ENVIRONMENT AGENCY AUSTRIA umweltbundesamt

Introduction to Provenance

Part 2: PROV DM

Doron Goldfarb EAA

Keith G Jeffery BGS/UKRI

ENVRI-FAIR has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824068

Prospective Provenance

- Explicit representation of computational steps

  “Recipe to generate data product”

  Source code
  Orchestration of different tools, eg via Shell script
  Workflow description language

- Abstract description, usually not a concrete instantiation

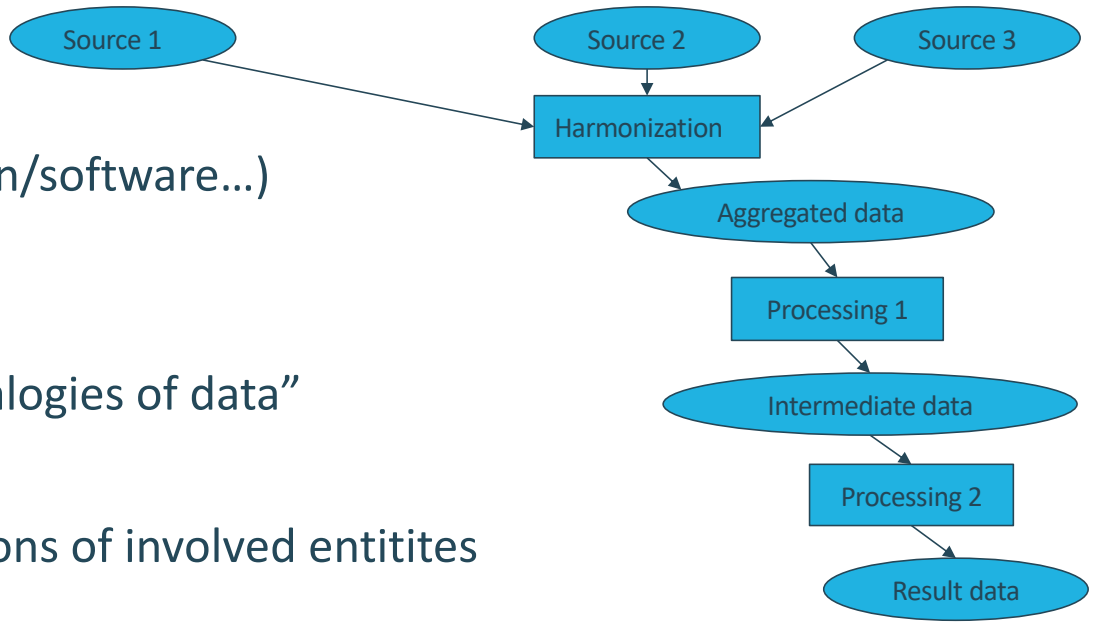
Retrospective Provenance

- Like a detailed log file, tracing history of dataset as far back as possible

- Used sources
  Responsible agents (human/software…)
  Intermediate states
  Applied transformations

  “Causal networks”, “Genealogies of data”

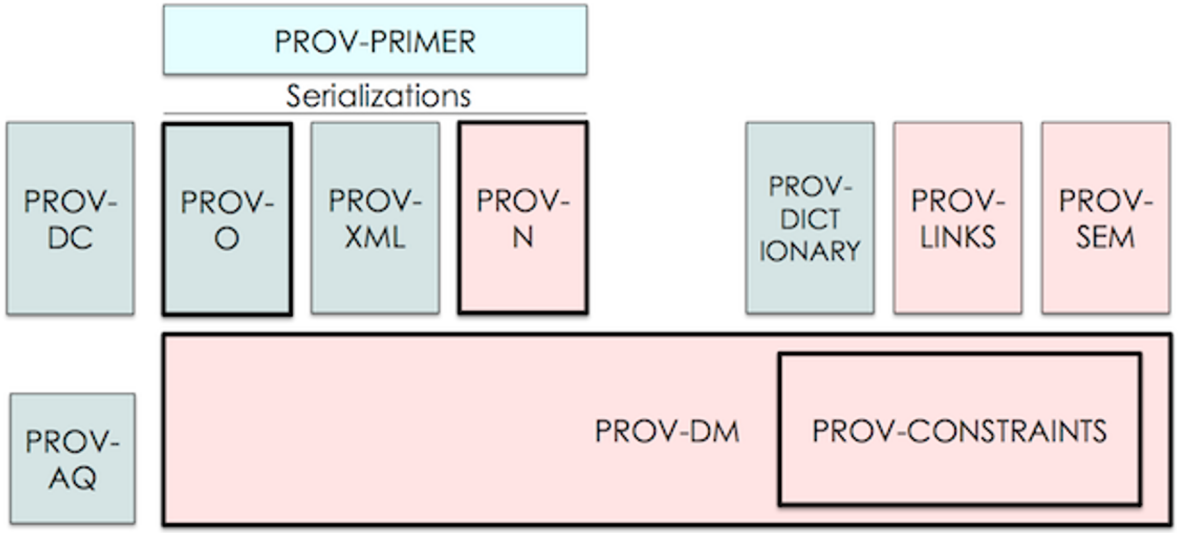
- Contains actual instantiations of involved entitites



W3C PROV

https://www.w3.org/TR/prov-overview/

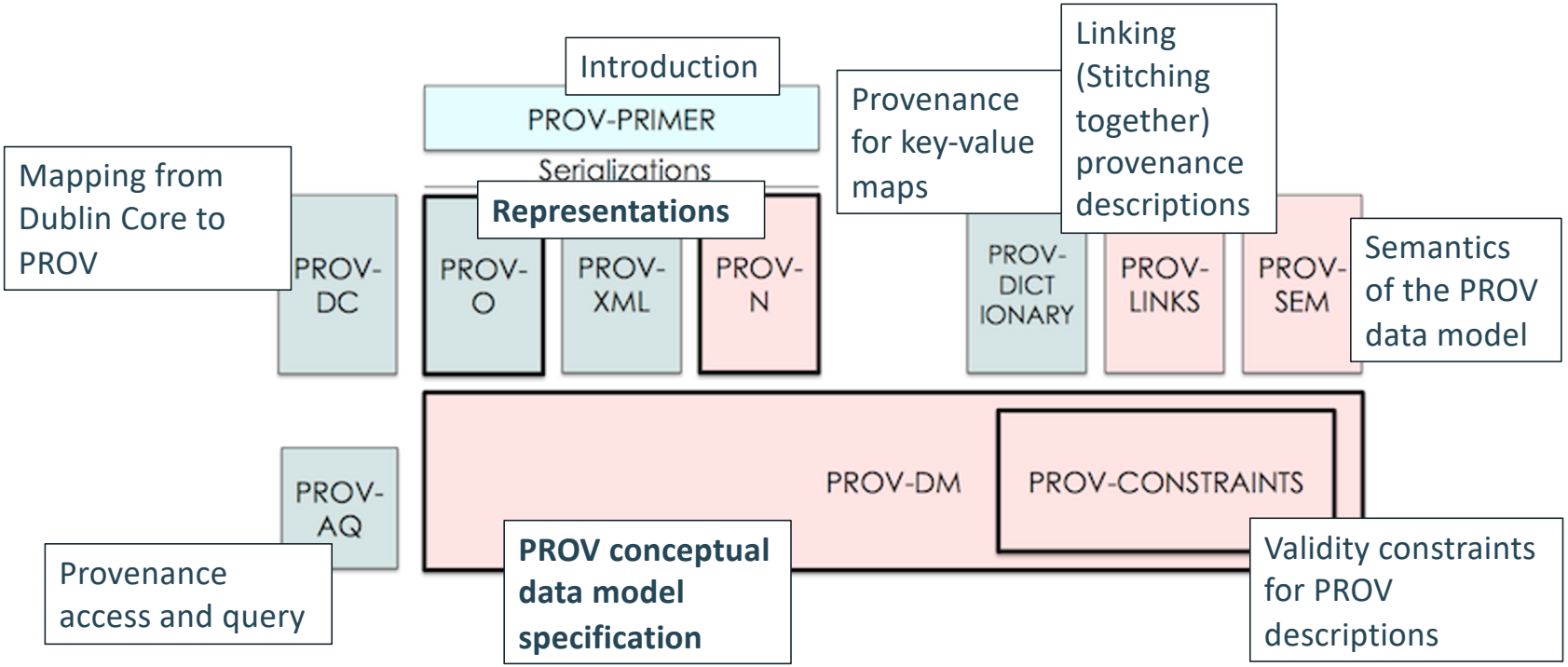
- Dedicated “ecosystem” of specifications centered around PROV data model (PROV-DM), est. 2013



W3C PROV

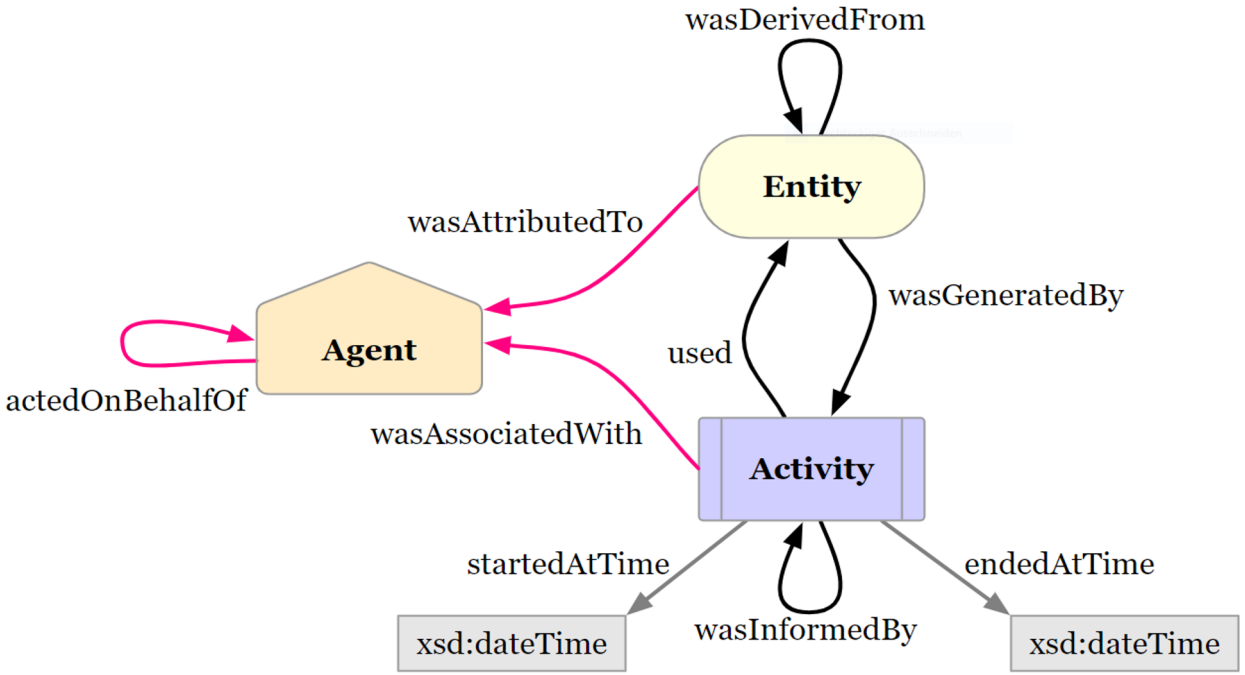
https://www.w3.org/TR/prov-overview/

- Dedicated “ecosystem” of specifications centered around PROV data model (PROV-DM)



PROV-DM

http://www.w3.org/TR/prov-dm/



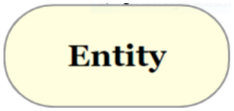
PROV-DM

http://www.w3.org/TR/prov-dm/

- “An entity is a physical, digital, conceptual, or other kind of thing with some fixed aspects; Entities may be real or imaginary”.

- Anything of interest for PROV documentation:

- Document, or part of it
- Dataset
- Concept
- Product
- etc.



PROV-DM

http://www.w3.org/TR/prov-dm/

- “An agent is something that bears some form of responsibility for an activity taking place, for the existence of an entity, or for another agent's activity”.

