Statistics Canada Statistique Canada

Modelling Provenance in GSIM/CSPA-LIM

www.statcan.gc.ca

Francine Kalonji
Information Architecture
Systems Engineering Division

April 2018

Canada

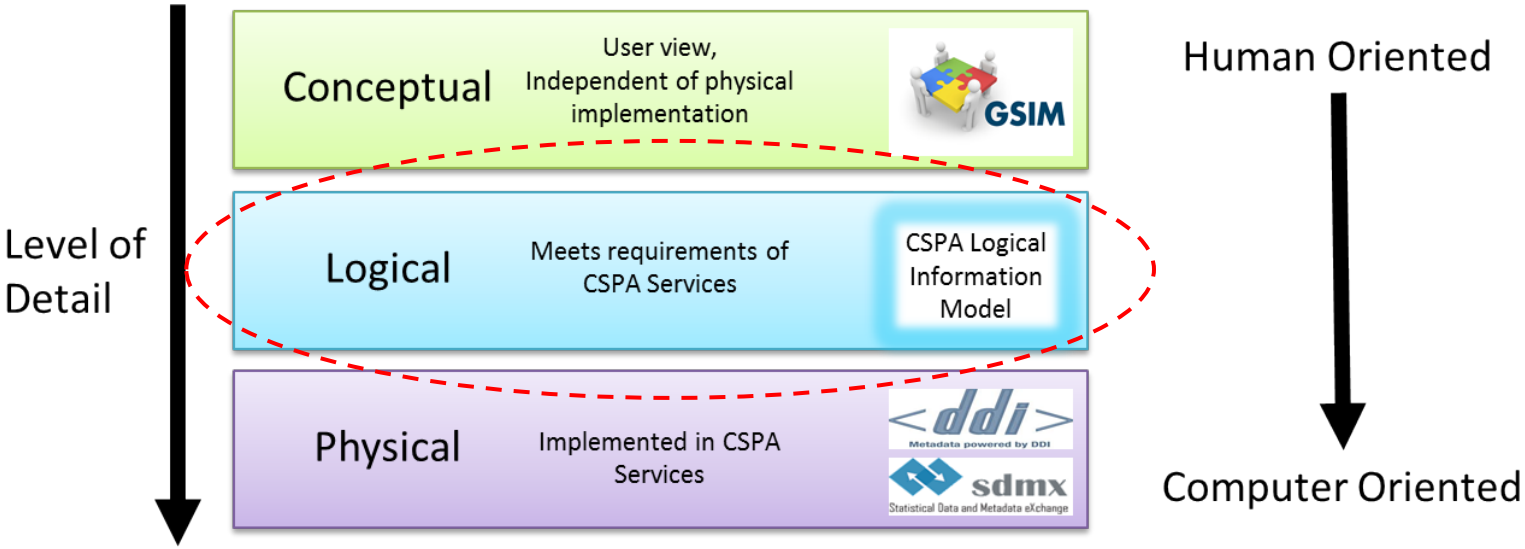
