SemStats

Using PROV-O to represent lineage in statistical processes: a record linkage example

Flavio Rizzolo – Statistics Canada
Guillaume Duffes – Institut National de la Statistique et des Études Économiques
Franck Cotton – Institut National de la Statistique et des Études Économiques

Contents

Context and objectives

Statistical offices need to provide trusted data

Information on how data was produced helps doing that

Provenance and lineage metadata are information on

- Processes and methods used
- Actors involved (data providers, owners, publishers, etc.)
- Relations between data outputs and data sources

That metadata should

- Use a standard model in order to be easily understandable
- Be accessible and (machine-)usable

Context and objectives

Main goal of the paper: proof of concept about using the PROV model to represent lineage information on statistical processes

Record linkage chosen as example process

- Sufficiently complex, but not too much
- Widely used in statistical production
- Formal descriptions already available
- Lineage metadata can be defined at various levels of detail
- Various software packages exist

Context and objectives

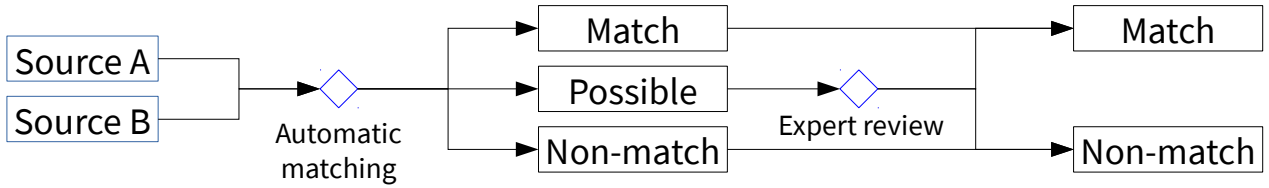
This is a very practical work, not groundbreaking research

Record linkage and lineage metadata

Record linkage

Matching of data about real-world entities (people, businesses, products…) coming from different data sources

Typical process

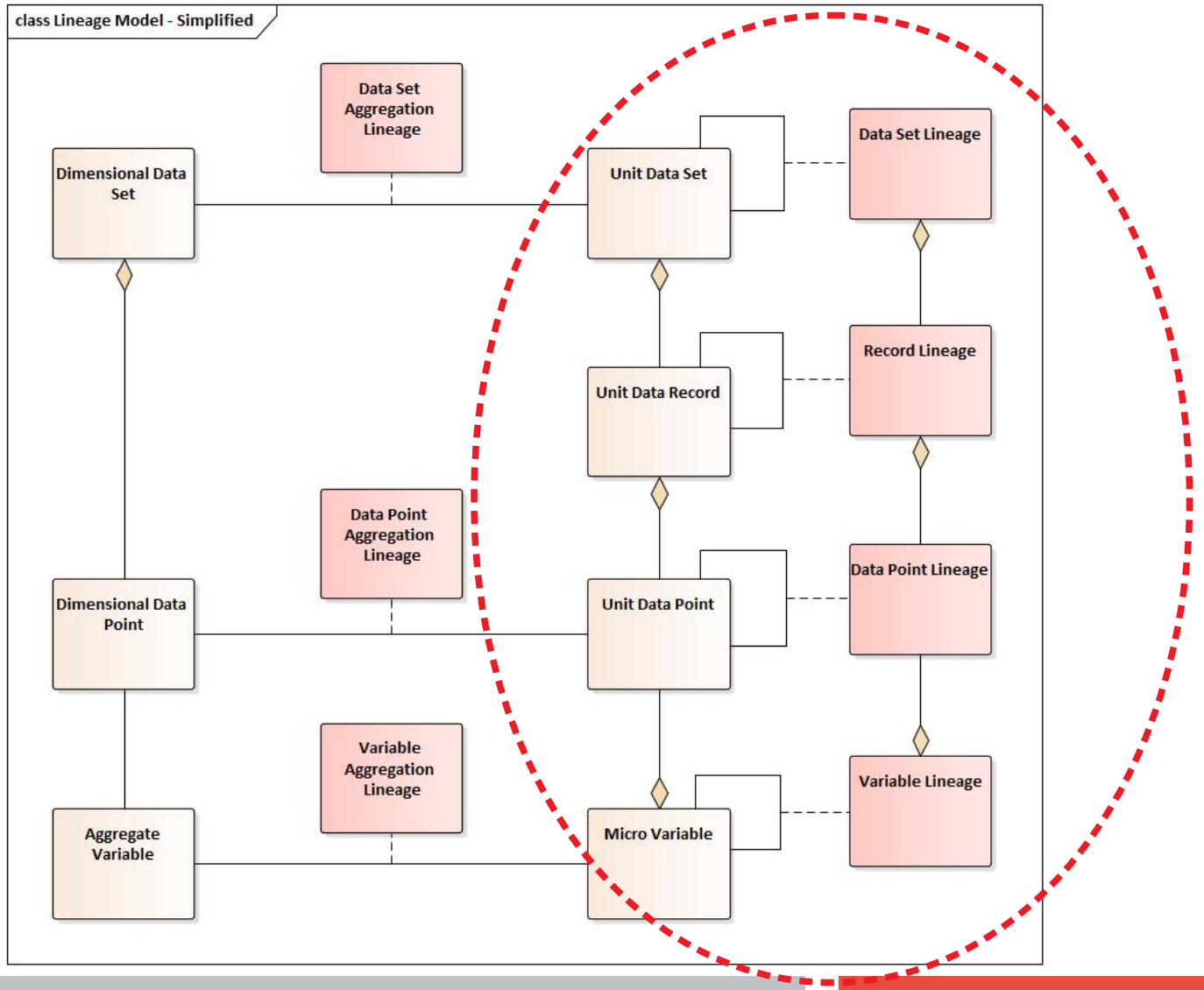


Widely used (e.g. data integration), lots of methodological work

Even a dedicated record linkage process model (Statistics Canada)

Record linkage and lineage metadata

Lineage model



Record linkage and lineage metadata

Types of lineage metadata

Dataset lineage

A dataset is derived from others by record linkage: keep track of sources and transformations applied

Record lineage

Track where the record comes from or which records are its contributors and what integration was applied

Variable lineage

Track how a variable (e.g. linkage key) is derived from variables in source datasets