- Agents receive attribution for entities and are responsible for activities:

- Creator of a document
- Web service processing request
- An organization
- Persons acting on behalf of organization
- etc

PROV-DM

http://www.w3.org/TR/prov-dm/

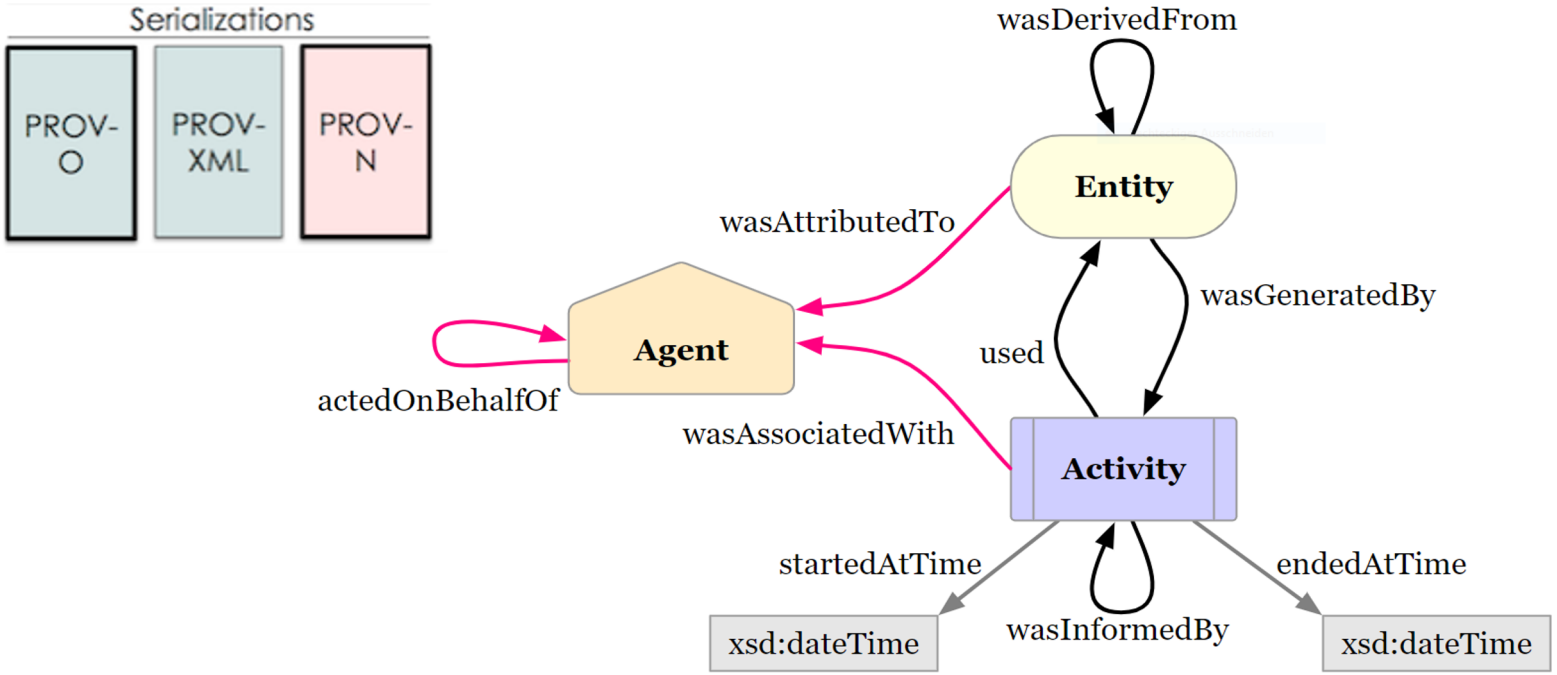
- “An activity is something that occurs over a period of time and acts upon or with entities; it may include consuming, processing, transforming, modifying, relocating, using, or generating entities”.

- Any processes that used or generated entities:

- Computing a result
- Downloading data
- Manually collecting data following a specific procedure / field manual
- etc

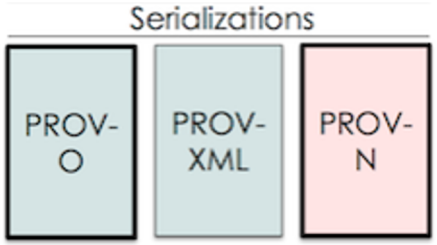
PROV-DM

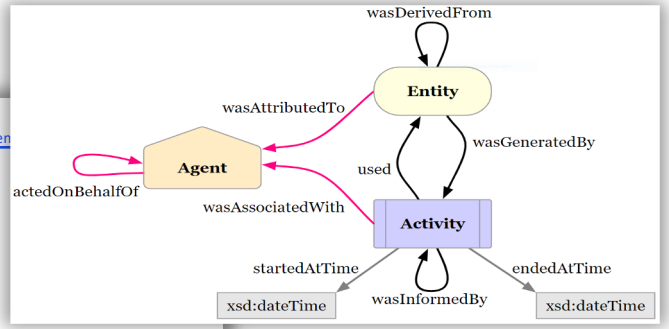
http://www.w3.org/TR/prov-dm/



PROV-DM

http://www.w3.org/TR/prov-dm/





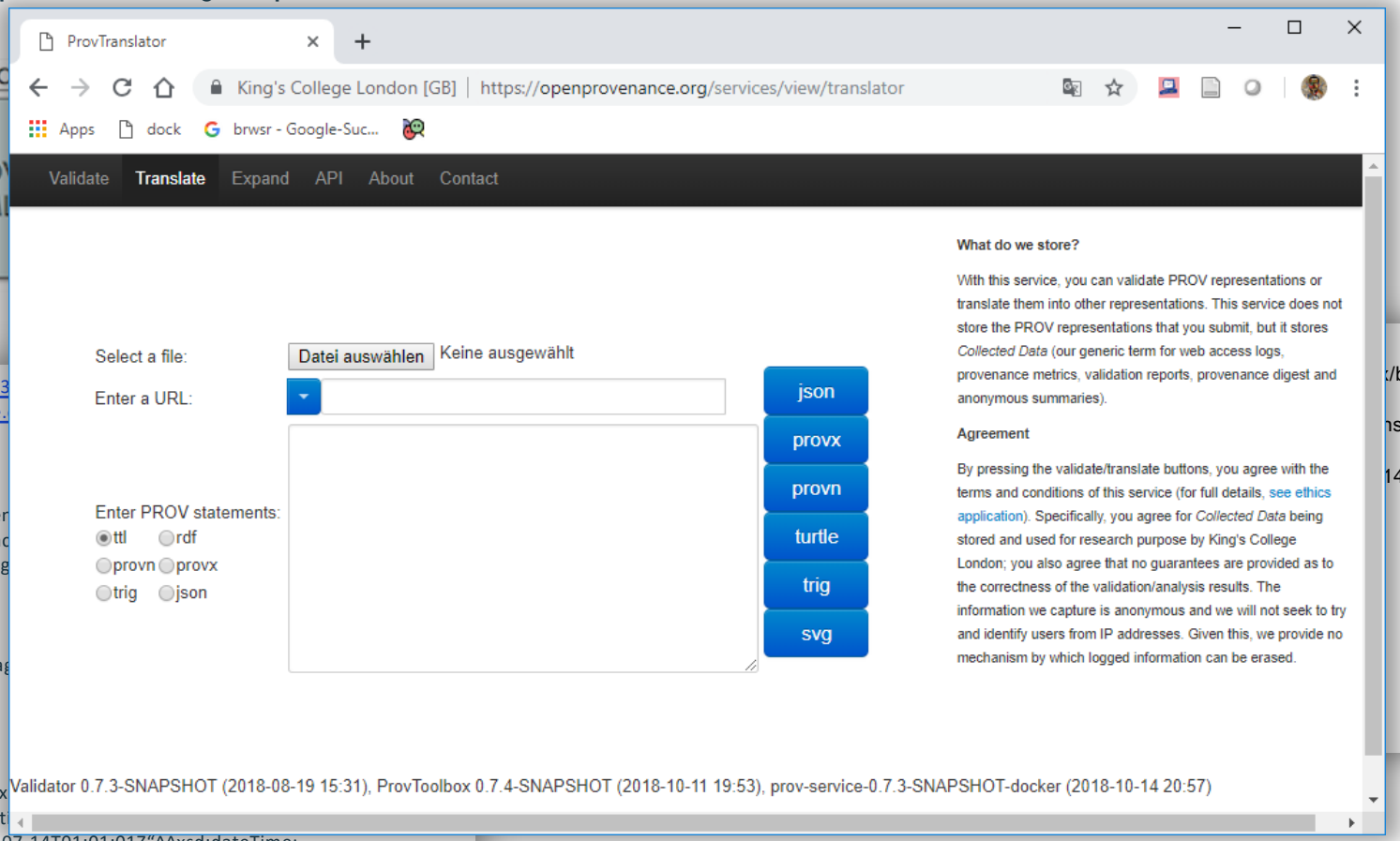

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<prov:document      xmlns:prov="http://www.w3.org/ns/prov#" xmlns:xsi=http://www.w3.org/2001/XMLSchem
                    xmlns:xsd="http://www.w3.org/2001/XMLSchema" xmlns:ex="http://example.org#">
<prov:activity prov:id="ex:activity1">
                    <prov:startTime>2011-07-14T01:01:01Z</prov:startTime>
                    <prov:endTime>2011-08-14T01:01:01Z</prov:endTime>
</prov:activity>
<prov:agent prov:id="ex:agent1"/>
<prov:entity prov:id="ex:entity1"/>
<prov:wasDerivedFrom>
                    <prov:generatedEntity prov:ref="ex:entity1"/>
                    <prov:usedEntity prov:ref="ex:entity1"/>
</prov:wasDerivedFrom>
<prov:wasGeneratedBy>
                    <prov:entity prov:ref="ex:entity1"/>
                    <prov:activity prov:ref="ex:activity1"/>
</prov:wasGeneratedBy>
<prov:wasAttributedTo>
                    <prov:entity prov:ref="ex:entity1"/>
                    <prov:agent prov:ref="ex:agent1"/>
</prov:wasAttributedTo>
<prov:used>
                    <prov:activity prov:ref="ex:activity1"/>
                    <prov:entity prov:ref="ex:entity1"/>
</prov:used>
<prov:wasAssociatedWith>
                    <prov:activity prov:ref="ex:activity1"/>
                    <prov:agent prov:ref="ex:agent1"/>
</prov:wasAssociatedWith>
<prov:wasInformedBy>
                    <prov:informed prov:ref="ex:activity1"/>
                    <prov:informant prov:ref="ex:activity1"/>
</prov:wasInformedBy>
<prov:actedOnBehalfOf>
                    <prov:delegate prov:ref="ex:agent1"/>
                    <prov:responsible prov:ref="ex:agent1"/>
</prov:actedOnBehalfOf>
</prov:document>
```

```
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX ex:   <http://example.org#>

ex:entity1
  a prov:Entity;
  prov:wasDerivedFrom ex:entity1;
  prov:wasGeneratedBy ex:activity1;
  prov:wasAttributedTo ex:agent1;

ex:agent1
  a prov:Agent;
  prov:actedOnBehalfOf ex:agent1;

ex:activity1
  a prov:Activity;
  prov:used ex:entity1;
  prov:wasAssociatedWith ex:agent1;
  prov:wasInformedBy ex:activity1;
  prov:startedAtTime „2011-07-14T01:01:01Z“^^xsd:dateTime;
  prov:endedAtTime „2011-08-14T01:01:01Z“^^xsd:dateTime;
```

```
document
                    prefix bnode <http://openprovenance.org/provtoolbox/bnode/>
                    prefix ex <http://example.org#>
                    prefix rdf <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

                    activity(ex:activity1,2011-07-14T01:01:01Z,2011-08-14T01:01:01Z)
                    agent(ex:agent1)
                    entity(ex:entity1)
                    wasDerivedFrom(ex:entity1, ex:entity1)
                    wasGeneratedBy(ex:entity1,ex:activity1,-)
                    wasAttributedTo(ex:entity1, ex:agent1)
                    used(ex:activity1,ex:entity1,-)
                    wasAssociatedWith(ex:activity1,ex:agent1,-)
                    wasInformedBy(ex:activity1,ex:activity1)
                    actedOnBehalfOf(ex:agent1,ex:agent1,-)
endDocument
```

PROV-DM

http://www.w3.org/TR/prov-dm/




```
PREFIX prov: <http://www.w3
PREFIX ex:   <http://example.

ex:entity1
  a prov:Entity;
  prov:wasDerivedFrom ex:er
  prov:wasGeneratedBy ex:ac
  prov:wasAttributedTo ex:ag