Context

- 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

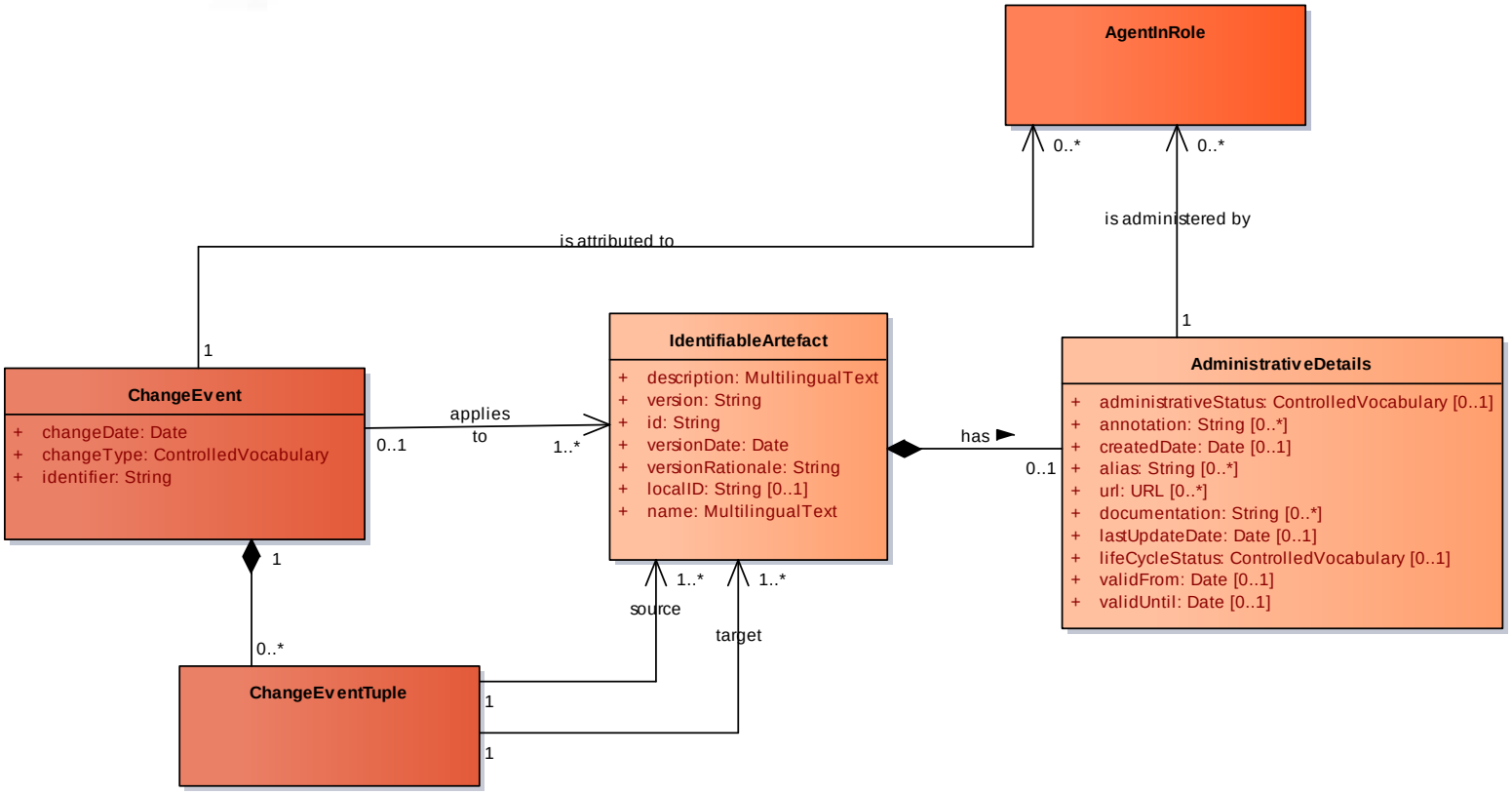
Some definitions

- 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 it life cycle (transformations)