Data point lineage

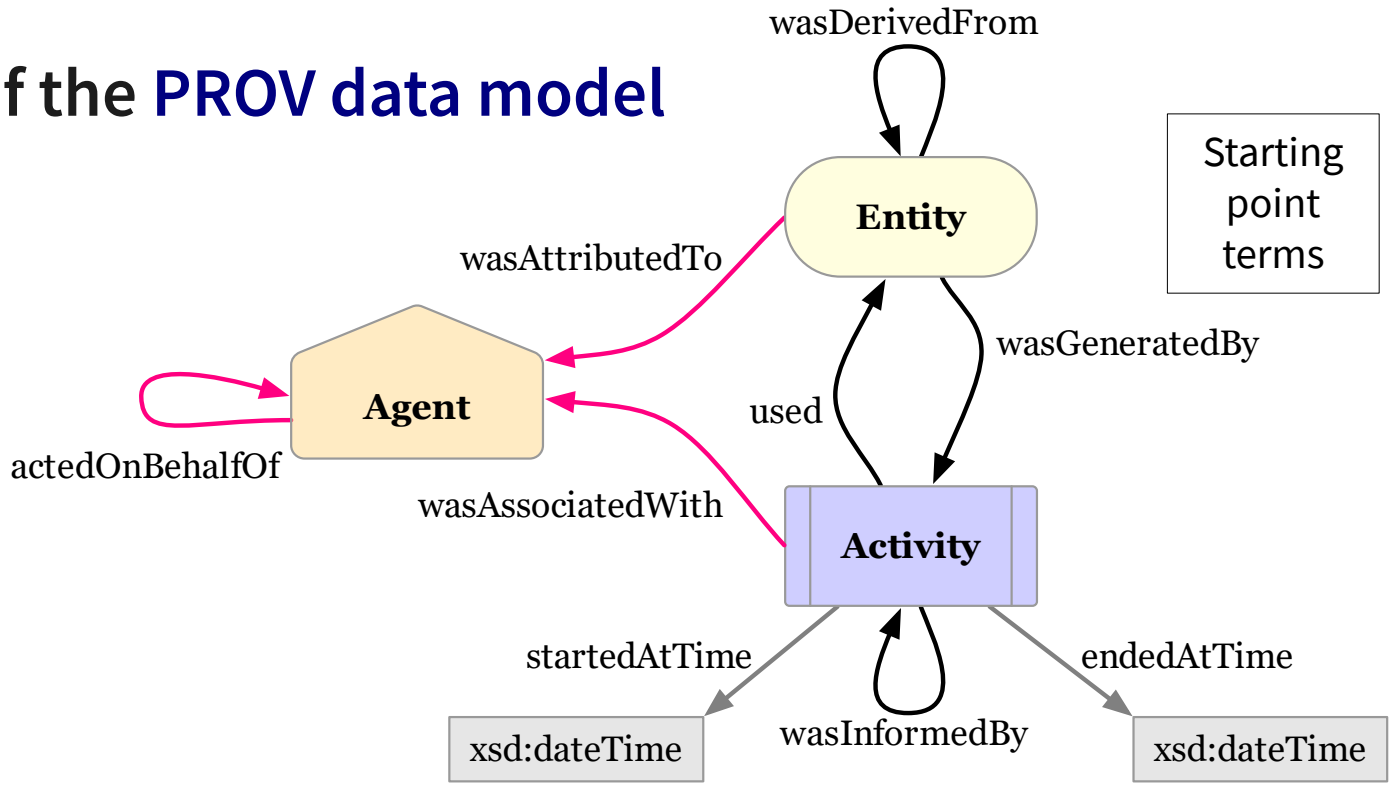
Not used for record linkage but heavily used in upstream tasks like data cleansing

What is PROV-O?

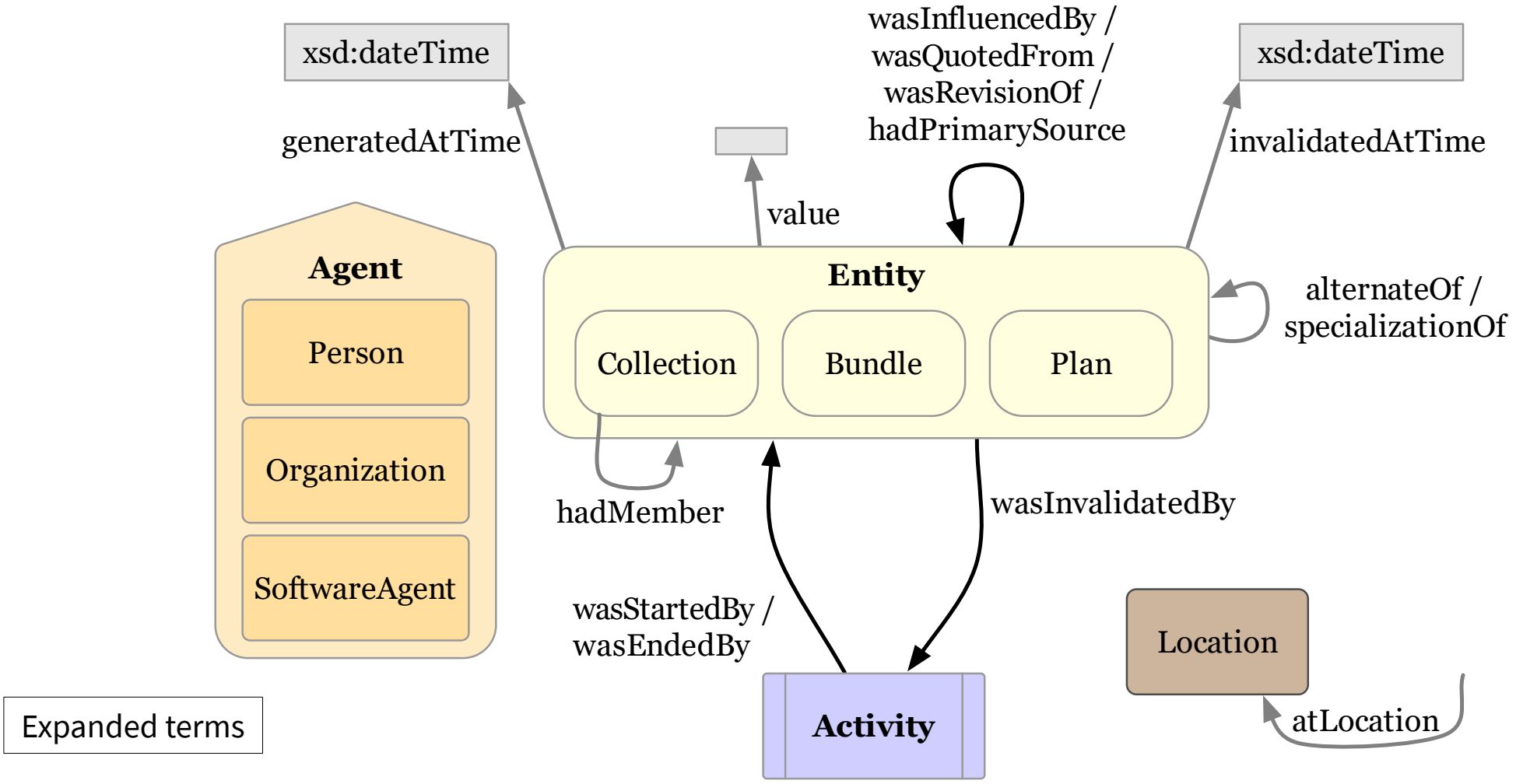
W3C recommendation part of the PROV familly (provenance metadata)