ex:agent1
  a prov:Agent;
  prov:actedOnBehalfOf ex:ag

ex:activity1
  a prov:Activity;
  prov:used ex:entity1;
  prov:wasAssociatedWith ex
  prov:wasInformedBy ex:act
  prov:startedAtTime „2011-07-14T01:01:01Z“^^xsd:dateTime;
  prov:endedAtTime „2011-08-14T01:01:01Z“^^xsd:dateTime;
```

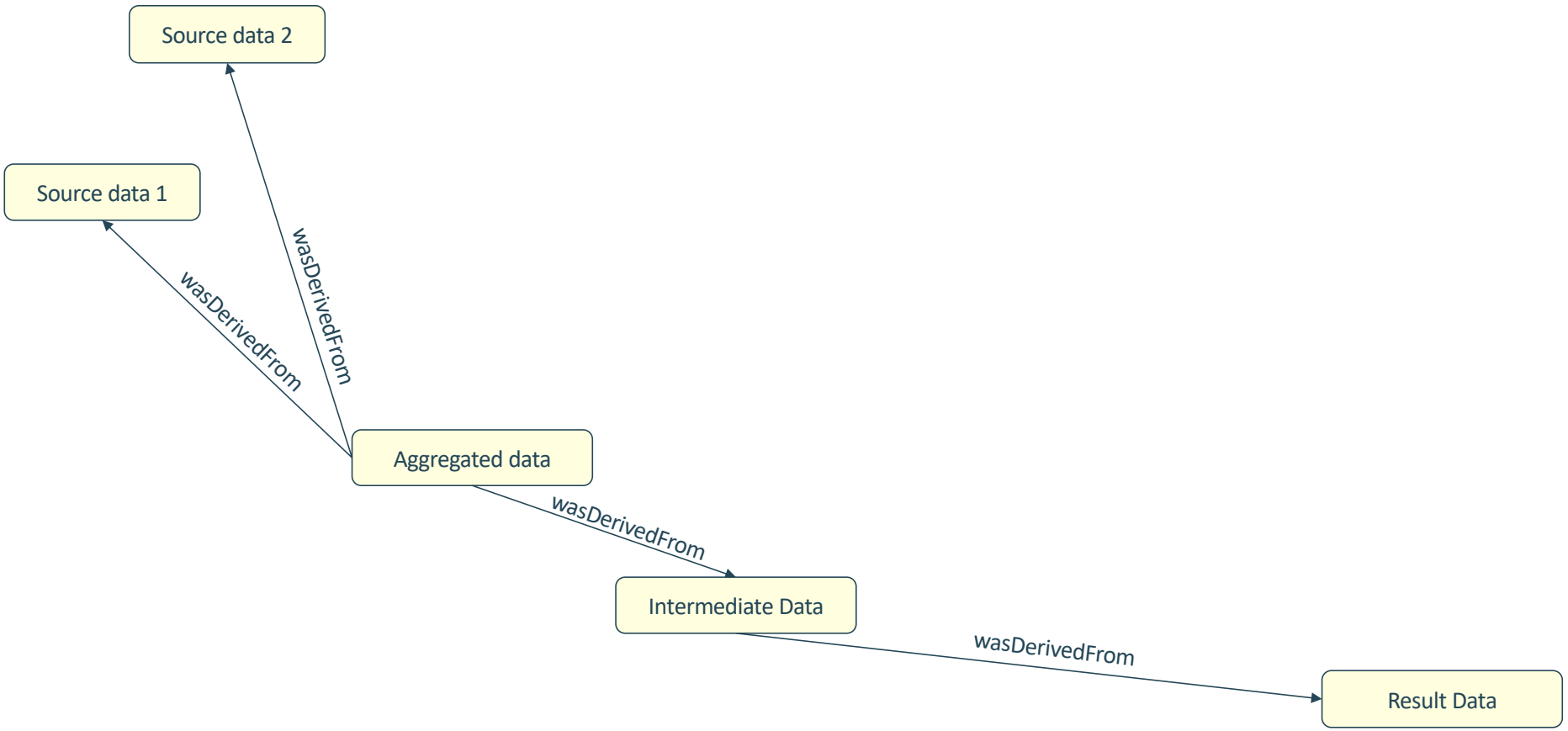
```
/bnode/>