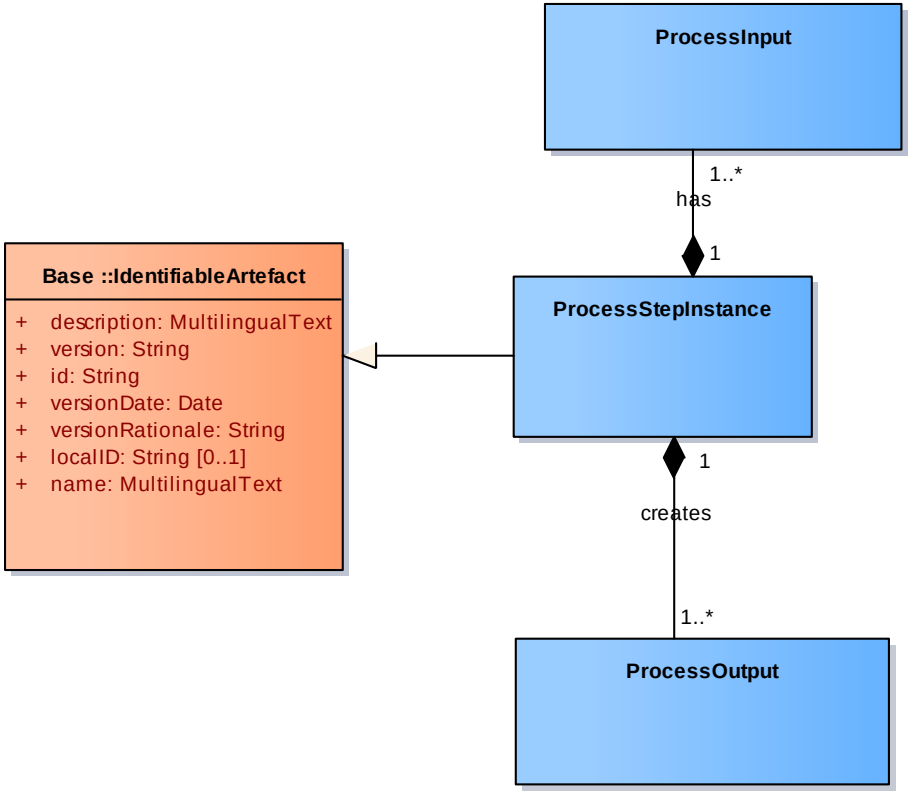
Logical level



Provenance/Lineage in CSPA-LIM



Provenance/Lineage in CSPA-LIM



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

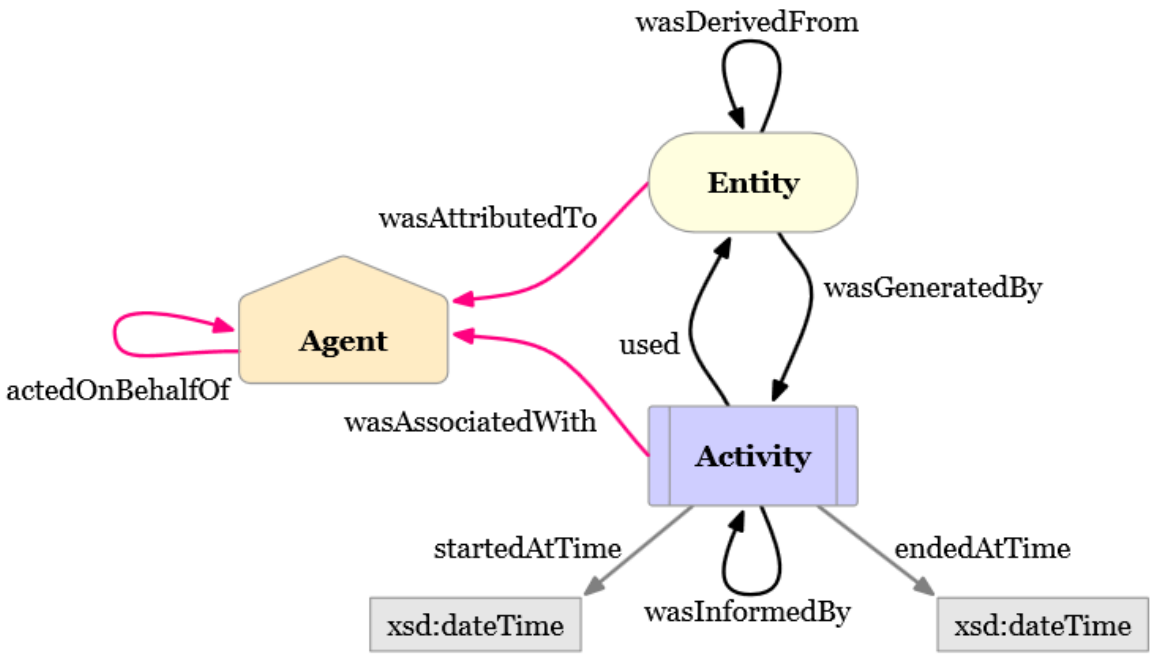
Provenance/Lineage in CSPA-LIM

- 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