OWL2 expression of the PROV data model

Simple "Starting point" model, expanded terms and qualification mechanism

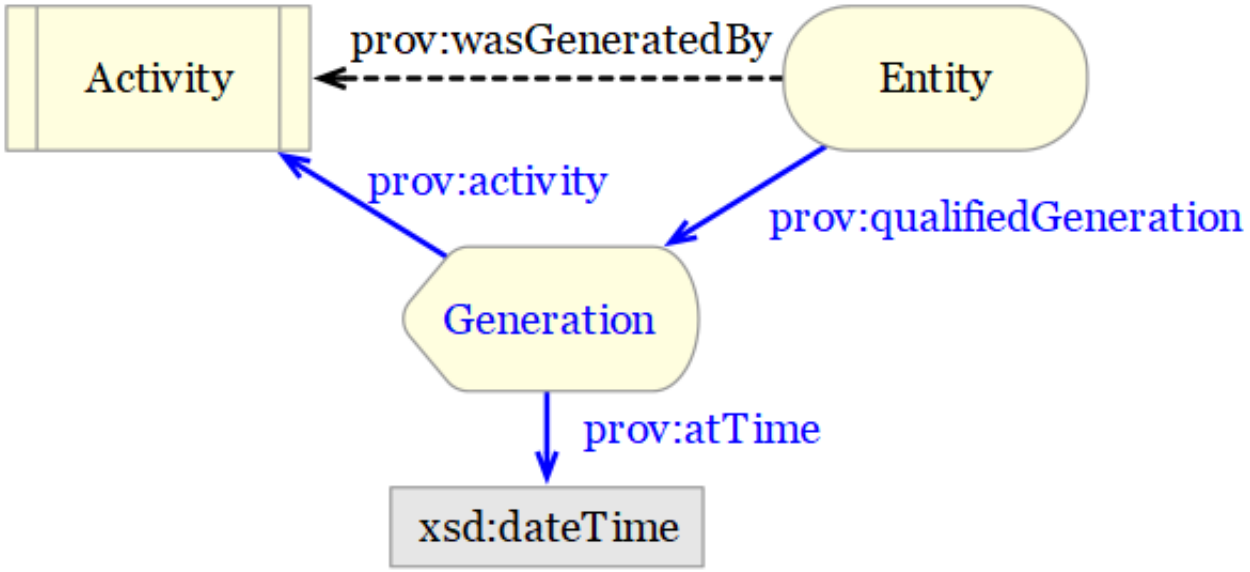


What is PROV-O?



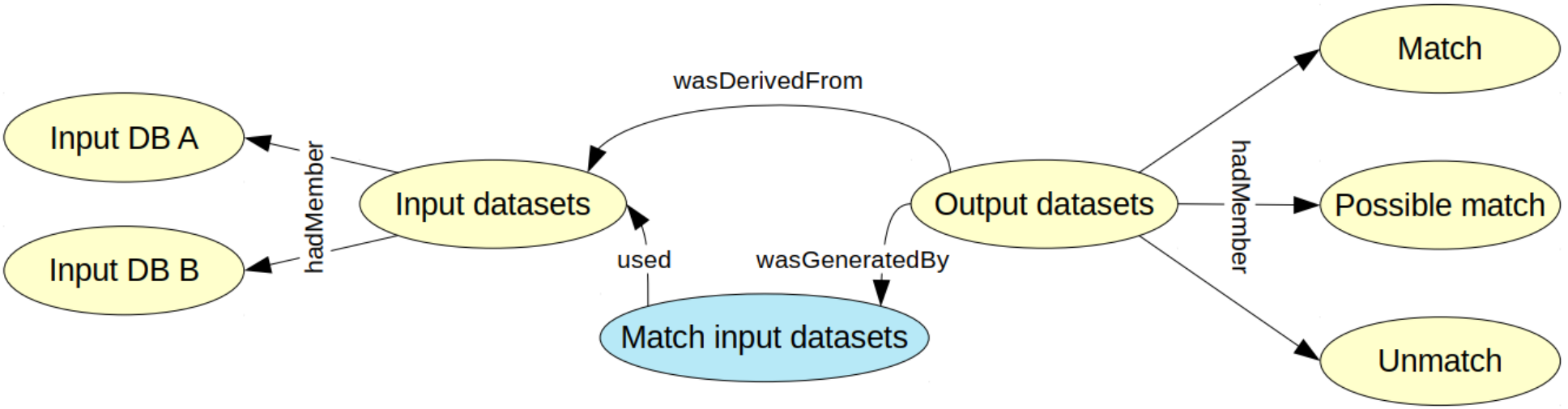
What is PROV-O?