ns#>

14T01:01:01Z)
```

W3C Prov

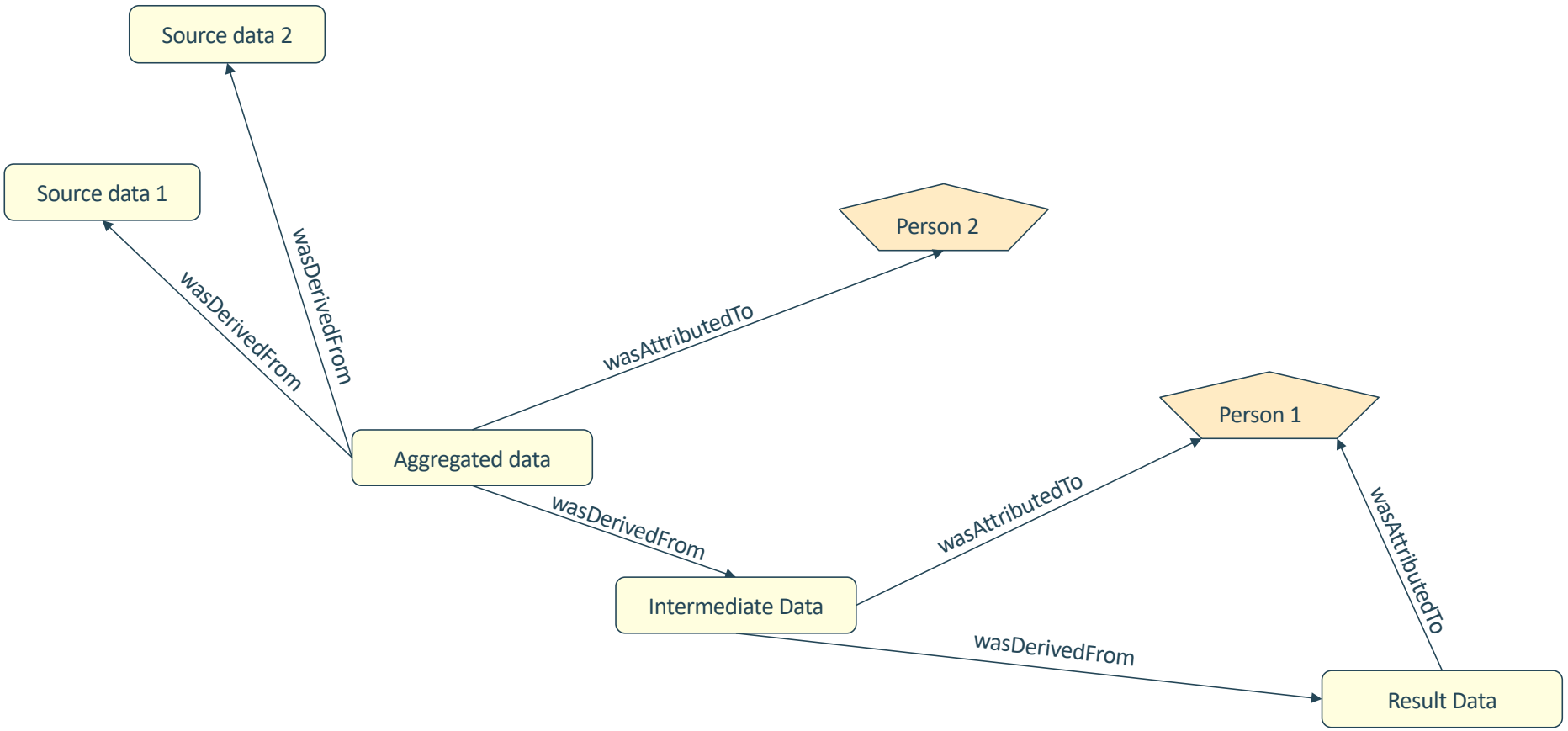
Documentation and Specifications

- PROV Overview(http://www.w3.org/TR/prov-overview/)
- PROV Primer (http://www.w3.org/TR/prov-primer/)
- PROV Data Model(*) (http://www.w3.org/TR/prov-dm/)
- PROV Constraints(*) (http://www.w3.org/TR/prov-constraints/)
- PROV Semantics(http://www.w3.org/TR/prov-sem/)
- PROV Notation(*) (http://www.w3.org/TR/prov-n/)
- PROV Ontology(*) (http://www.w3.org/TR/prov-o/)
- PROV XML Serialization(http://www.w3.org/TR/prov-xml/)
- PROV Access and Query (http://www.w3.org/TR/prov-aq/)
- PROV DC Mapping (http://www.w3.org/TR/prov-dc/)
- PROV Links (http://www.w3.org/TR/prov-links/)
- PROV Dictionary (http://www.w3.org/TR/prov-dictionary/)
- PROV Implementations(http://www.w3.org/TR/prov-implementations/)

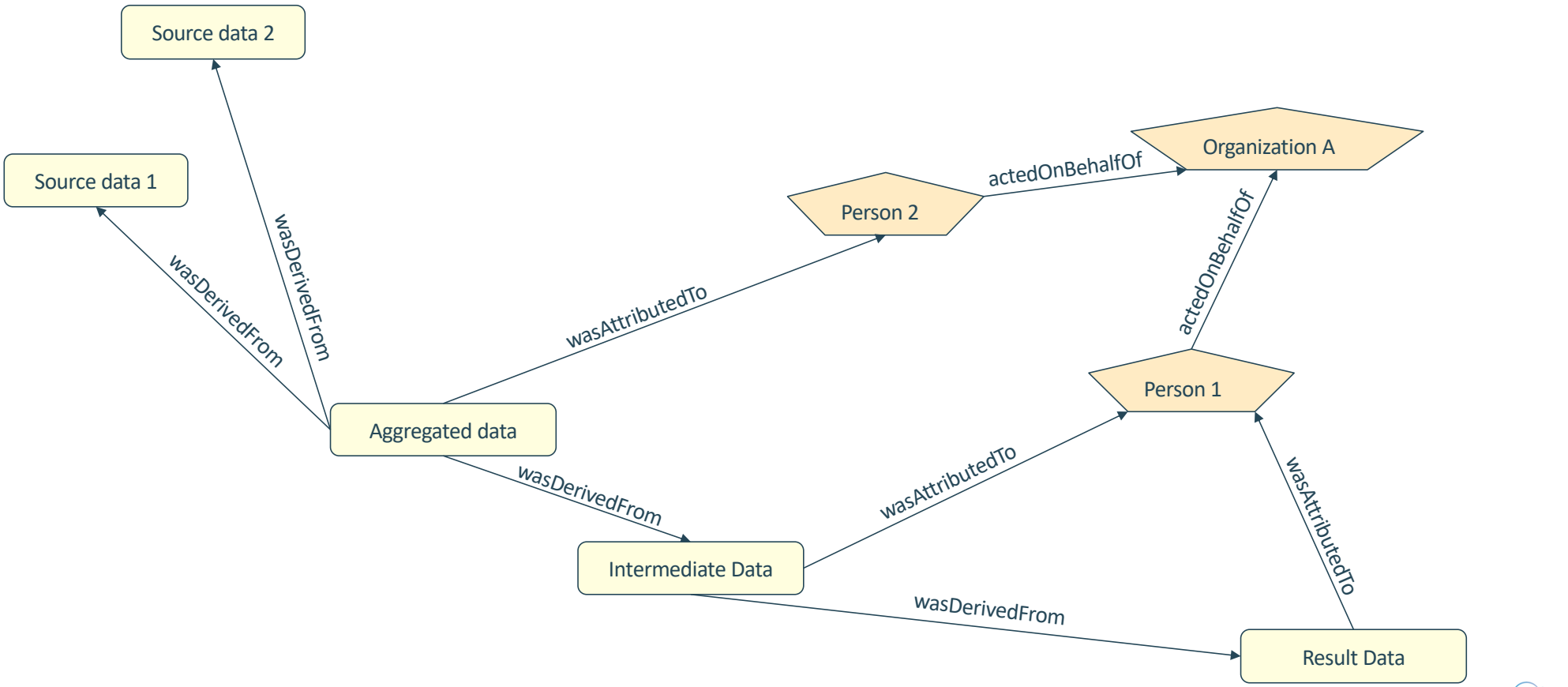
Data flow view



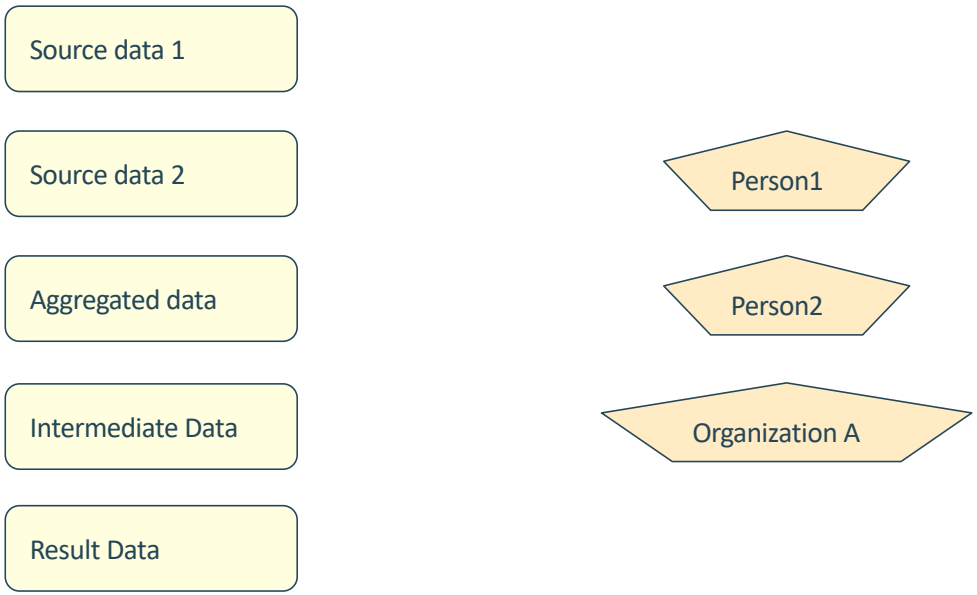
Data flow & Responsibility view



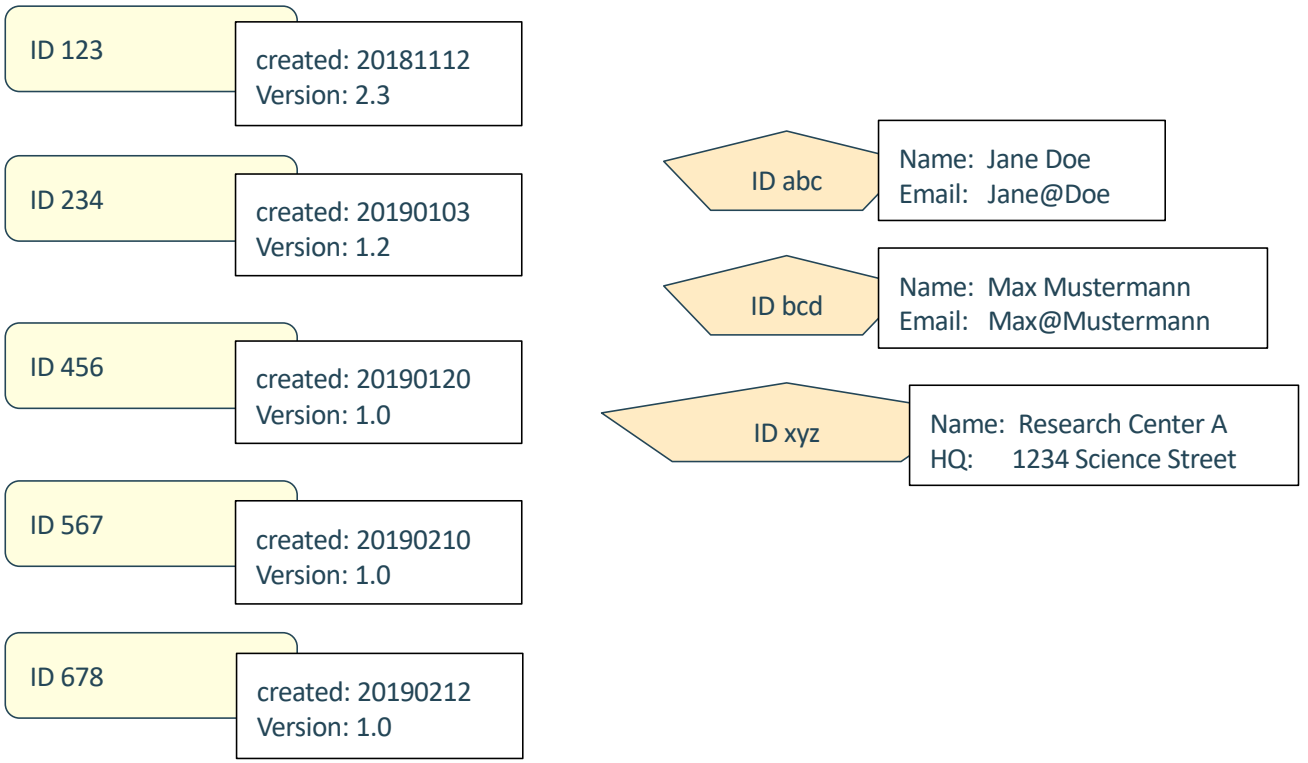
Data flow & Responsibility view

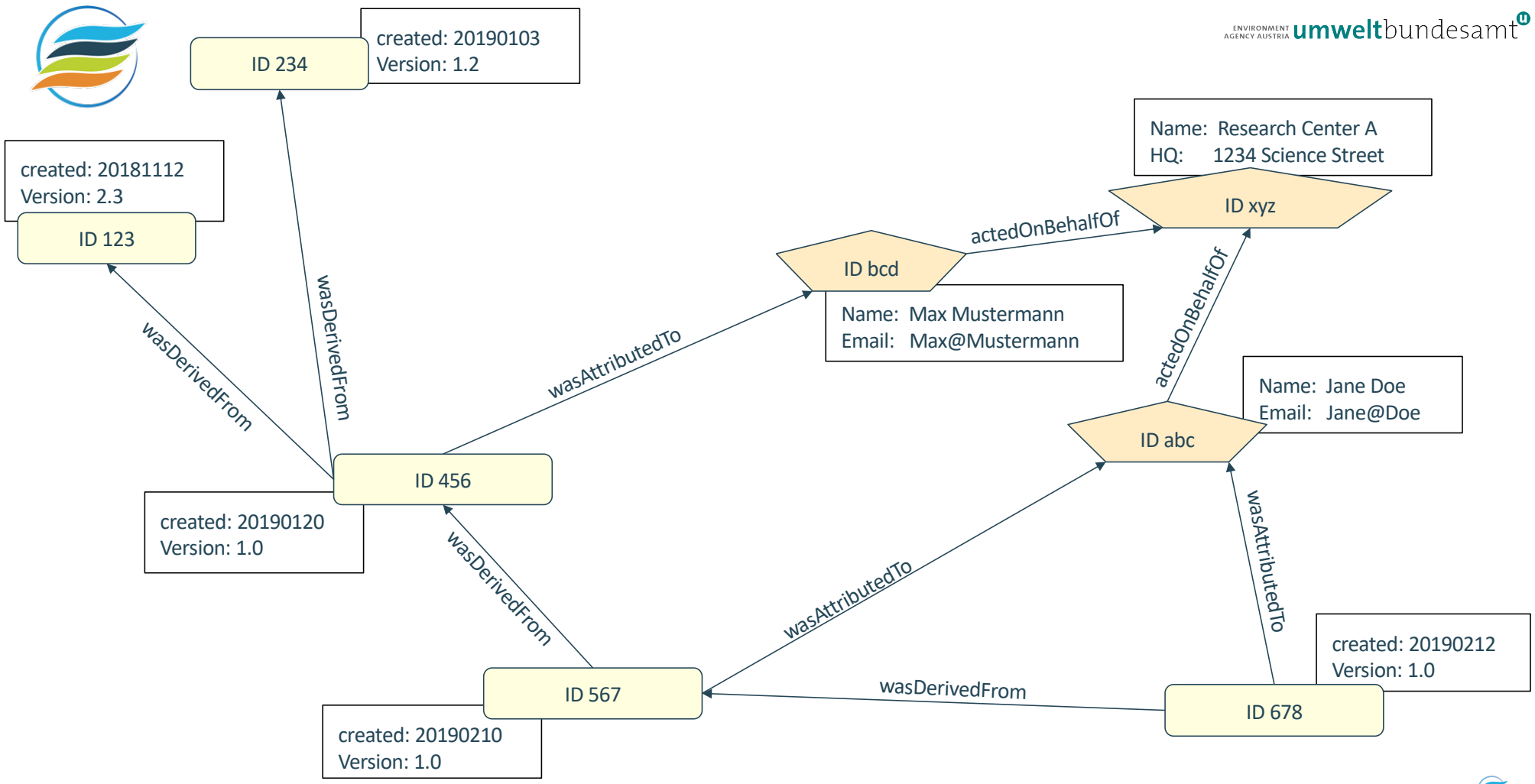


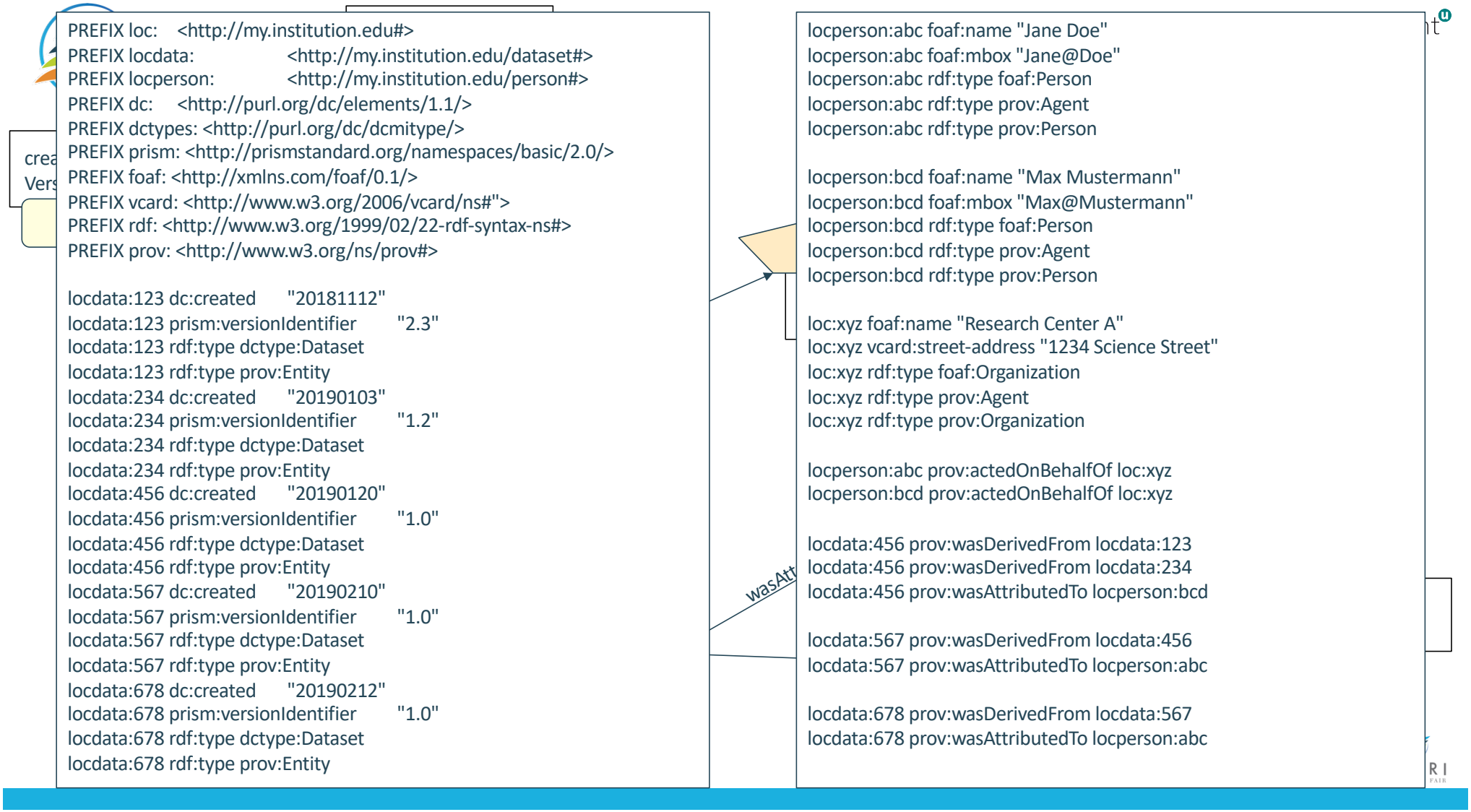
Linking existing descriptions

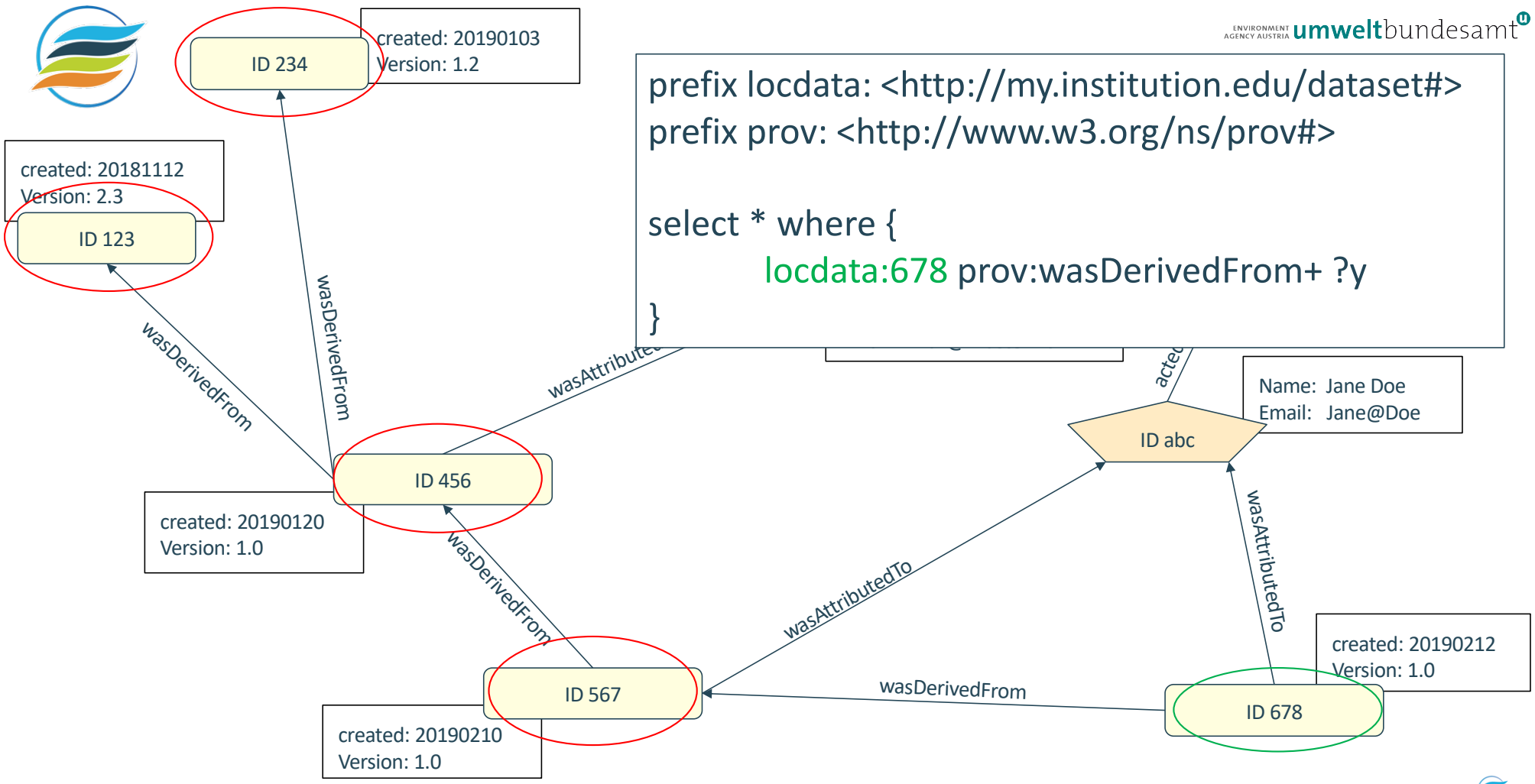


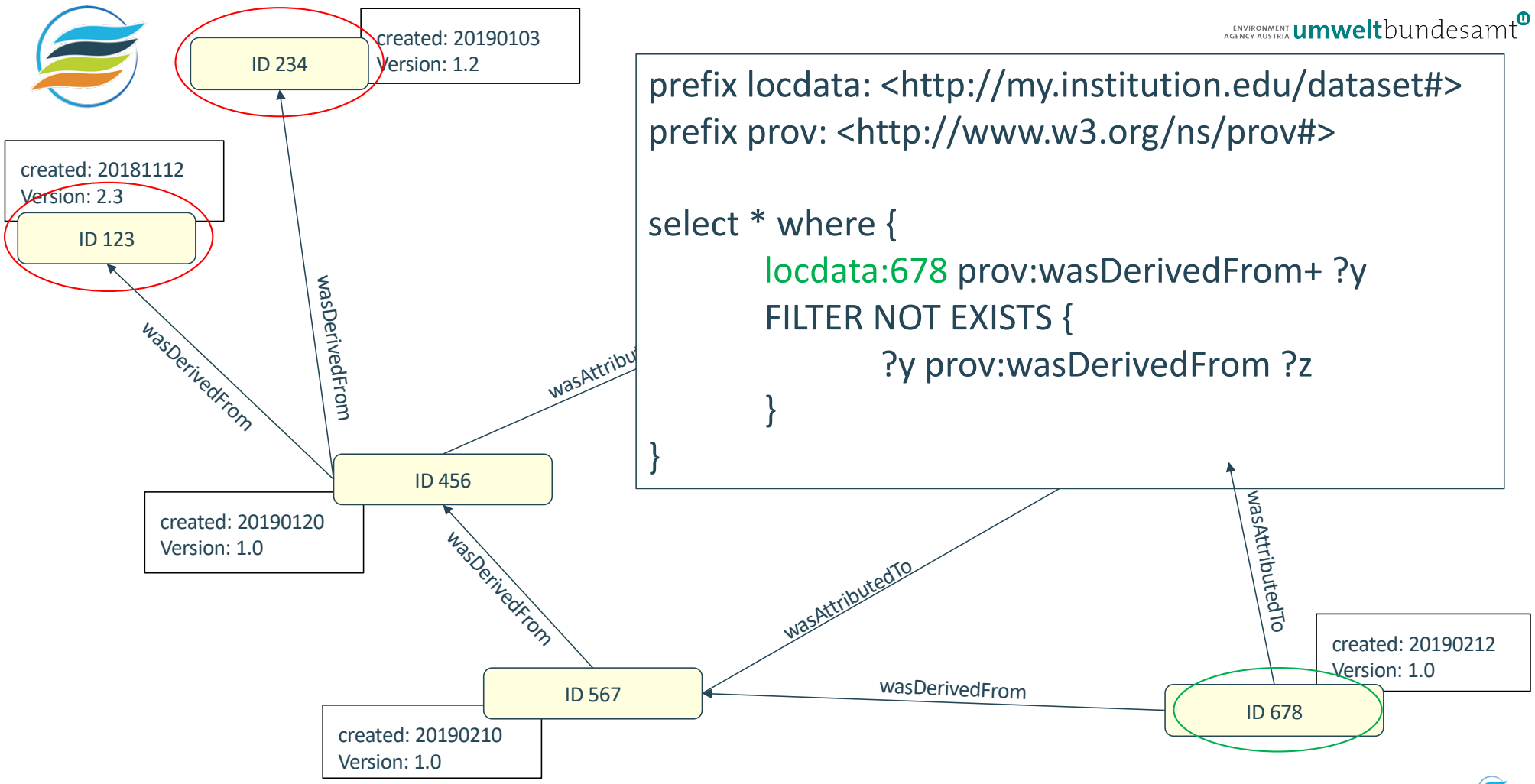
Linking existing descriptions



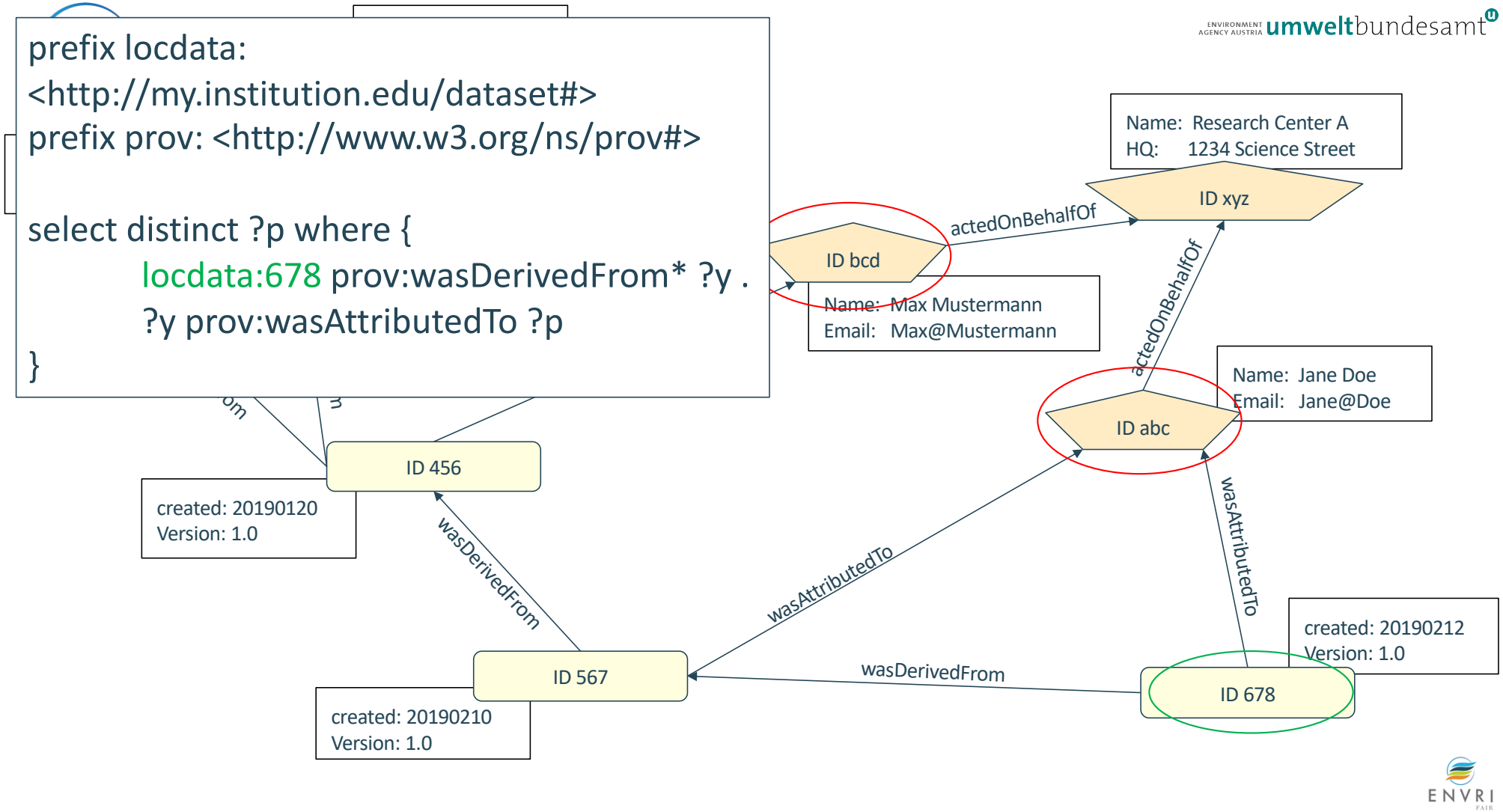

ID 234
created: 20190103
Version: 1.2
created: 20181112
Version: 2.3
ID 123
wasDerivedFrom
wasDerivedFrom
wasAttributedTo
ID bcd
actedOnBehalfOf
Name: Research Center A
HQ: 1234 Science Street
ID xyz
Name: Max Mustermann
Email: Max@Mustermann
actedOnBehalfOf
Name: Jane Doe
Email: Jane@Doe
ID abc
ID 456
created: 20190120
Version: 1.0
wasDerivedFrom
wasAttributedTo
wasAttributedTo
created: 20190212
