To be RE-USABLE
Basic aspects on data provenance

Barbara Magagna barbara.magagna@umweltbundesamt.at

Lecce, July 4 2019 *Ecosystem Research and Environmental Information Management*

umweltbundesamt
PERSPEKTIVEN FÜR UMWELT & GESELLSCHAFT

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

Outline

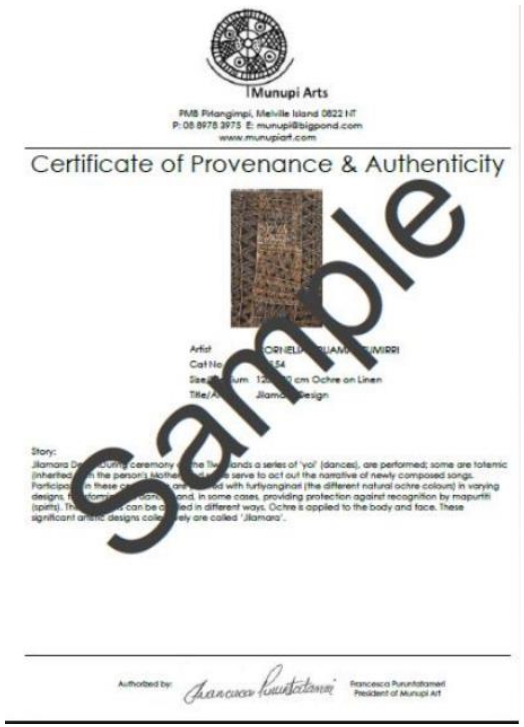
- Part 1: Provenance definition, uses and challenges
- Part 2: PROV standard and extensions
- Part 3: Provenance Template Service
- Part 4: Jupiter Notebook Provenance Implementation

To be RE-USABLE

Provenance definition, uses and challenges

why it is important

Provenance:
Any associations from everdays life?



FAIRTRADE INTERNATIONAL

BIO nach EG-Öko-Verordnung

D.O.C.G. DENOMINAZIONE DI ORIGINE CONTROLLATA E GARANTITA

TECH · TRANSPARENCY · TRUST · PROVENANCE.ORG

Provenance:
Any associations from everdays life?

- **Arts** - Documented evidence of provenance for an object helps to establish that it has not been altered and is not a forgery, a reproduction, stolen or looted art. Furthermore it supports to assign the work to a known artist and to prove ownership. Without provenance the artifact is auctioned for a lower price.
- **Food Provenance** – supply chain is crucial for the quality of food products
  -> FairTrade products, carbon neutral goods, sustainable productions
  Origin and quality of manufacturing process are essential for their branding -> competitive advantage

Origin of provenance

- French term ‘**provenir**’ meaning ‘**to come from**’, was originally used to keep track of the chain of ownership of cultural artifacts
- Goes back to the principle “Respect de fonds”, also known as the “Principle of provenance”, established in 1841 by a commission of historians in France.
  After the French Revolution the need emerged to merge public and private collections of property, legal and historical records into a single national archive.
  It means that the archives of one creator had to remain in the same original order and was considered to be a stable entity.

Definition of data provenance

Data provenance is information about entities, activities, and people involved in producing (influencing or delivering) a piece of data

This information is used for assessments about the data regarding:

- its quality
- its reliability
- Its trustworthiness

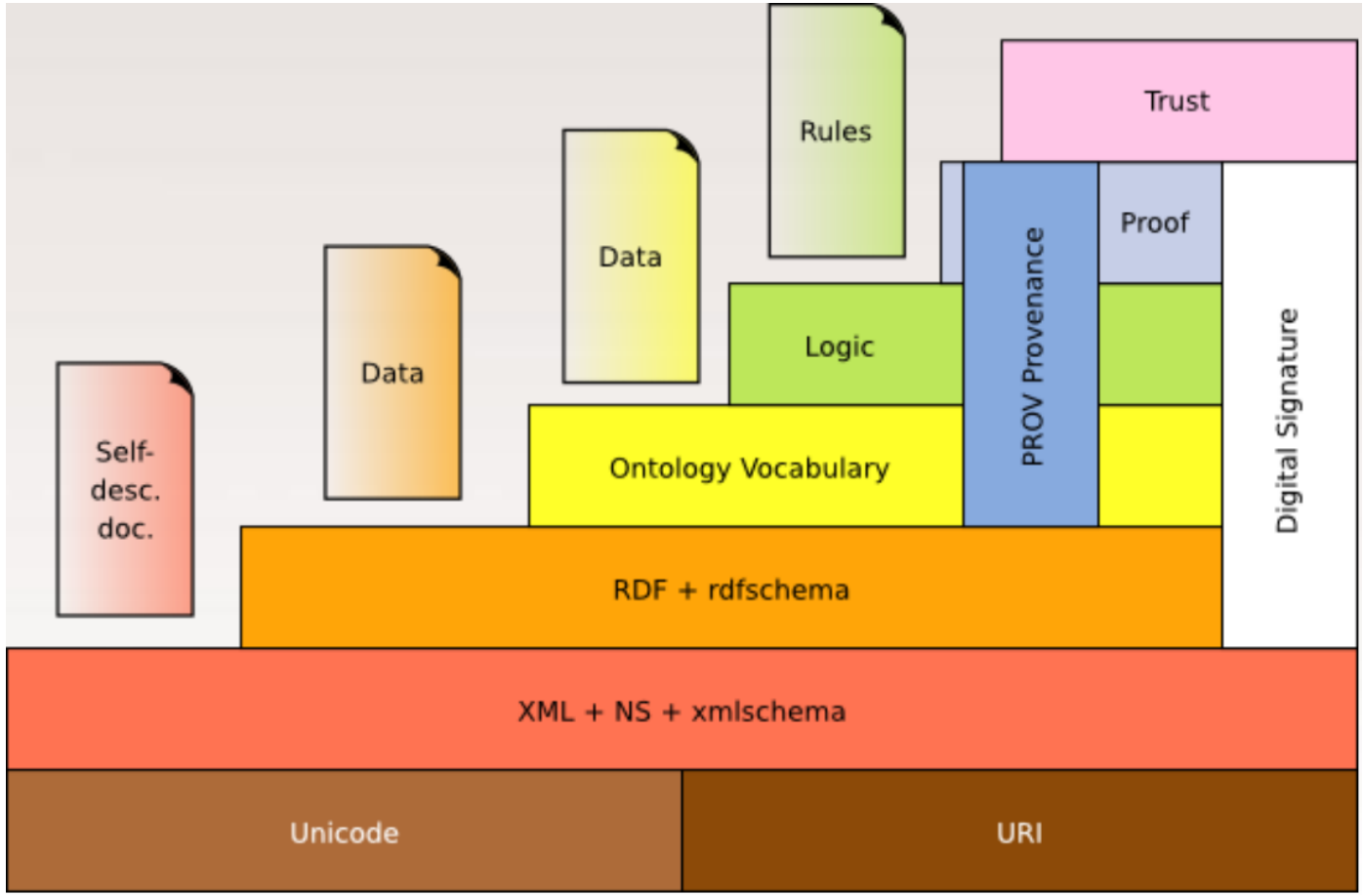
Provenance versus metadata

- Provenance is often conflated with metadata.
  They are related but not the same.
- **Provenance is a kind of metadata,**
  BUT not all metadata is provenance
  - The title or format of a book is metadata, but it is not part of its provenance.
  - The date of creation, the author, the publisher or the license of a book are part of its provenance.

Provenance versus logs

- Activity logs have been maintained in many ICT systems in the past and the present. A great variety in structure and detailed contents
  -> no standard form
- In general all the sciences have moved from handwritten to computer-based logs.
  Challenge -> reuse of log files for the reconstruction of data provenance

Provenance in the semantic web architecture



[Moreau, 2010]

PROV spans multiple layers:

- Ontology
- Logic: Inferencing rules
- Proof: tracks of logical inferences
- Trust: supported by assertions about origins

Provenance versus FAIR principles

- **R1. Meta(data) are richly described with a plurality of accurate and relevant attributes**
  - **R1.2. (Meta)data are associated with detailed provenance**

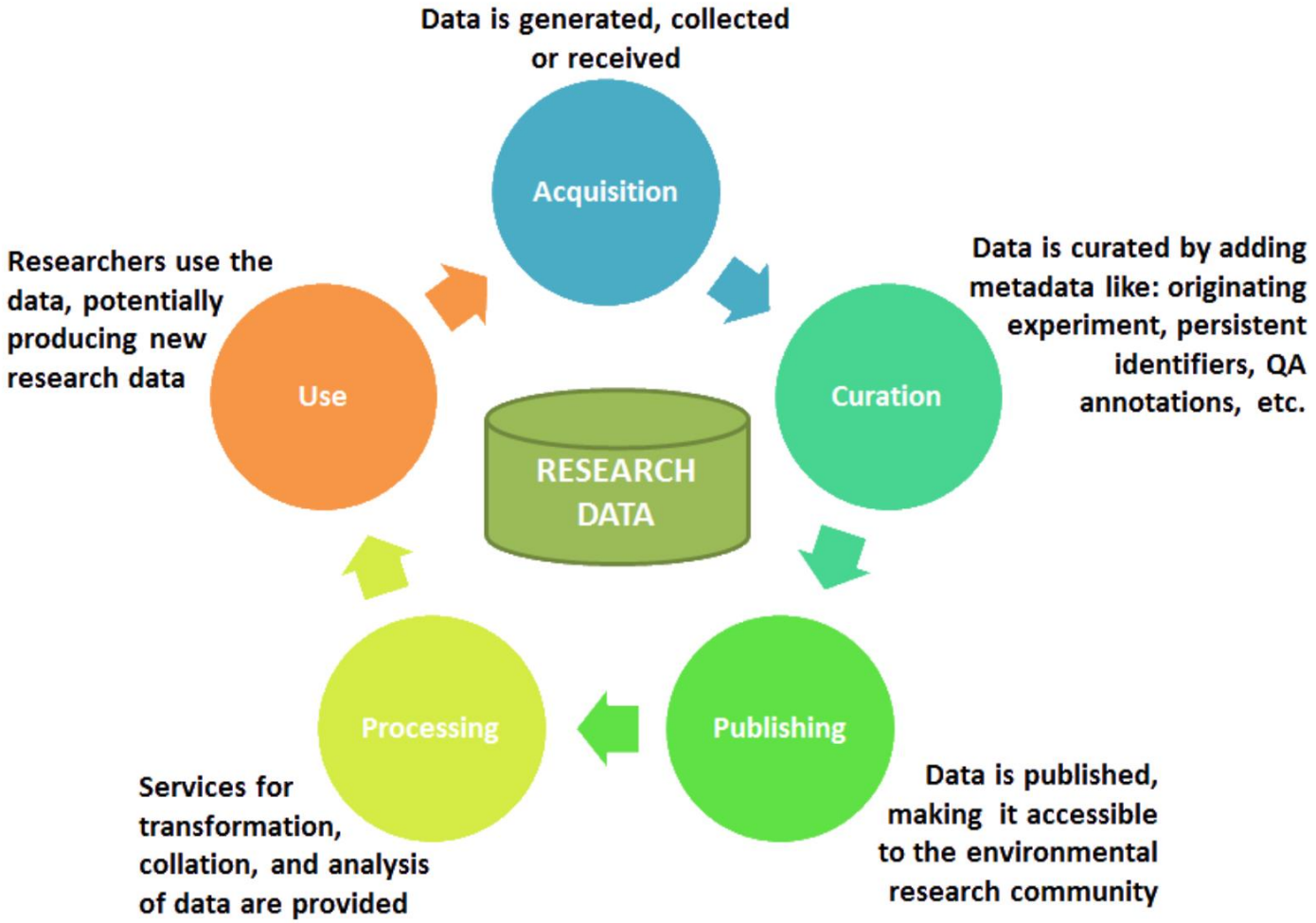
Example:

https://commons.wikimedia.org/wiki/File:Sampling_coral_microbiome_(27146437650).jpg
includes authorship details, and uses the Creative Commons Attribution Share Alike license, which indicates exactly how the data author wishes to be cited.