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

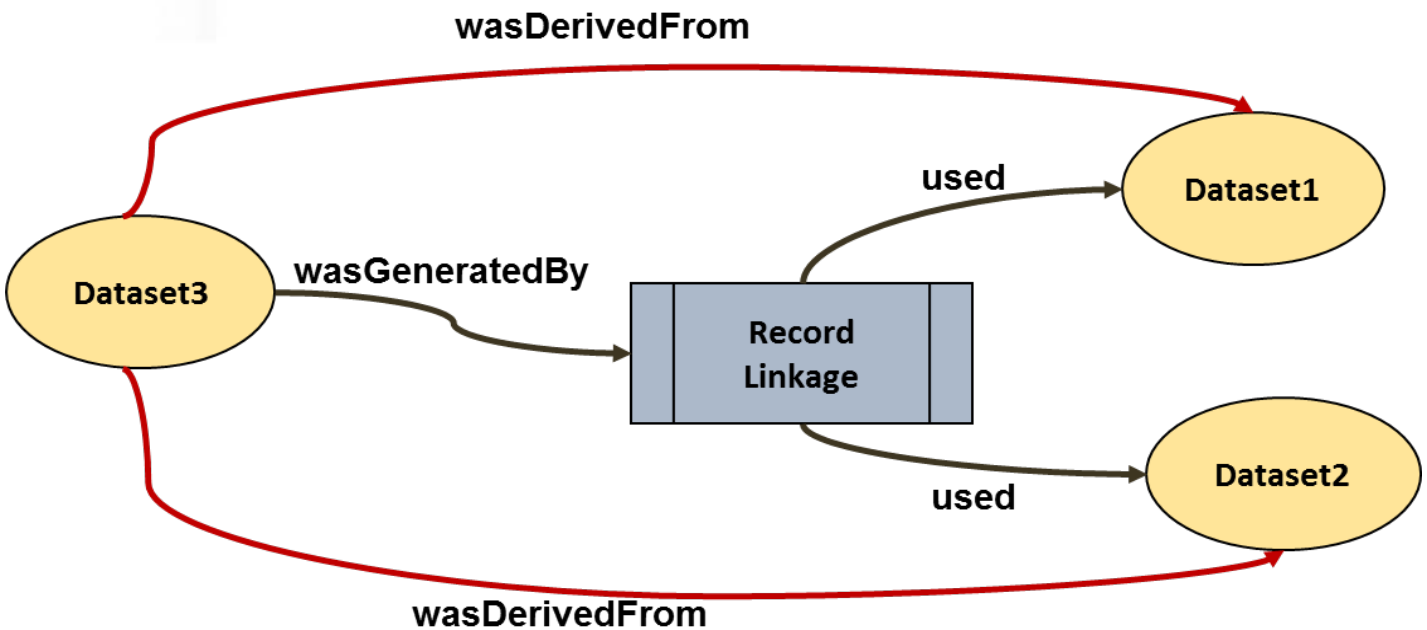
PROV-O

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

Statistics Canada

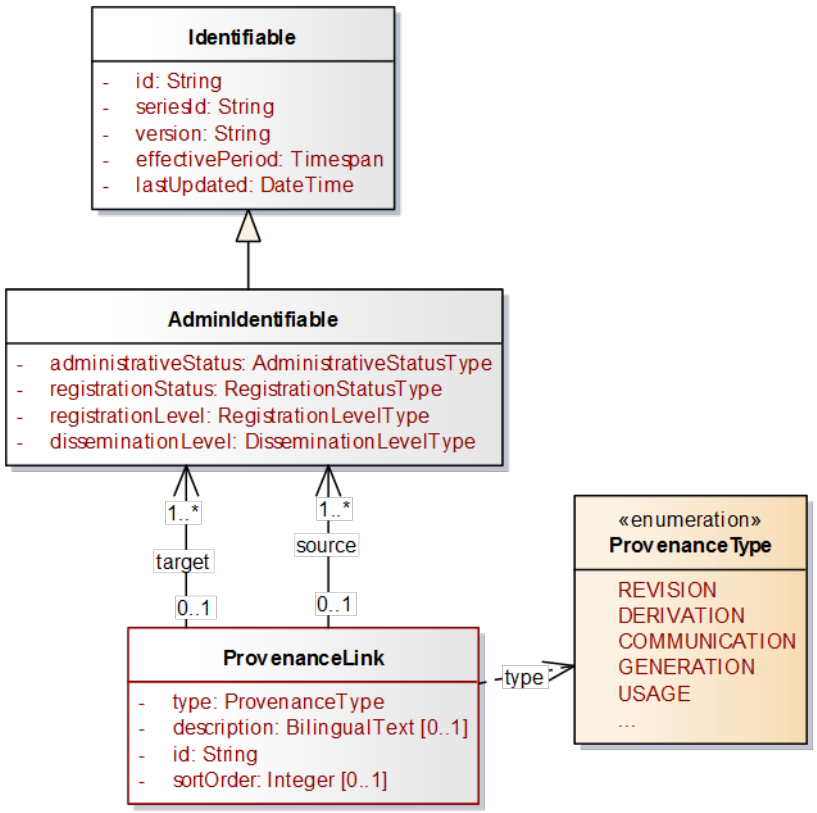
- Other entities
  - Core concepts to conduct statistical activities
    - Variable
    - Data Asset
    - Questionnaire