Version: 1.0
ID 567
wasDerivedFrom
ID 678
created: 20190210
Version: 1.0


PREFIX loc: <http://my.institution.edu#>
PREFIX locdata: <http://my.institution.edu/dataset#>
PREFIX locperson: <http://my.institution.edu/person#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX dctypes: <http://purl.org/dc/dcmitype/>
PREFIX prism: <http://prismstandard.org/namespaces/basic/2.0/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX vcard: <http://www.w3.org/2006/vcard/ns#">
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX prov: <http://www.w3.org/ns/prov#>
locdata:123 dc:created "20181112"
locdata:123 prism:versionIdentifier "2.3"
locdata:123 rdf:type dctype:Dataset
locdata:123 rdf:type prov:Entity
locdata:234 dc:created "20190103"
locdata:234 prism:versionIdentifier "1.2"
locdata:234 rdf:type dctype:Dataset
locdata:234 rdf:type prov:Entity
locdata:456 dc:created "20190120"
locdata:456 prism:versionIdentifier "1.0"
locdata:456 rdf:type dctype:Dataset
locdata:456 rdf:type prov:Entity
locdata:567 dc:created "20190210"
locdata:567 prism:versionIdentifier "1.0"
locdata:567 rdf:type dctype:Dataset
locdata:567 rdf:type prov:Entity
locdata:678 dc:created "20190212"
locdata:678 prism:versionIdentifier "1.0"
locdata:678 rdf:type dctype:Dataset
locdata:678 rdf:type prov:Entity
locperson:abc foaf:name "Jane Doe"
locperson:abc foaf:mbox "Jane@Doe"
locperson:abc rdf:type foaf:Person
locperson:abc rdf:type prov:Agent
locperson:abc rdf:type prov:Person
locperson:bcd foaf:name "Max Mustermann"
locperson:bcd foaf:mbox "Max@Mustermann"
locperson:bcd rdf:type foaf:Person
locperson:bcd rdf:type prov:Agent
locperson:bcd rdf:type prov:Person
loc:xyz foaf:name "Research Center A"