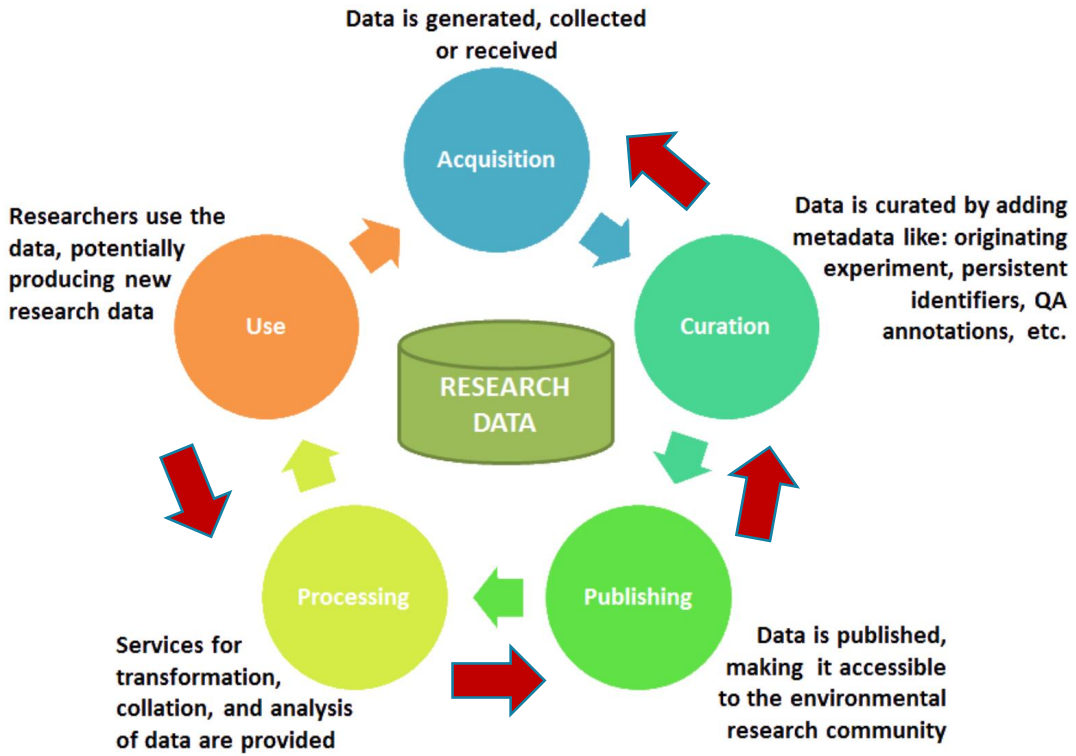
What provenance is about

- Who played a role when creating the data?
- Who owned the data?
- Who contributed to the data?
- How data was modified from its first revision?
- How other data affected the current data?
- Which tools where used to generate each version of the data?
- etc

Data life cycle (ENVRI Reference Model)