Qualification mechanism



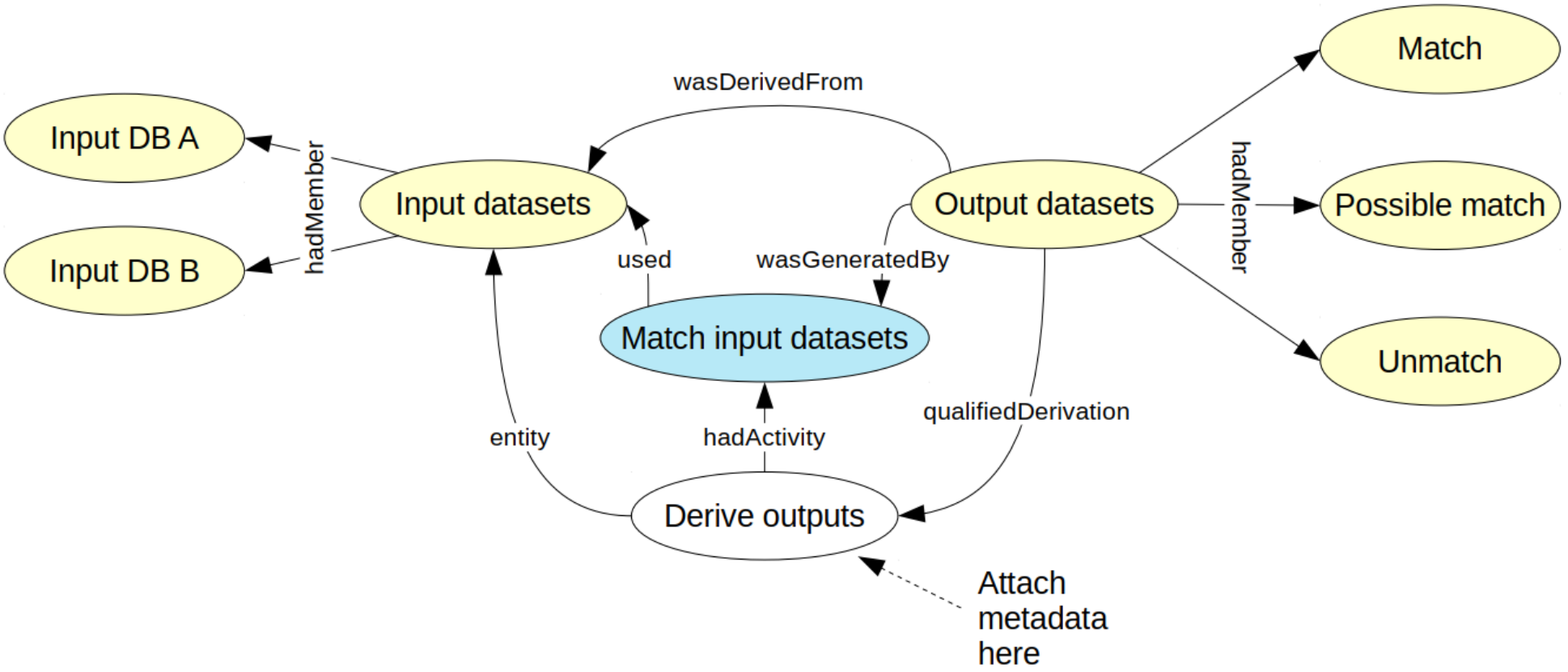
PROV-O representations for record linkage lineage

Simple example: the high-level view



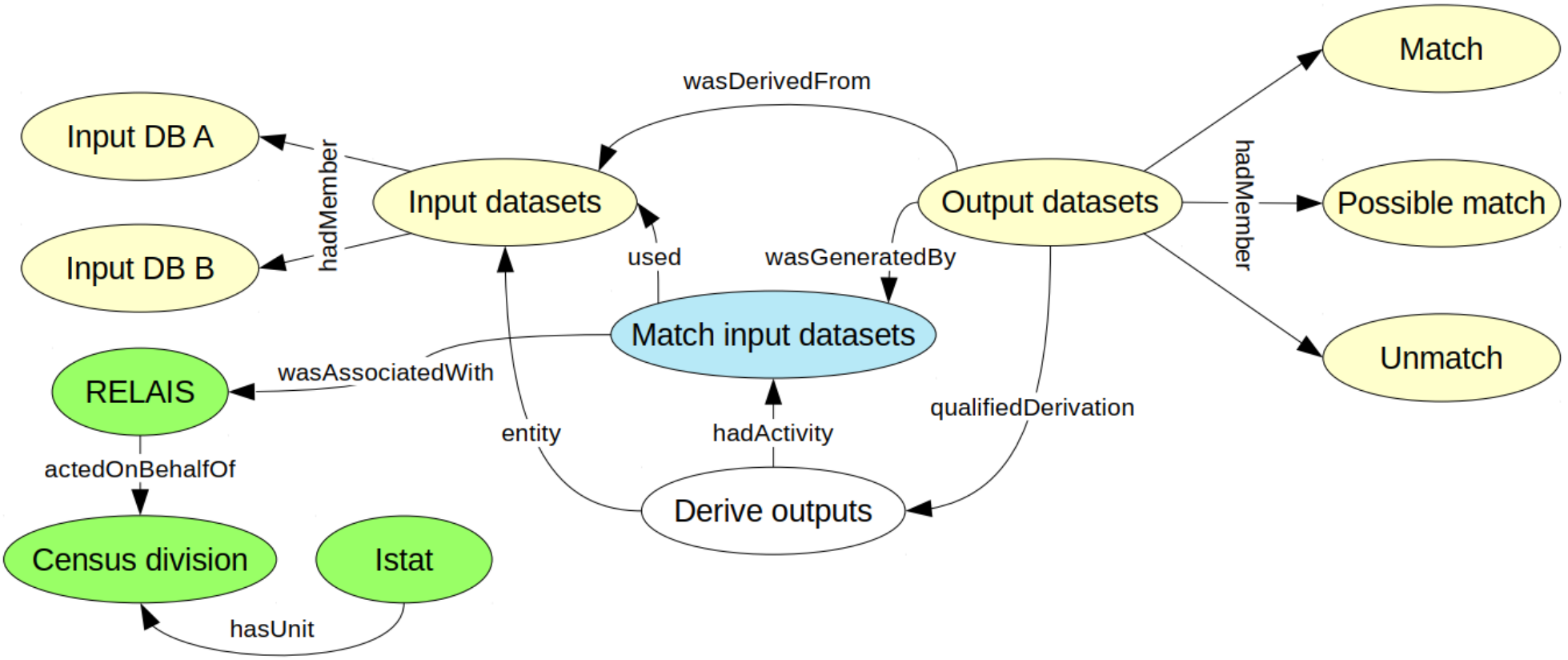
PROV-O representations for record linkage lineage