loc:xyz vcard:street-address "1234 Science Street"
loc:xyz rdf:type foaf:Organization
loc:xyz rdf:type prov:Agent
loc:xyz rdf:type prov:Organization
locperson:abc prov:actedOnBehalfOf loc:xyz
locperson:bcd prov:actedOnBehalfOf loc:xyz
locdata:456 prov:wasDerivedFrom locdata:123
locdata:456 prov:wasDerivedFrom locdata:234
locdata:456 prov:wasAttributedTo locperson:bcd
locdata:567 prov:wasDerivedFrom locdata:456
locdata:567 prov:wasAttributedTo locperson:abc
locdata:678 prov:wasDerivedFrom locdata:567
locdata:678 prov:wasAttributedTo locperson:abc

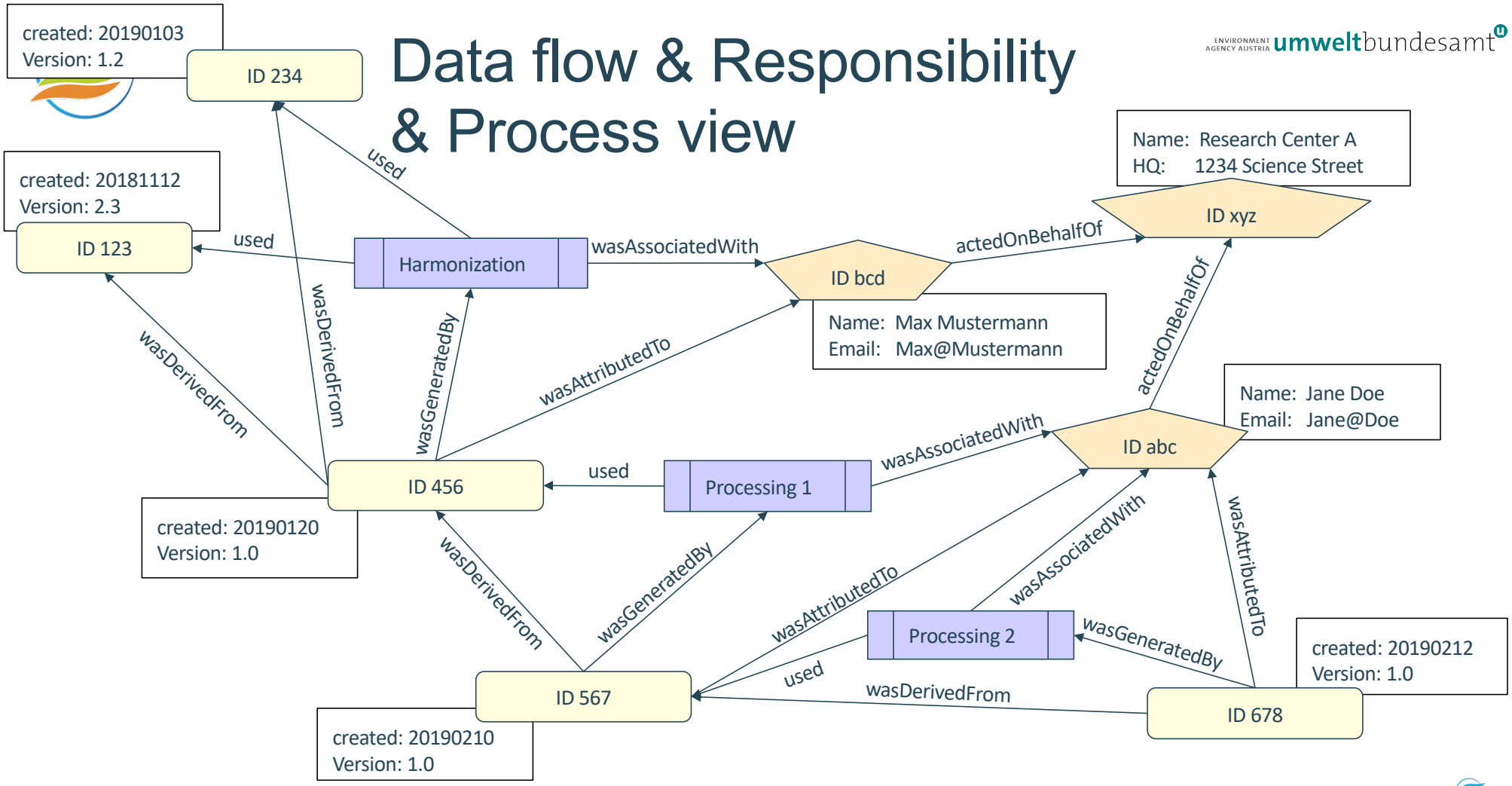

ID 234
created: 20190103
Version: 1.2
created: 20181112
Version: 2.3
ID 123
wasDerivedFrom
wasDerivedFrom
prefix locdata: <http://my.institution.edu/dataset#>
prefix prov: <http://www.w3.org/ns/prov#>
select * where {
locdata:678 prov:wasDerivedFrom+ ?y
}
Name: Jane Doe
Email: Jane@Doe
ID abc
ID 456
created: 20190120
Version: 1.0
wasDerivedFrom
wasAttributedTo
wasAttributedTo
created: 20190212
Version: 1.0
ID 567
wasDerivedFrom
ID 678
created: 20190210
Version: 1.0


ID 234
created: 20190103
Version: 1.2
created: 20181112
Version: 2.3
ID 123
wasDerivedFrom
wasDerivedFrom
wasAttribu
ID 456
created: 20190120
Version: 1.0
wasDerivedFrom
wasAttributedTo
wasAttributedTo
created: 20190212
Version: 1.0
ID 567
wasDerivedFrom
ID 678
created: 20190210
Version: 1.0
prefix locdata: <http://my.institution.edu/dataset#>
prefix prov: <http://www.w3.org/ns/prov#>
select * where {
locdata:678 prov:wasDerivedFrom+ ?y
FILTER NOT EXISTS {
?y prov:wasDerivedFrom ?z
}
}

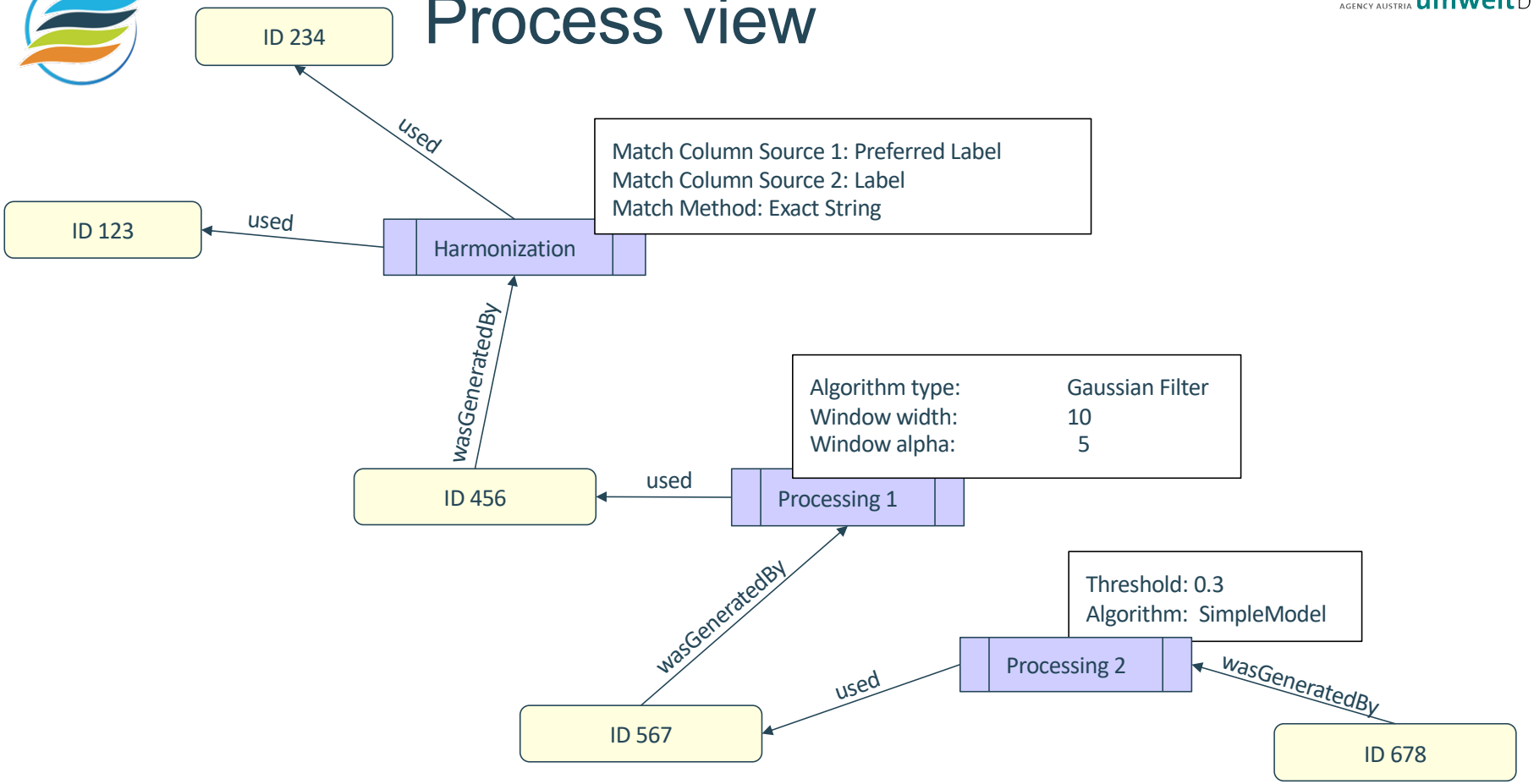
```
prefix locdata:
<http://my.institution.edu/dataset#>
prefix prov: <http://www.w3.org/ns/prov#>