Data life cycle and provenance

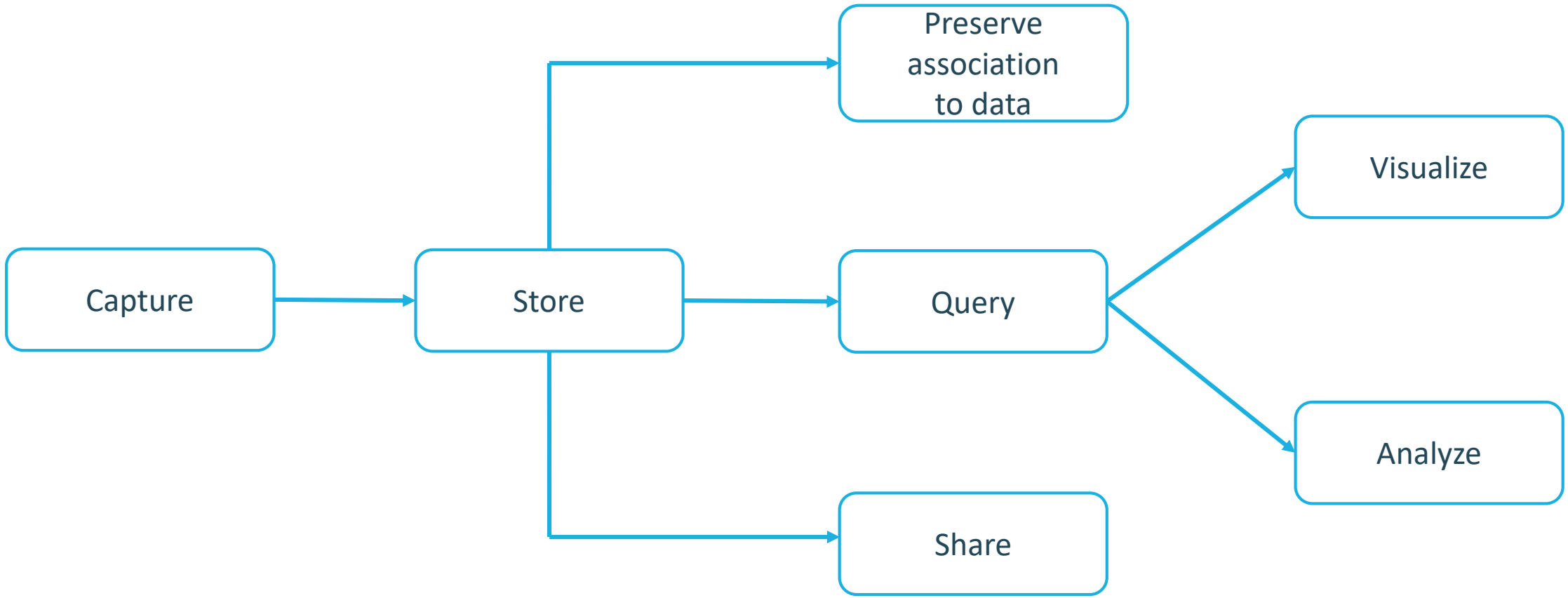


Provenance is the entire information that ran through the whole history of data process, including all data sources and all processes generating these data.

Requirements for data provenance

- **Identification of the data object**: it must have a persistent unique identifier (PID) to be referable. The digital objects may be organized in collections, subgroups and thus provenance may refer to aspects or portions of an object
- Using a **common provenance model** for the provenance record: PROV is a W3C standard
- **Provenance store**: Storing and querying PROV compliant data requires special data storage facilities that have to be configured, set up and maintained accordingly.
- **Creation of PROV compliant data**: Existing workflows and other processes must be adapted in order to be able to create PROV compliant data.

Life Cycle of Data Provenance



From: P. Missier, 'The lifecycle of provenance metadata and its associated challenges and opportunities,' in Building Trust in Information, Springer, 2016, pp. 127–137.

Types of provenance

- Retrospective provenance
  -> generation of data
- Prospective provenance
  -> abstract representation of procedure, workflow specification
- Process provenance
  -> evolution of the workflow description
- Provenance of data structure
  -> inputs and outputs of tasks in workflow

Uses of provenance

- Data reuse
- Estimation of data quality
- Attribution
- Data discovery
- Comparison
- Debugging
- Audit trail

For good data provenance

it is essential that

- the evolutionary contexts are maintained according to the actual data lifecycle to be a **source of trust**,
- the additional **information** about the digital object is **as comprehensive as possible** and captured by the people directly involved in the data production to be reliable,
- **all fragments of provenance information** tracked in the life span of the data production in each of the different distributed comp. infrastructures are based on the same model enabling a provenance summary **to be interoperable**

Challenges

- Efficiency of provenance collections
- Granularity
- Representation of domain semantics
- Interoperability
- Incomplete and uncertain provenance
- Trust
- Ease of use
- Visualisation

