Preliminary cost estimate to ensure project is within client/stakeholder budget

- Outcome: Project Plan and budget

3. Approval

3.1 Consult and confirm approval process	3.2 Prepare approval documents	3.3 Submit for approval	3.4 Archive approvals

- Confirm approval process – consult the applicable policies, directives and guidelines;
- Prepare request for approval
- Submit for approval
- Approved proposal should be archived and easily accessible for future consultation

- Outcome: Record linkage approval

4. Prepare Data

4.1 Standardise linkage variables	4.2 Assess linkage variables	4.3 Identify in-scope records for linkage	4.4 Evaluate results of data preparation	4.5 Initiate record linkage report

- Subset linkage variables from source files – standardise format – consult data custodians to obtain relevant information;
- Assess linkage variables -
  - ✓ Quality – completeness, rate of missing and invalid values
  - ✓ Discriminatory power – differentiate entities
- Identify records from each source file that are eligible for linkage – informed by previous step – inform quality later on
- Evaluate results of data preparation – decide whether to move on with record linkage
- Initiate the record linkage report – document results of pre-processing

- Output: Linkage ready files

5. Link Data

5.1 Indexing (or Blocking)	5.2 Field and record comparison	5.3 Linkage rules	5.4 Finalise record linkage strategy	5.5 Document record linkage strategy

- Initial index (blocking) to limit the number of pairs to be evaluated – critical when linking large datasets – consider only pairs that match on an initial criteria
- Compare the pairs using information from the linkage variables;
- Use results of the comparison to derive the linkage rules – to determine which pairs are a match or non-match;
- Iterative process – evaluate initial strategy, make adjustments, evaluate results – until finalise strategy
- Document strategy and results in the record linkage report

- Output: Preliminary linkage keys

6. Assess Quality

6.1 Internal Validation	6.2 External Validation	6.3 Adjust record linkage strategy	6.4 Produce linkage keys	6.5 Finalise record linkage report

- Internal validation – assessing quality and accuracy of the linkage
  - ✓ Assess initial linkage rates – sub-groups for expected patterns
  - ✓ Assess non-linked records – evidence of bias
  - ✓ Measures of accuracy – false positive and false negative rates
- External validation – assessing "fitness for use" of linked data – may require additional variables from source files – compare with external sources
- Make adjustments to the linkage strategy based on results of quality assessment – if required
- Document results of quality and finalise record linkage report

- Output: Final linkage keys; Record Linkage Report

7. Integrate and Analyze

7.1 Integrate data, review and validate	7.2 Apply quality adjustments	7.3 Derive variables	7.4 Finalize linked data set and document	7.5 Analyze, validate and feedback

- Integrate the source file using the linkage keys to produce a linked analytical file – review to ensure merging conducted correctly;
- Apply quality adjustments anticipated in sub-process 2.3 and others if necessary – e.g. weighting, imputation
- Derive new variables, if necessary
- Document the linked data file – refer to existing documentation of source files – avoid duplication!
- First analysis of the linked data – focus on objectives of the analysis/ research specified in Phase 1 – further validation of the linked data file, provide feedback to record linkage if there are issues:
  - ✓ Assess linkage rates for their cohort of interest
  - ✓ Test associations between variables across source files
- Output: Linked data file; Documentation; Analytical products

8. Access and Disseminate

8.1 Establish access process	8.2 Establish disclosure control protocols	8.3 Store and manage access	8.4 Destruction of files

- Review access requirements from planning phase (2.4) – revise if required – there may broader interest in the linked data;
- Establish disclosure protocols – i.e. minimum cell sizes, restrictions on level of geography, treatment of output (e.g. random rounding);
- Consult with linked data custodian and/or custodians of source data to ensure meet requirements under which data are available;
- Manage access to the linked data file in secure environment
- At the end of the retention period, ensure file is either immediately destroyed or request extension

- Output: Disclosure and Access protocols

9. Evaluate Linkage Project

9.1 Gather evaluation inputs	9.2 Conduct evaluation	9.3 Agree on action plan	9.4 Add to the record linkage "toolbox"

- ***Evaluation of the project NOT the linked data – already did that!***
- Inputs to the evaluation can come at any stage – look for "lessons learned";
- Synthesise inputs in an evaluation or "close-out" report highlighting quality issues, lessons learned, evaluation of project assumptions (time lines and costs), recommendations for improvement;
- Consult with senior managers to agree on any actions or changes to the process as a result of the evaluation
- Add tools, programs, processes to the record linkage "toolbox" for use in subsequent linkage projects – improve efficiency, standardisation, and quality

- Output: Evaluation Report; Toolbox

Benefits

- Mapping the process of a record linkage project:
  - ✓ defines and describes the process in a coherent way making it easier to plan and perform future record linkage projects;
  - ✓ is essential to creating reliable and complete information on the methodology and data quality for the data users to make appropriate use of the linked data;
  - ✓ highlights the fact that a record linkage project is more than a simple merging exercise;
  - ✓ establishes solid benchmarks to facilitate comparison of various processes within and between statistical agencies.

- Using the Record Linkage Project Process Model could help in ensuring that record linkage activities are more efficient and can potentially improve data quality in much the same way as the production of official statistics has been improved by the Generic Statistical Business Process Model.

Comments and feedback welcome

Working Group co-chairs:

Claudia.Sanmartin@canada.ca

Richard.Trudeau@canada.ca

For more information on record linkage at Statistics Canada:

http://www.statcan.gc.ca/eng/record/gen