Simple example: the high-level view



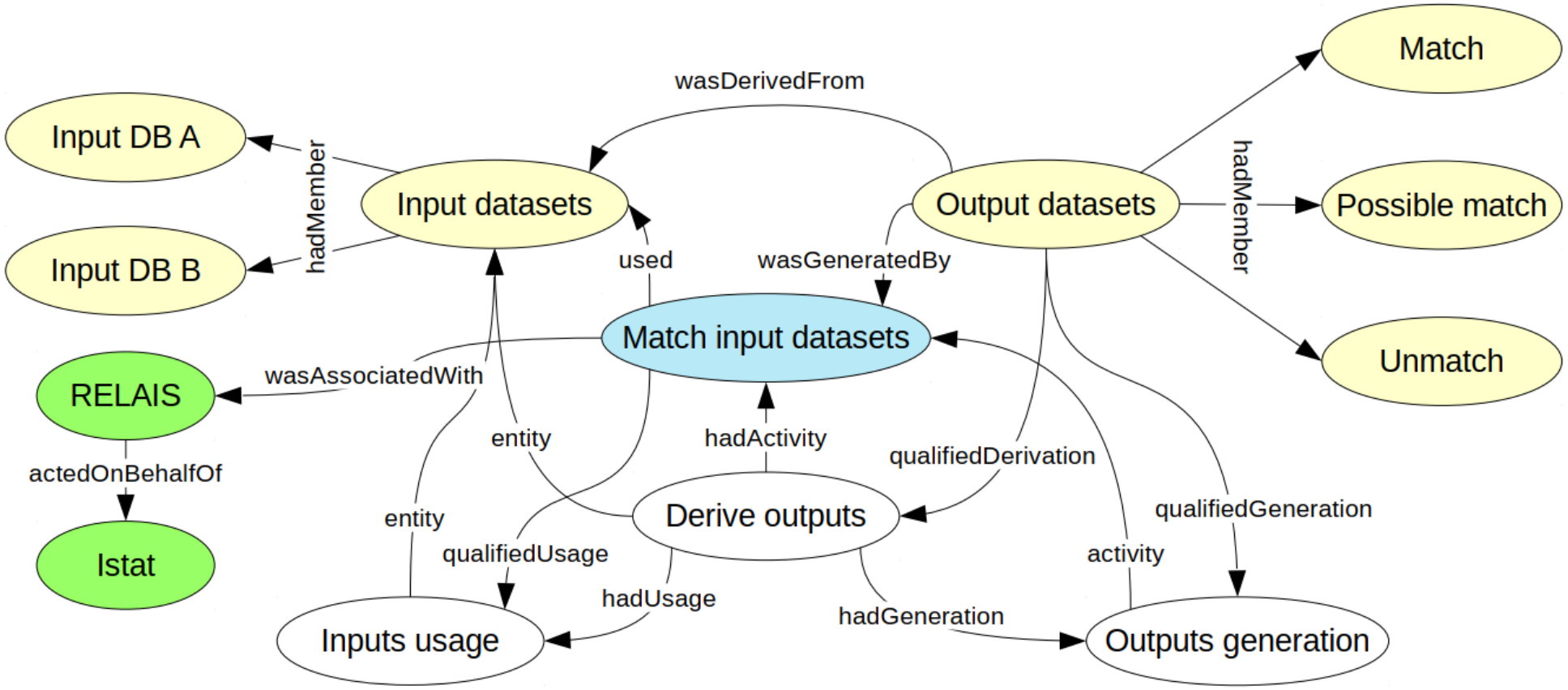
PROV-O representations for record linkage lineage

Simple example: the high-level view



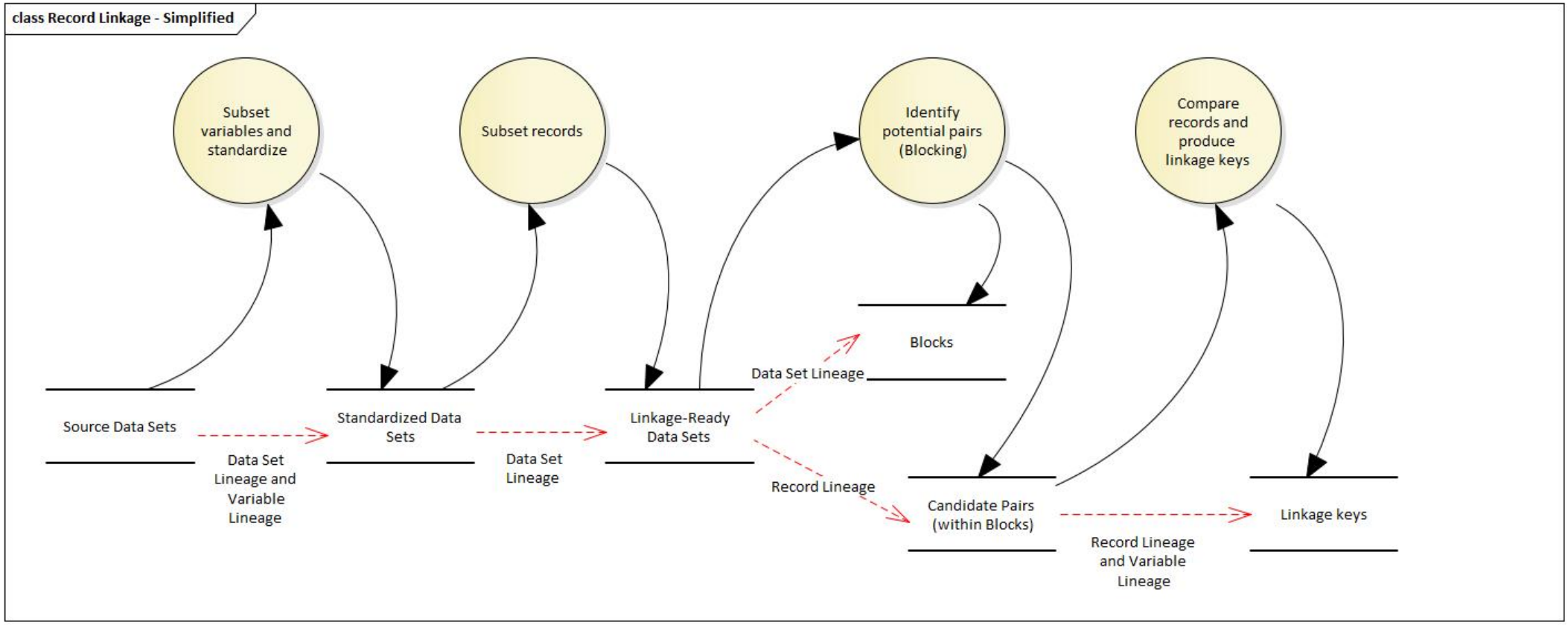
PROV-O representations for record linkage lineage