To be RE-USABLE

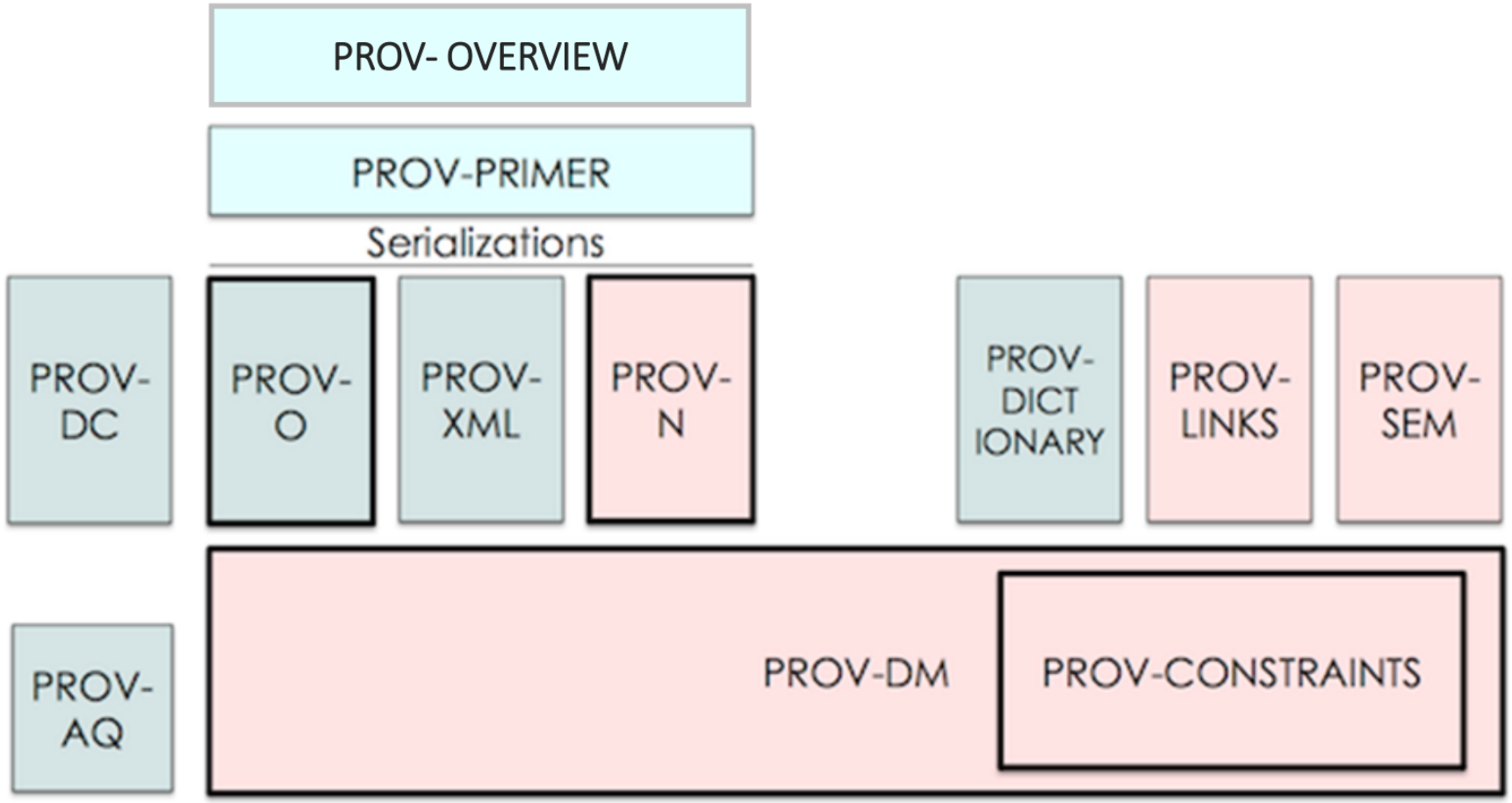
PROV standard and extensions

how it is represented

Provenance of PROV

- International Provenance and Annotation Workshop 2006 – Challenge
- 3rd Provenance Challenge (2010): OPM (the Open Provenance Model), first standard adopted by many workflow systems
- W3C Provenance Incubator Group -> W3C Provenance Working Group
- 2013: W3C PROV specifications, influenced by OPM and built upon the recommendations of the W3C Provenance Incubator Group

PROV documents



12 documents:

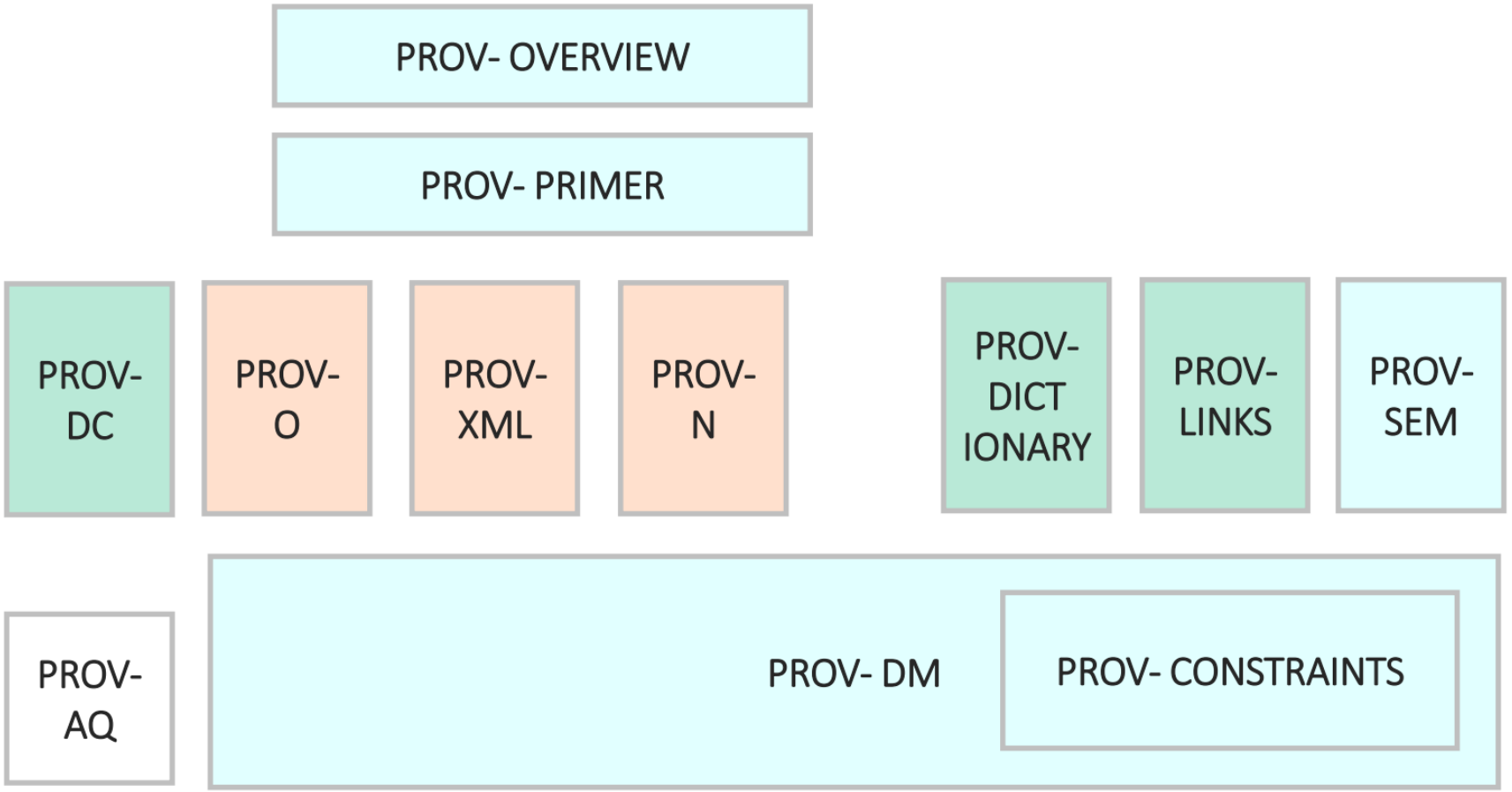
USERS

ADVANCED

DEVELOPERS

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

PROV documents



12 documents:

MODEL

SERIALIZATIONS

EXTENSIONS

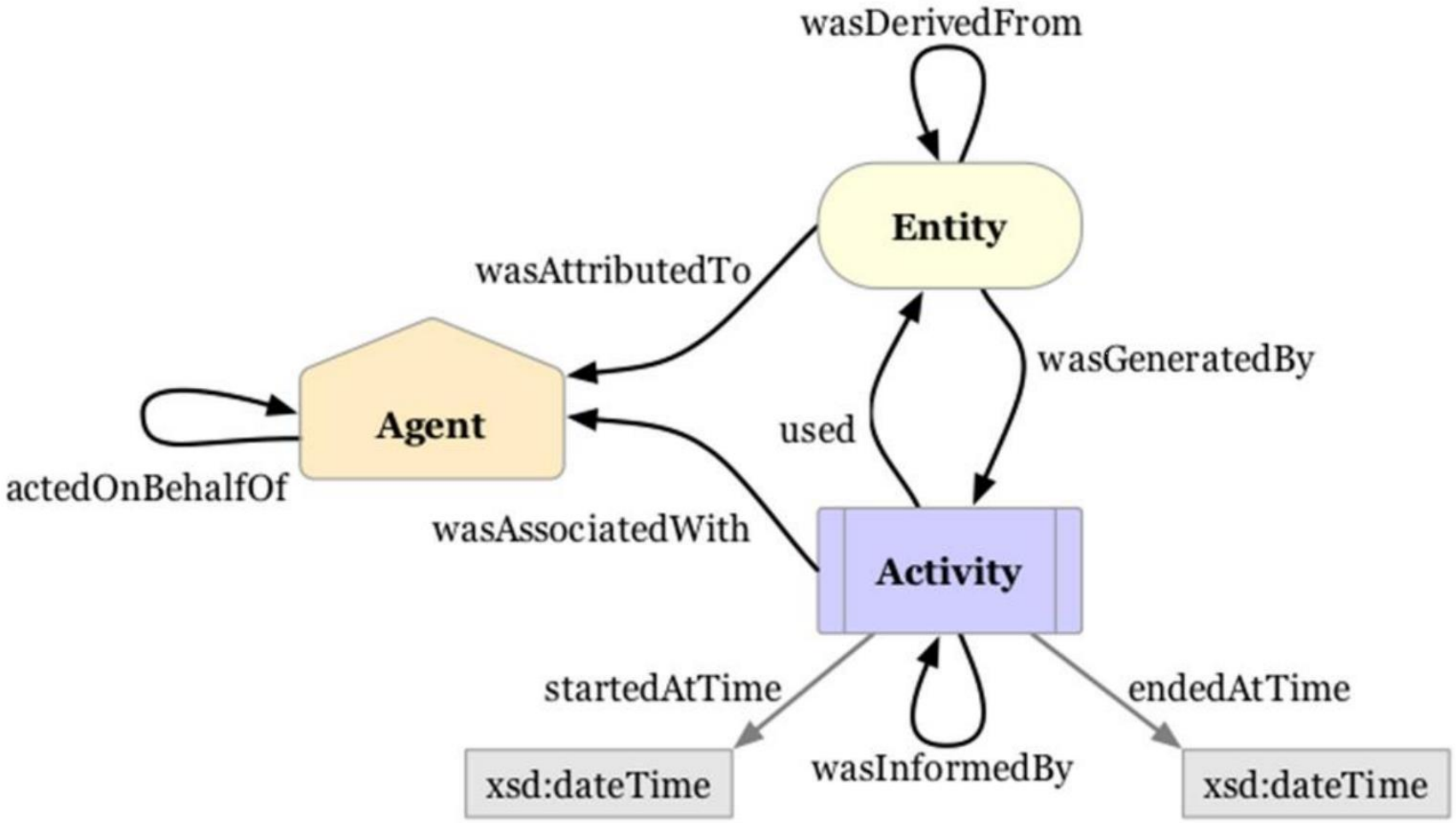
PROV documents

- 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/)

Categories of PROV terms

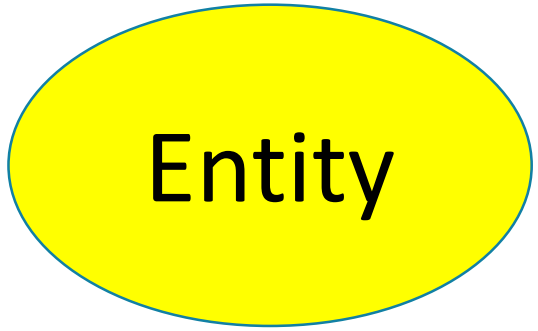
- **Starting Point classes and properties:** the basics.
- **Expanded classes and properties:** additional terms around the starting point terms for richer descriptions.
- **Qualified classes and properties:** for advanced provenance descriptions.

Starting Points



Starting Points: Entities

- Anything of interest for PROV documentation:
- A document
- A part of a document
- An idea
- A product
- A result
- Etc.



"An *entity* is a physical, digital, conceptual, or other kind of thing with some fixed aspects; entities may be real or imaginary".

Starting Points: Activities

Any processes that used or generated entities:

- Computing a result
- Making a request
- Writing a book
- Giving a presentation
- Creation of car
- Etc

Activity

"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".