select distinct ?p where {
        locdata:678 prov:wasDerivedFrom* ?y .
        ?y prov:wasAttributedTo ?p
}
```



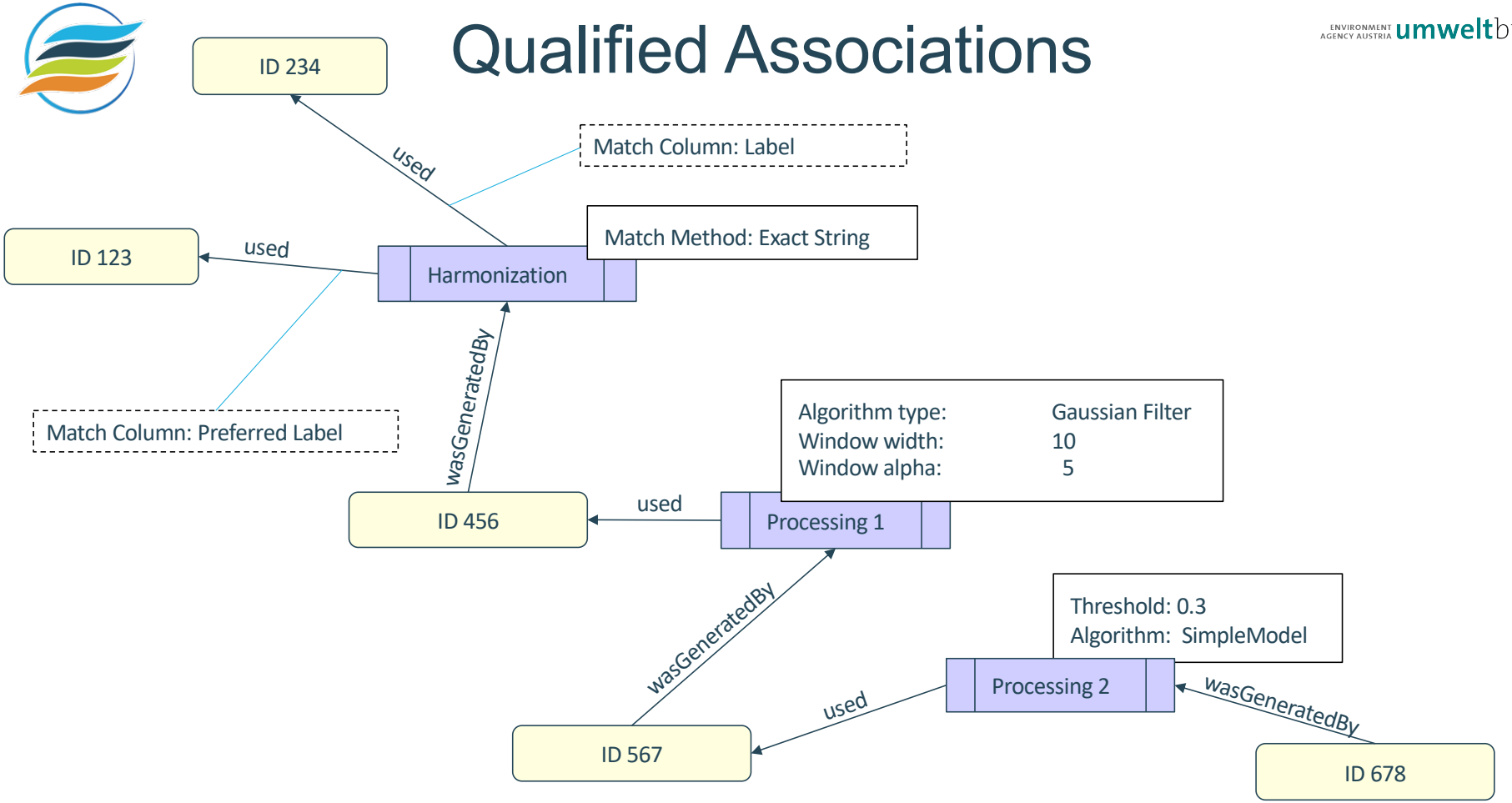
Data flow & Responsibility & Process view



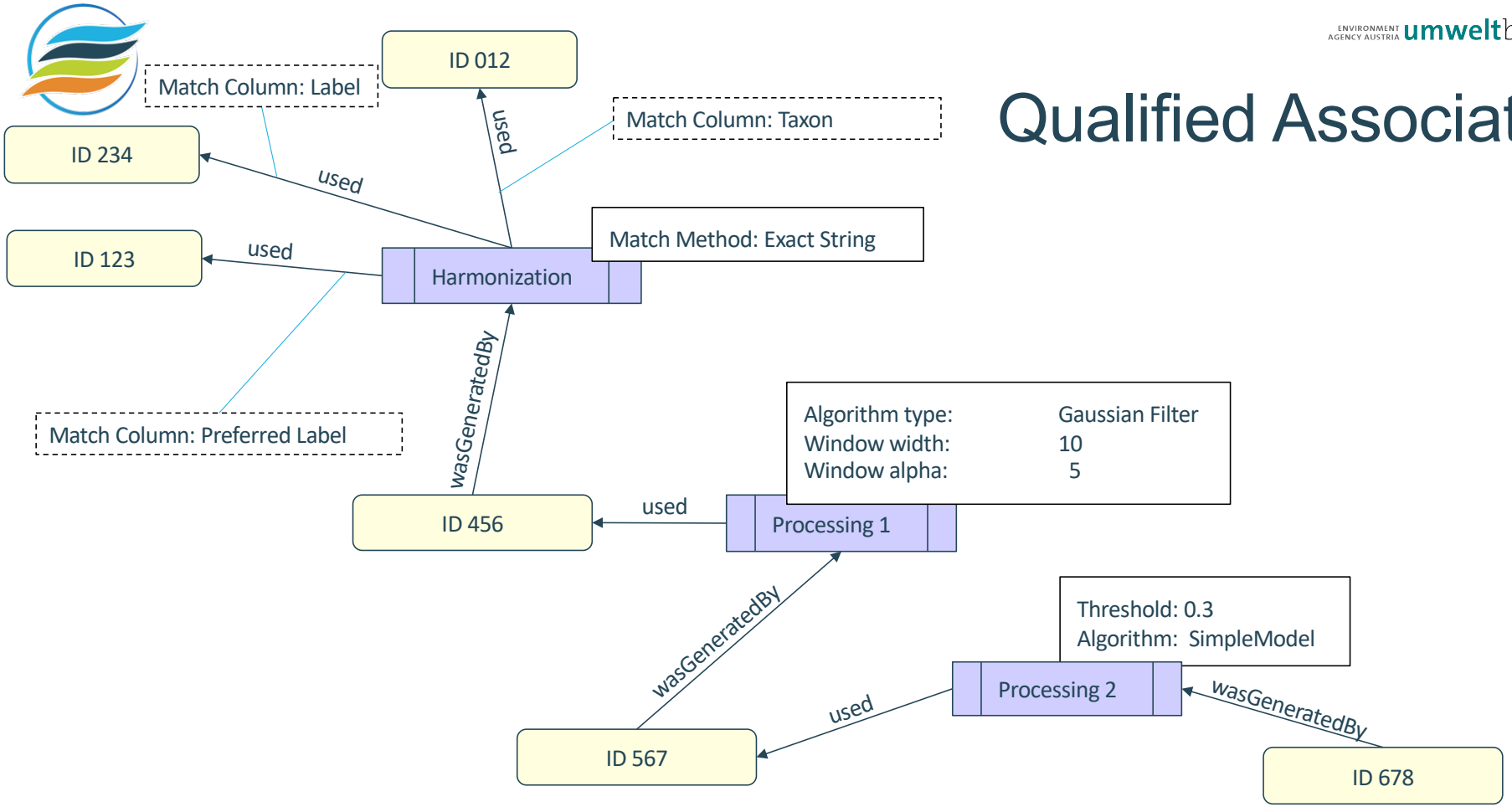
Process view



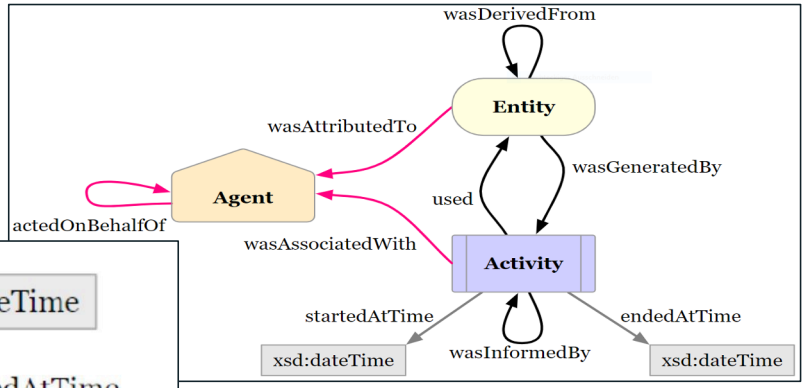
Qualified Associations

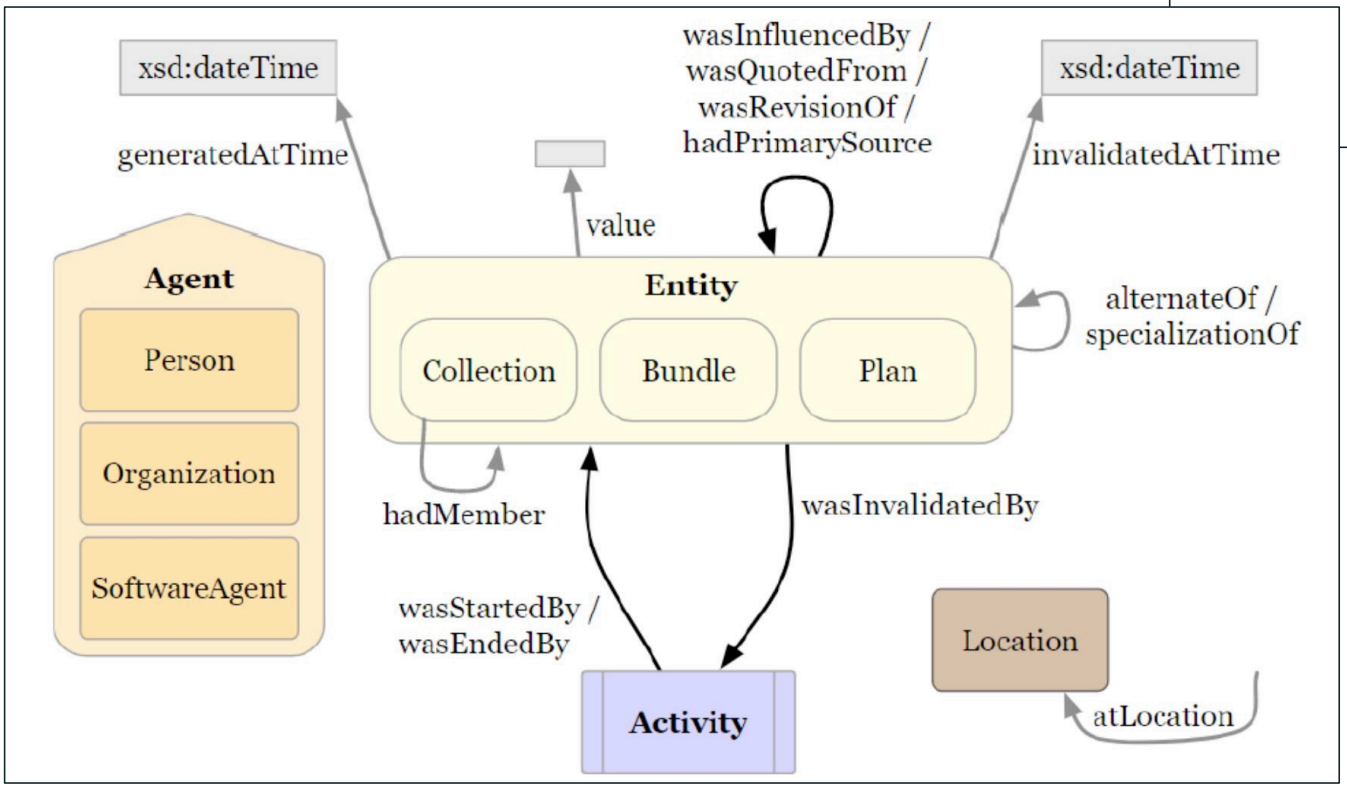


Qualified Associations

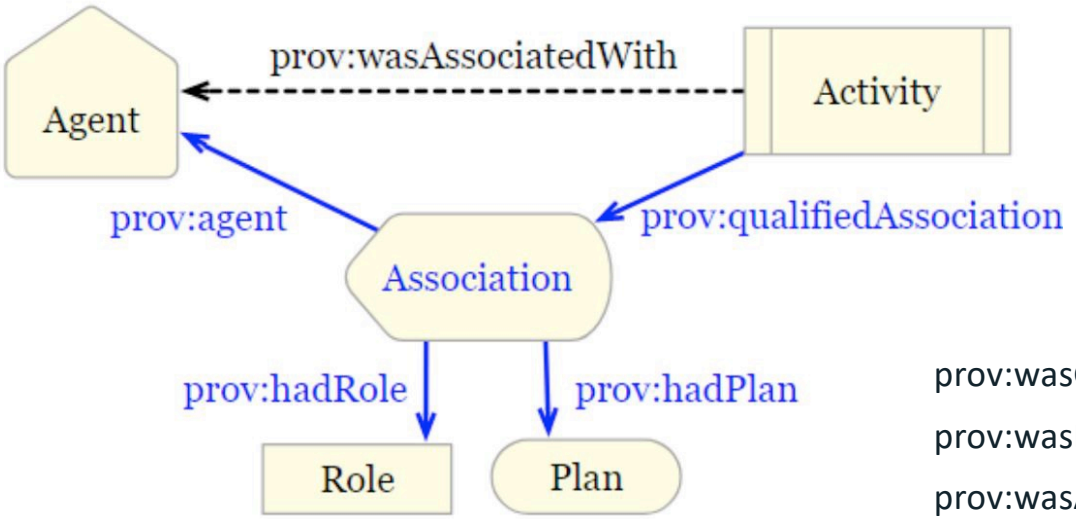


Expanded Terms





Qualified Terms



prov:wasGeneratedBy	→ prov:Generation
prov:wasDerivedBy	→ prov:Derivation
prov:wasAttributedTo	→ prov:Attribution
prov:wasInformedBy	→ prov:Communication
prov:actedOnBehalfOf	→ prov:Delegation
prov:wasDerivedBy	→ prov:Derivation
prov:used	→ prov:Usage

Modeling Provenance

Iterative Modeling

- Identify main data elements within workflow
- Model relationships using PROV (Start at high level)
- Modify workflow implementation to provide PROV
- Query/Refine – Test augmented workflow, query results and refine model if necessary

Modeling Provenance

Identification

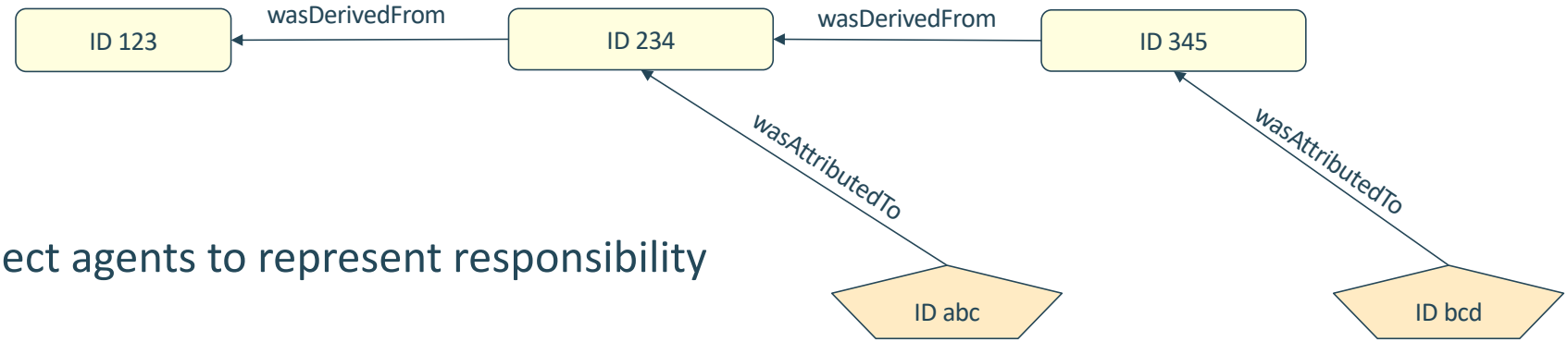
- Create identifier scheme → Ideally, use resolvable http-URIs
- Specify fixed aspects of each item, e.g. version number, identifier, location, type, context
- Create type hierarchy for annotating data items
- Create identifier for each data item and use it for provenance trace
  (Ideally for other things too)

Modeling Provenance

- Start with data flow view, end with process flow view
  - Describe data flow using prov:wasDerivedFrom relations between prov:Entitites

Modeling Provenance

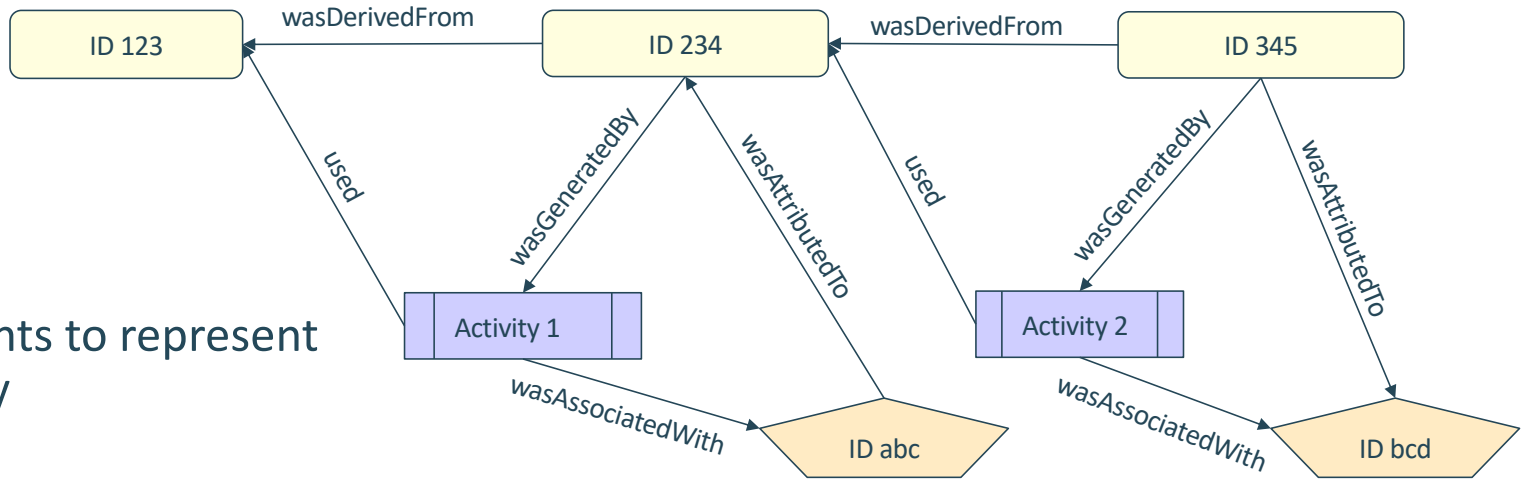
- Start with data flow view, end with process flow view
  - Describe data flow using prov:wasDerivedFrom relations between prov:Entitites



  - Connect agents to represent responsibility

Modeling Provenance

- Start with data flow view, end with process flow view
  - Describe data flow using prov:wasDerivedFrom relations between prov:Entitites
  - Connect agents to represent responsibility
  - Add activities where required