Simple example: the high-level view



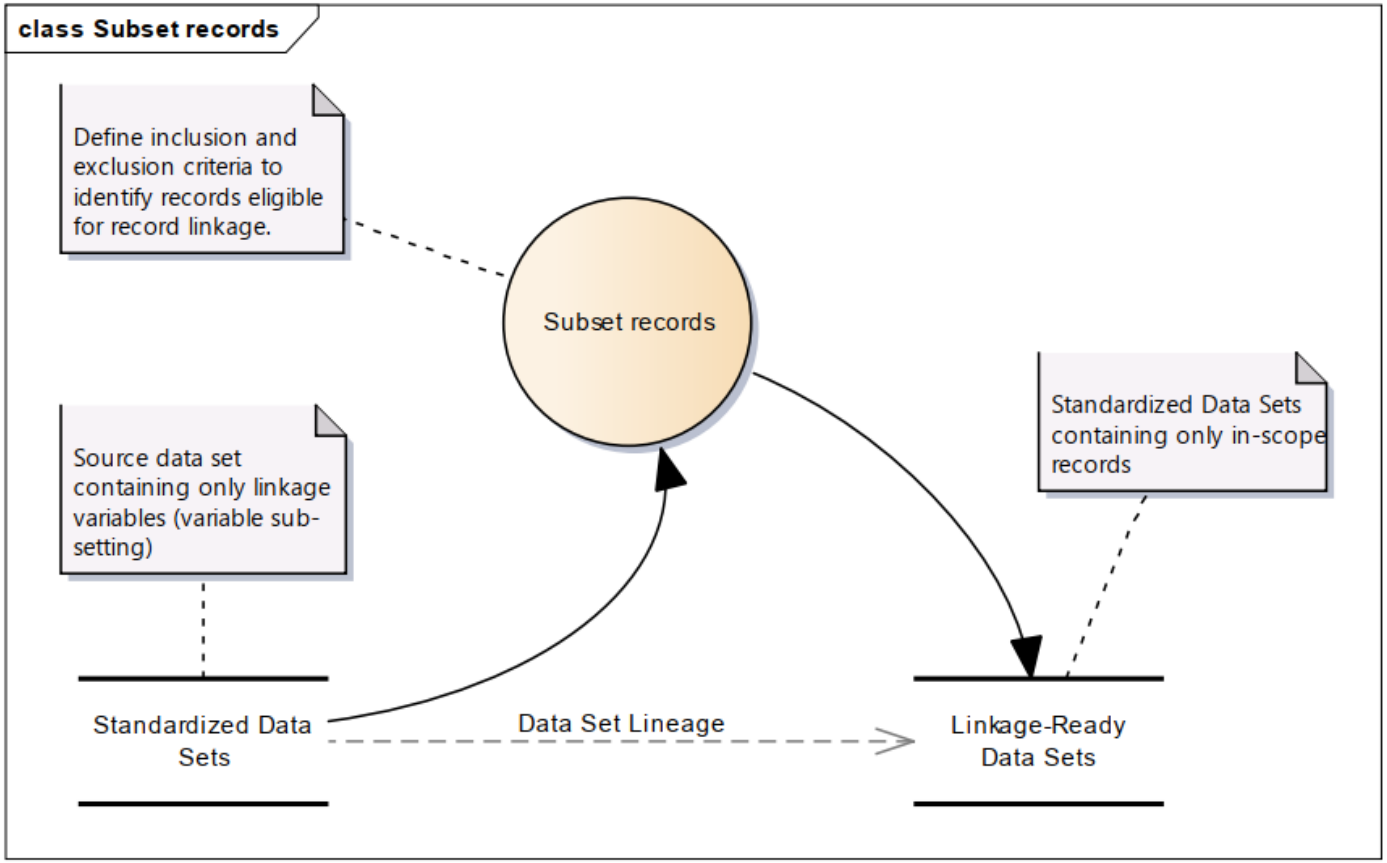
PROV-O representations for record linkage lineage

The record linkage process (simplified)



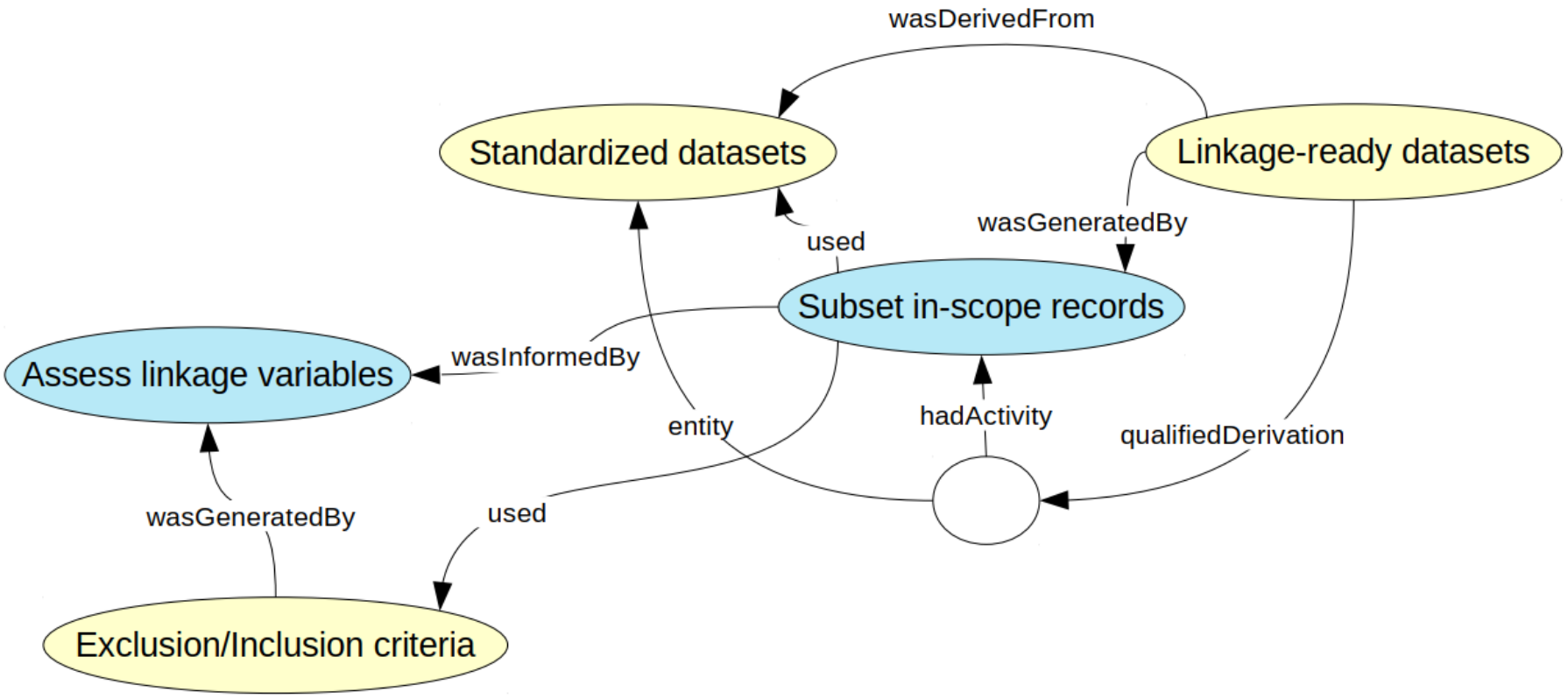
PROV-O representations for record linkage lineage