Starting Points: Agents

Agents receive attribution for entities and are responsible for activities:

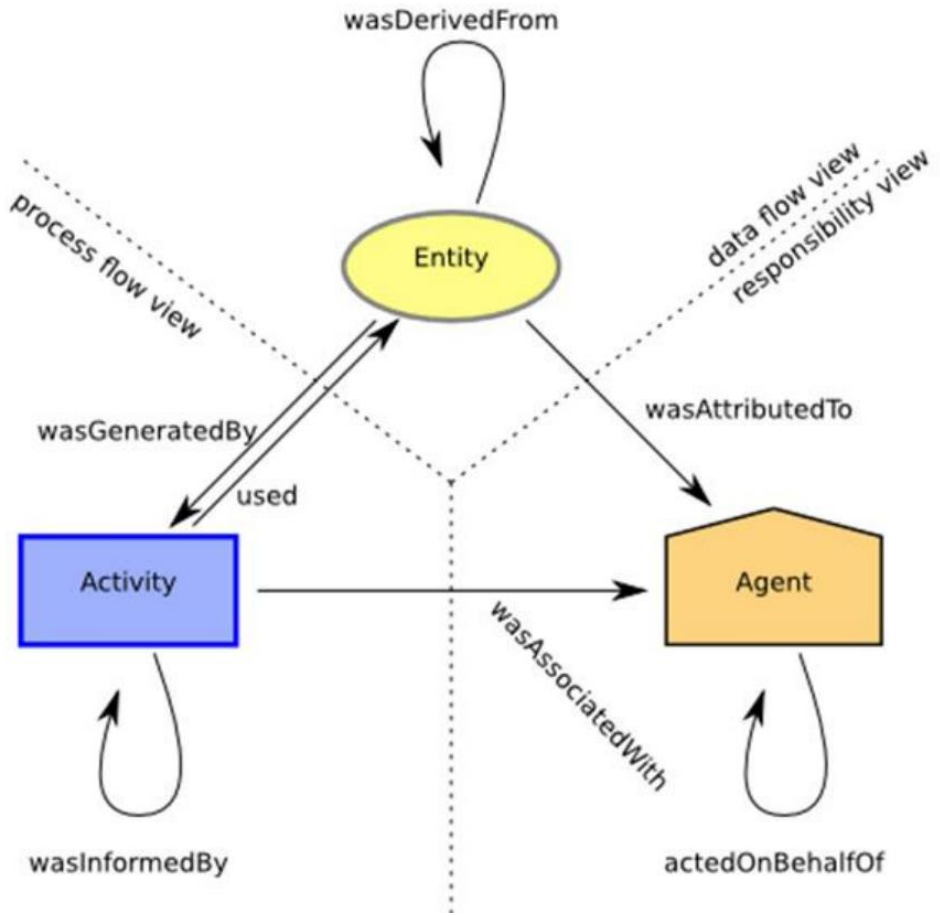
- Creator of a document
- Web service accepting request
- Tool or managing system
- An organization
- The student acting on behalf of the organization
- Etc.

"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".