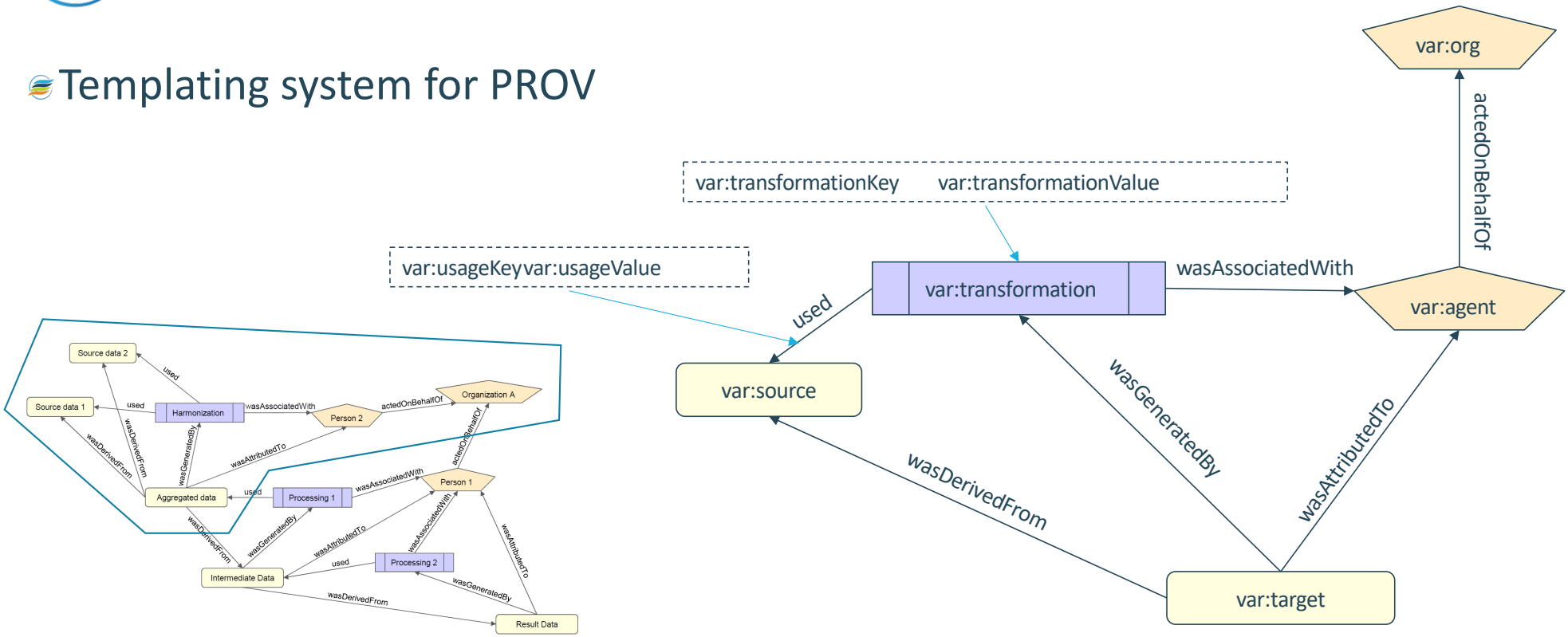


Generating Provenance Data

- Supported by existing workflow management systems
  e.g. Taverna plugin
- Dedicated libraries to annotate scripting environments
  e.g. “YesWorkflow”
- Direct implementation in own software
- Extraction from existing log output
  e.g. PROV-TEMPLATE

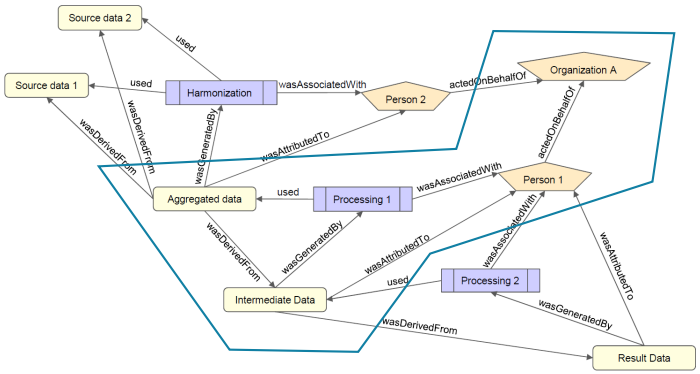
PROV-TEMPLATE

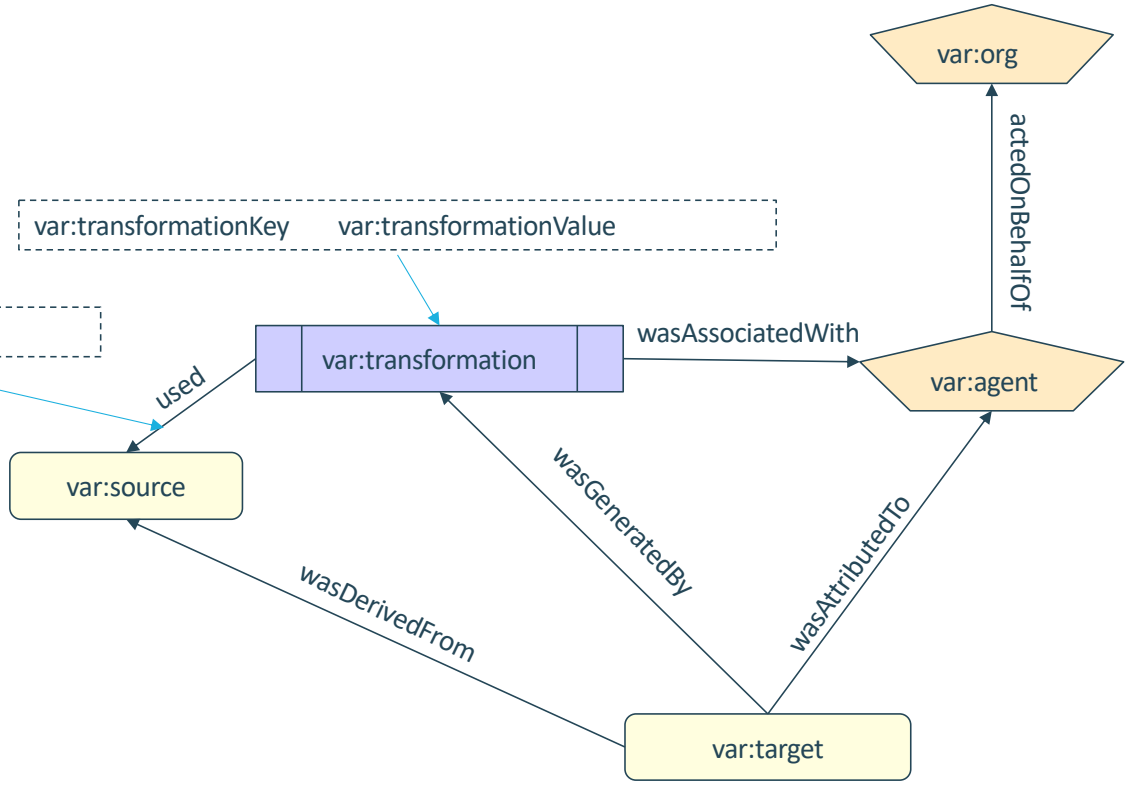
- Templating system for PROV



PROV-TEMPLATE

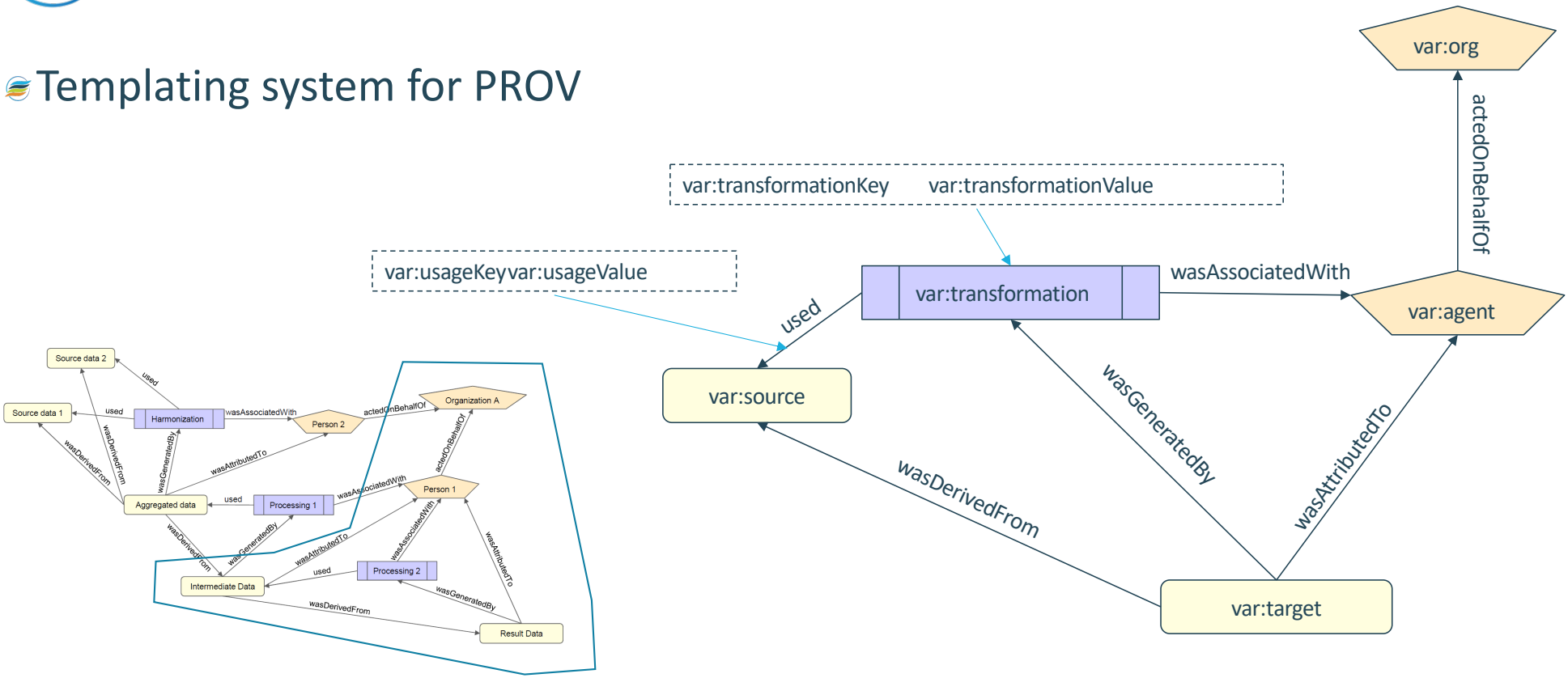
- Templating system for PROV





PROV-TEMPLATE

- Templating system for PROV



Making Provenance Available

- Provenance should be made available together with data product

- PROV-AQ specification (https://www.w3.org/TR/prov-aq/)

  Direct Access → Resolve provided provenance-URI
  Indirect Access → Retrieve via provided query-service URI (e.g. SPARQL)

  Mechanisms/semantics for provision of provenance-URI and query information via HTTP headers

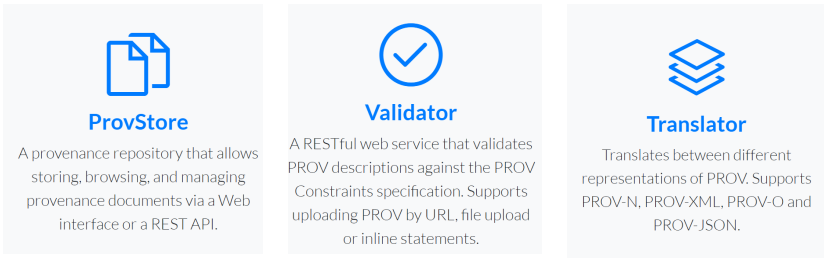
- General considerations: How to treat “inherited” provenance

  Provenance document for current dataset **links to provenance documents** of source datasets
  Provenance document for current dataset **contains provenance information** of source datasets

Provenance Tools

https://openprovenance.org

ProvToolbox	Java package
ProvPython	Python lib
ProvJS	Javascript lib
ProvExtract	Extract PROV from Web pages
ProvVis	Visualization
PROV-N Editor	Editor



Recipes and Patterns

- L. Moreau and P. Groth,
  'Provenance: An Introduction to PROV'.
  Morgan & Claypool Publishers, 2013.

- RDA WG Provenance patterns: http://patterns.promsns.org

Wrap-Up

- Prospective vs Retrospective Provenance
- Tracing Data Products back to their origin
- W3C PROV Ecosystem of Specifications
- Data Flow / Responsibility / Process views
- Querying PROV
- Generating Provenance
- Publishing Provenance Documents
- Tools & Guides

ENVIRONMENT AGENCY AUSTRIA umweltbundesamt

ENVRI

FAIR

doron.goldfarb@umweltbundesamt.at

keith.jeffery@keithgjefferyconsultants.co.uk