Produce linkage-ready datasets – process



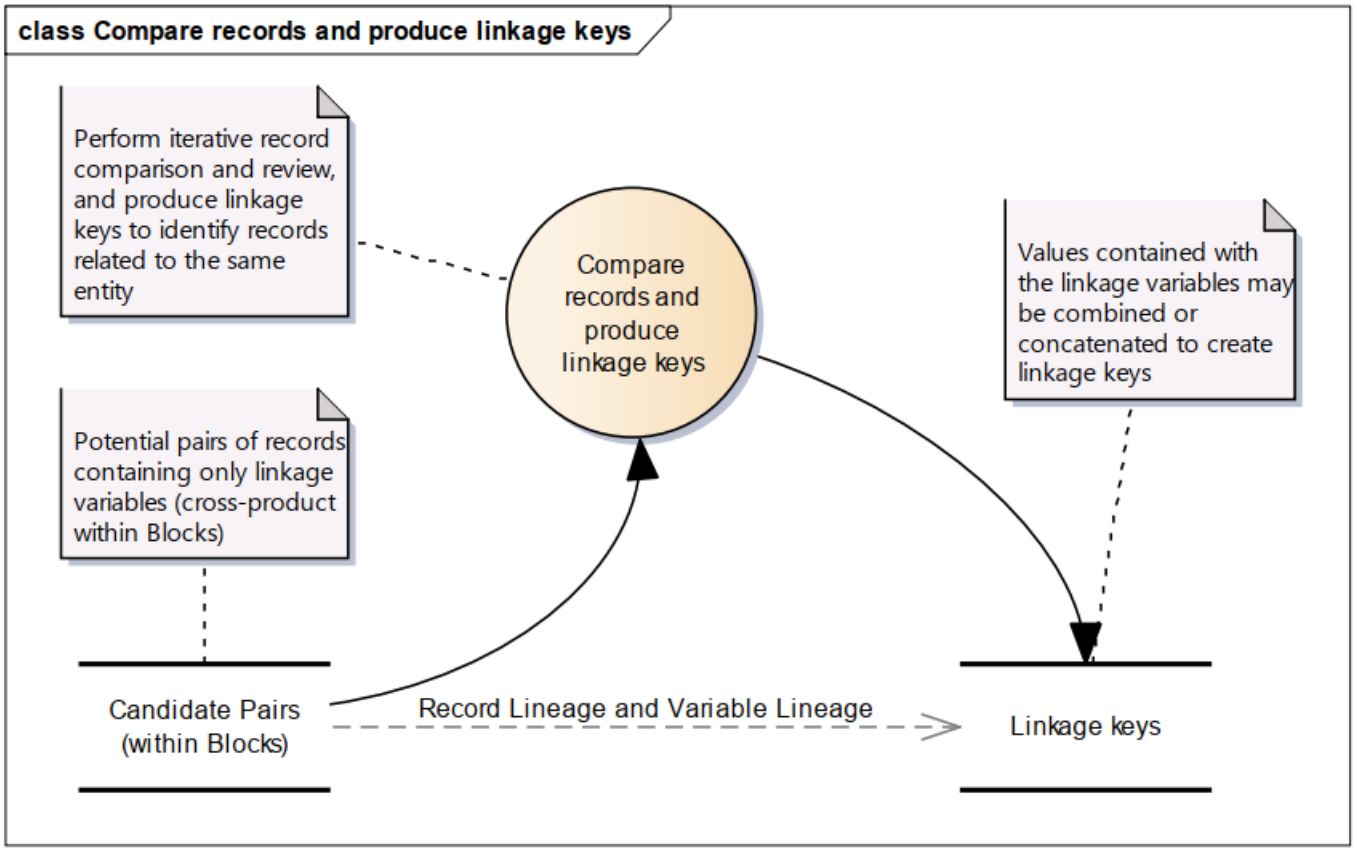
PROV-O representations for record linkage lineage