Modeling provenance



1. Dataset3 is derived from Dataset1 and Dataset2
2. Dataset3 was generated by Record Linkage activity using Dataset1 and Dataset2

Modeling provenance



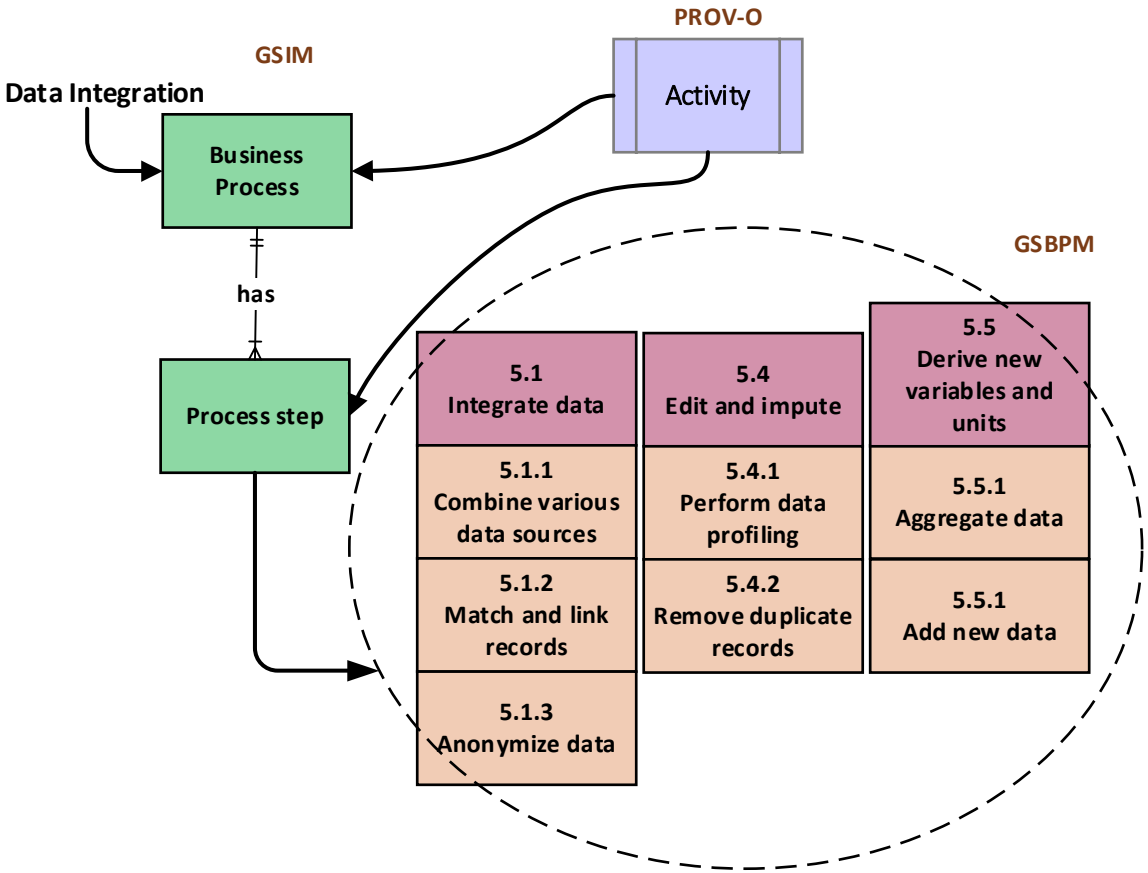
- *type*: for provenance type
- *description*
- *id*: to be able to hold items together in the case of a provenance chain
- *sortOrder*: to establish the sequence

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



5 Process	6 Analyse
5.1 Integrate data	6.1 Prepare draft outputs
5.2 Classify and code	6.2 Validate outputs
5.3 Review and validate	
5.4 Edit and impute	
5.5 Derive new variables and units	

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?

Comments/Feedback

Thank you!

Francine.Kalonji@Canada.ca