Modelling Provenance in GSIM/CSPA-LIM

Francine Kalonji
Information Architecture
Systems Engineering Division

April 2018
Context

- Single point of access for corporate metadata
- Emphasis on administrative data and other source of data to reduce cost and response burden
- Data exchanged internally and externally
  - Provenance and lineage become very important
    - To assess value, reliability and pertinence of data
  - Need for proper documentation of provenance and lineage
    - Use of a standard (common language) necessary to understand provenance and lineage
Some definitions

- **Provenance**: information about
  - source of data,
  - how it was captured/accessed/processed/produced,
  - its description,
  - its location/storage
    - along its path
- **Lineage**: information about
  - where/how data moves over its life cycle (transformations)
Logical level

Level of Detail

Conceptual

User view, Independent of physical implementation

Logical

Meets requirements of CSPA Services

CSPA Logical Information Model

Physical

Implemented in CSPA Services

Human Oriented

Computer Oriented
Provenance/Lineage in CSPA-LIM
Provenance/Lineage in CSPA-LIM

Base ::IdentifiableArtefact
+ description: MultilingualText
+ version: String
+ id: String
+ versionDate: Date
+ versionRationale: String
+ localID: String [0..1]
+ name: MultilingualText

ProcessInput
1..*
has
1

ProcessStepInstance
1
creates
1..*

ProcessOutput
Provenance/Lineage in CSPA-LIM

- in Administrative Details
  - created date
  - valid from/valid to
  - url
- in Identifiable Artefact
  - version
  - process input/process output
- as relationships between
  - change event source(s)/change event target(s)
  - Change Even Tuple and Identifiable Artefact
  - administrator Agent
  - In Role and Administrative Details
Provenance/Lineage in CSPA-LIM

- Process flows
  - established with process input and process output
- date attribute of the Process Step Instance (Inherited from Identifiable Artificat)

- Links
  - between change event source(s) and change event target(s) with related change types.
    - For instance, if the changeType attribute is versioning, then the source and target can help provide a version chain.
Limitations

- **ChangeEvent object**
  - manages changes in states on information objects
  - can be leveraged to represent some provenance
- **PROV-O is broader and models**
  - exchange between activities: communication
  - association between agent and activity
  - delegation between two agents
- **PROV-O provides a simple provenance framework**
PROV-O basic classes

- Agent
  - actedOnBehalfOf
  - wasAttributedTo
  - wasAssociatedWith
- Activity
  - startedAtTime: xsd:dateTime
  - endedAtTime: xsd:dateTime
  - wasInformedBy
- Entity
  - wasDerivedFrom
  - wasGeneratedBy
  - used
PROV-O and Statistics Canada objects

PROV-O
- Activity
  - something that occurs over a period of time and acts upon or with entities

Statistics Canada
- Statistical activity / process
  - action steps to produce statistical outputs
PROV-O and Statistics Canada objects

**PROV-O**
- Agent
  - something that bears some form of responsibility

**Statistics Canada**
- Agent
  - performs a role in relation to a statistical business process/entity
PROV-O and Statistics Canada objects

**PROV-O**
- Entity
  - physical, digital, conceptual, or other kind of thing with some fixed aspects

**Statistics Canada**
- Other entities
  - Core concepts to conduct statistical activities
    - Variable
    - Data Asset
    - Questionnaire
1. Dataset3 is derived from Dataset1 and Dataset2
2. Dataset3 was generated by Record Linkage activity using Dataset1 and Dataset2
Modeling provenance

- **type**: for provenance type
- **description**
- **id**: to be able to hold items together in the case of a provenance chain
- **sortOrder**: to establish the sequence
Challenges

• Reconciling
  • similar terms and their definitions
  • different terms with same definition
    ➢ Essential in mapping objects from one standard to another

• Process
  • PROV-O: activity => GSIM: business process, process step
  • GSIM: process step => GSBPM: sub-process or sub-sub-process etc.
Challenges

Data Integration

Business Process

Process step

5.1 Integrate data
- 5.1.1 Combine various data sources
- 5.1.2 Match and link records
- 5.1.3 Anonymize data

5.4 Edit and impute
- 5.4.1 Perform data profiling
- 5.4.2 Remove duplicate records

5.5 Derive new variables and units
- 5.5.1 Aggregate data
- 5.5.1 Add new data

5 Process
- 5.1 Integrate data
- 5.2 Classify and code
- 5.3 Review and validate
- 5.4 Edit and impute
- 5.5 Derive new variables and units

6 Analyse
- 6.1 Prepare draft outputs
- 6.2 Validate outputs
Future considerations

- Extend PROV-O to include statistical domain properties
  - International collaboration
    - We are interested in sharing knowledge with other organizations looking at this
- Consider different levels of granularity
  - From Datasets to data records to data points
- Look at other provenance models?
Thank you!

Francine.Kalonji@Canada.ca