Produce linkage-ready datasets – PROV-O representation



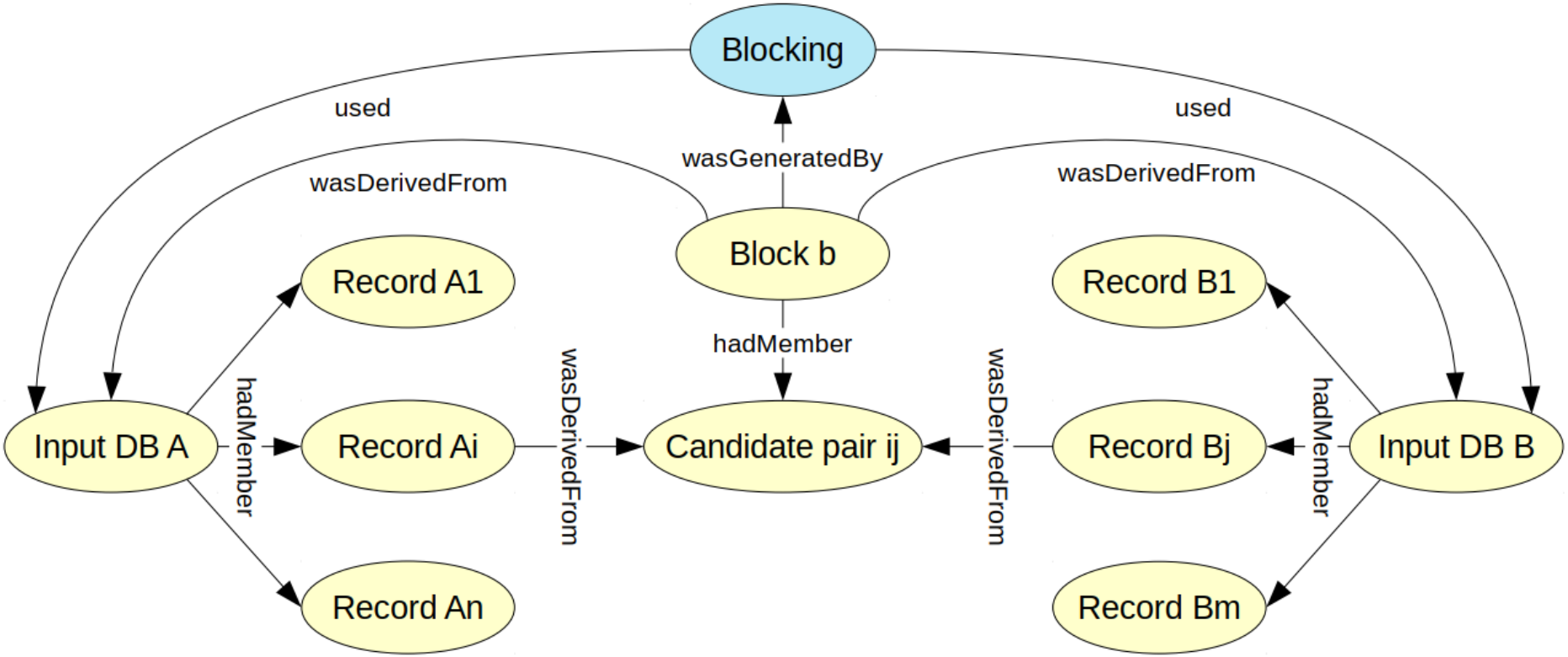
PROV-O representations for record linkage lineage

Produce linkage keys – process



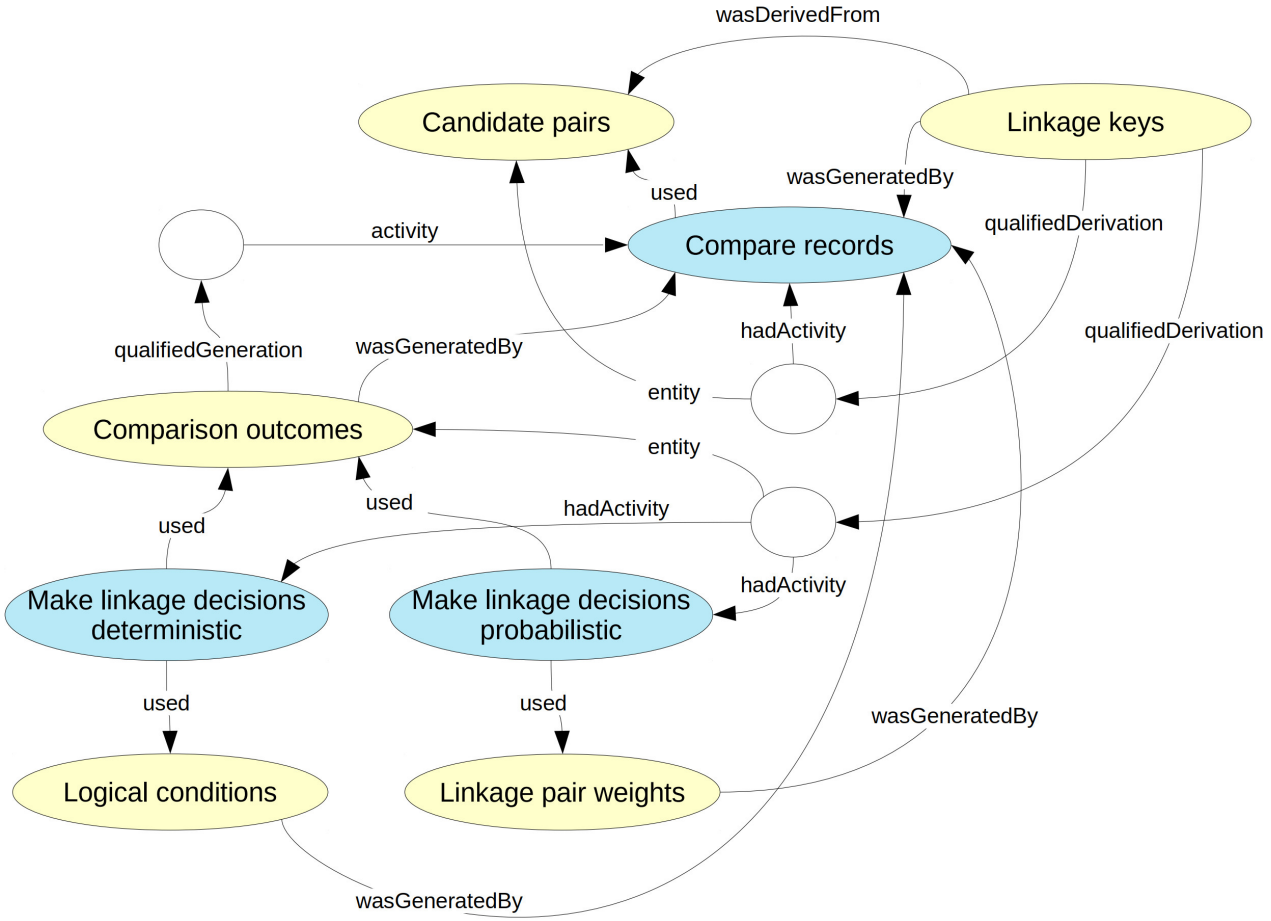
PROV-O representations for record linkage lineage

Produce linkage keys – PROV-O representation – blocking



PROV-O representations for record linkage lineage

Produce linkage keys – PROV-O representation – linking



Conclusions and future work

Proof of concept conclusive

PROV-O can be used to represent the process

Using PROV-O allows to represent coherently the different levels of lineage metadata

The “russian dolls” nature of the PROV-O model implies that metadata can be produced at different levels

Example of queries that can be made

- List output datasets produced from a given data sources
- Which dataset(s) does this record come from?

Conclusions and future work

Future work

Continue work on record linkage, in particular on the representation of methodology

Test how to automate the production of metadata in usual software

Study the possibility to activate metadata (i.e. use it as specification)

Adapt to other statistical operations (e.g. data editing, variable derivation...)

Promote the work in the Official Statistics community

Thank you for your attention

Any questions?