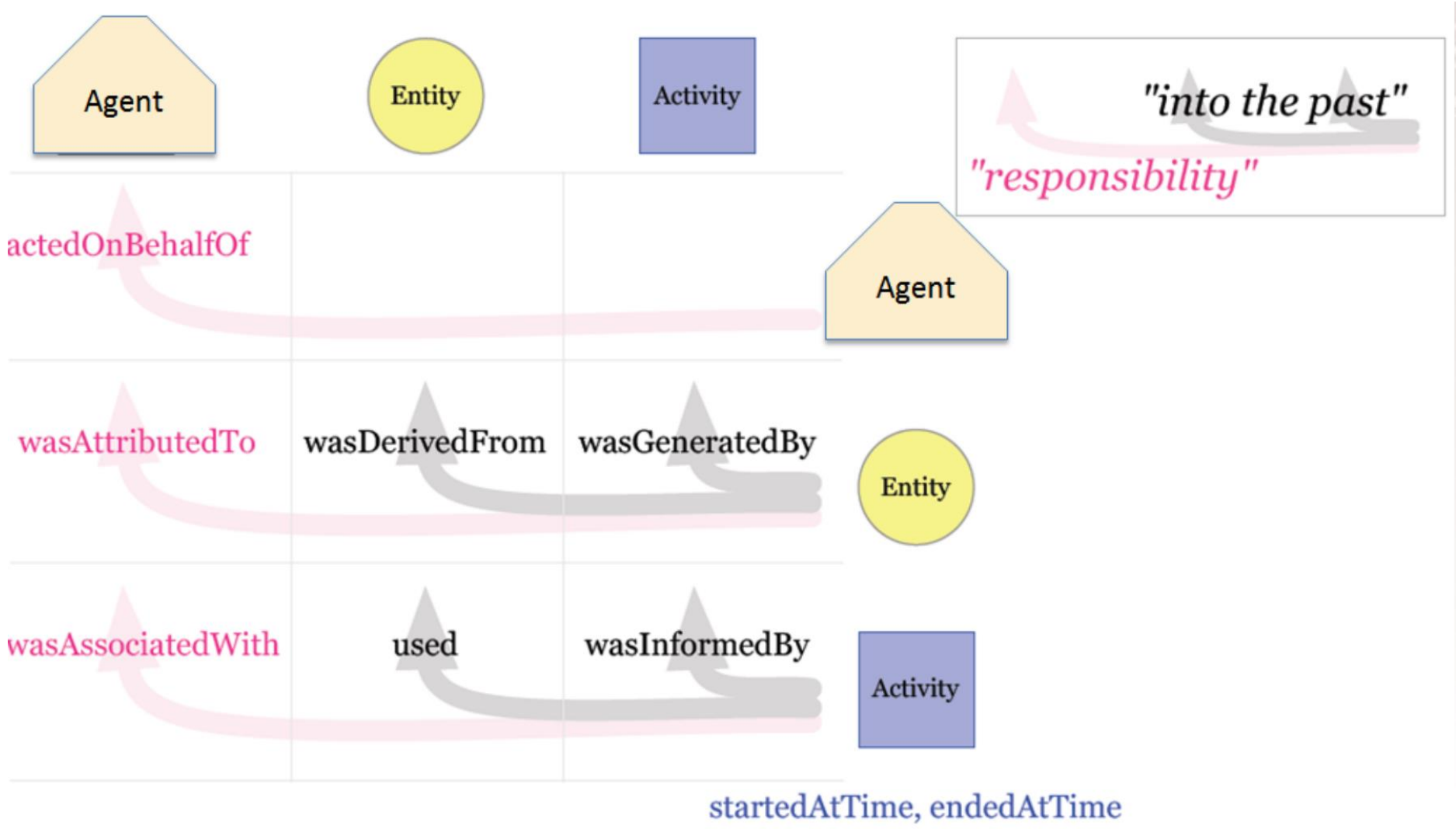
Starting Points: Properties

- **prov:startedAtTime** and **prov:endedAtTime** start and end points in time of Activities and Entities
- **prov:used** an activity used some Entity
- **prov:wasGeneratedBy** an Activity generated an Entity
- **prov:wasInformedBy** an Activity used an entity produced by another Activity, which allows Activity provenance chains
- **prov:wasDerivedFrom** an Entity was derived from another Entity, without mentioning the Activities involved. This expresses derivation, the provenance among entities, which is a transformation of an Entity into another.
- **prov:wasAttributedTo** an Entity is ascribed to an Agent
- **prov:wasAssociatedWith** an Activity lies in the responsibility of an Agent
- **prov:actedOnBehalfOf** an Agent can be responsible for the Activity of another Agent, who may have been less involved

Three views on provenance



Responsibility versus time



To be RE-USABLE

PROV Template Registry and Expansion Service

An implementation of Doron Goldfarb (EAA)

PROV Template

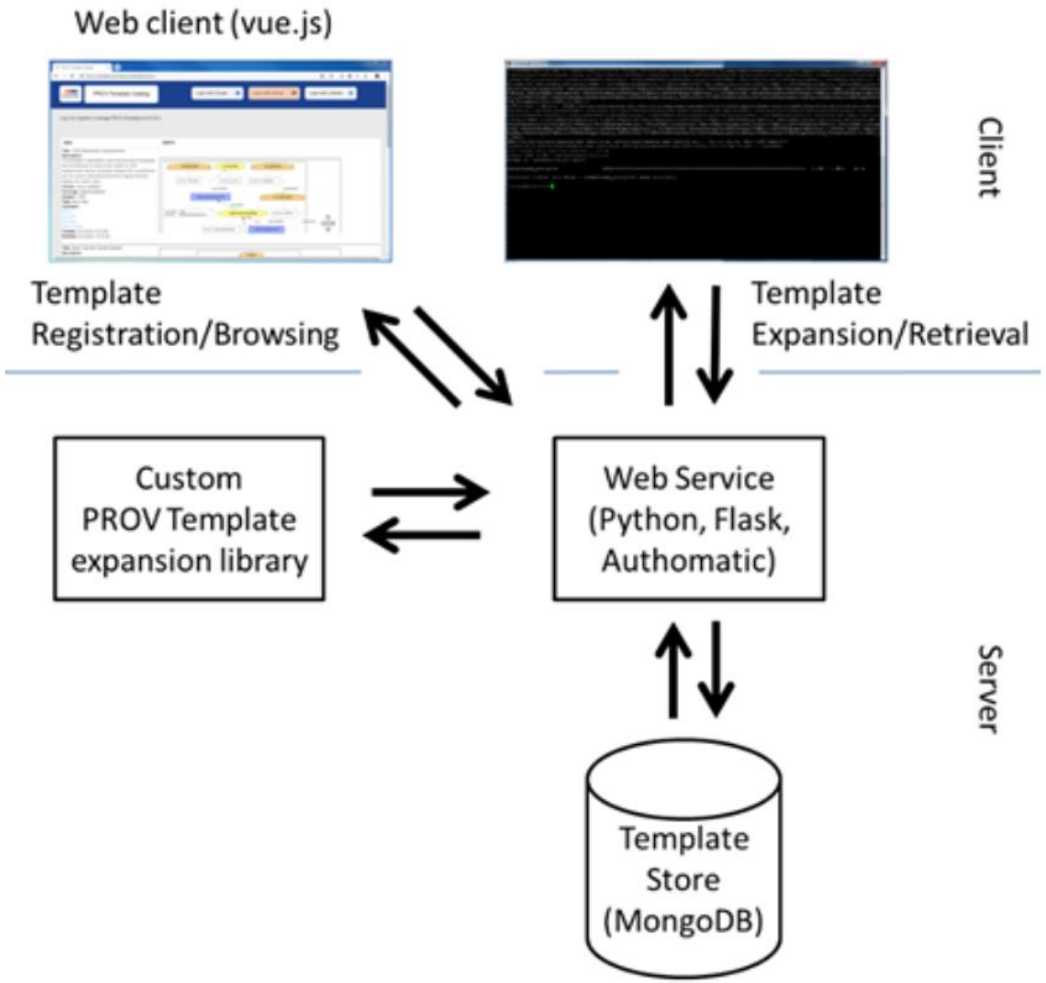
- **PROV-Template** is a proposed standard for converting existing process output such as log files into representations following the PROV Data Model (PROV-DM) specification for describing provenance of electronic resources in machine readable, structured form.
- The **ENVRI Template Service** is a public platform for describing, storing and sharing PROV-Templates across members of different RI, including a dedicated Web API for instantiating stored templates with individual data.
- It creates templates which predefine the structure of the intended provenance information using variables which are later instantiated (via bindings) with appropriate data extracted from existing process output.

Literature: L. Moreau, B. V. Batlajery, T. D. Huynh, D. Michaelides, and H. Packer, 'A Templating System to Generate Provenance', IEEE Transactions on Software Engineering, vol. 44, no. 2, pp. 103–121, Feb. 2018.

Objectives of the Template service

- Provide convenient means for communities to start experimenting with PROV-DM
- Enable the sharing and re-use of registered PROV-Templates in order to foster interoperable provenance traces across communities.
- Provide a Python based implementation of the PROV-Template expansion mechanism.

Service Description

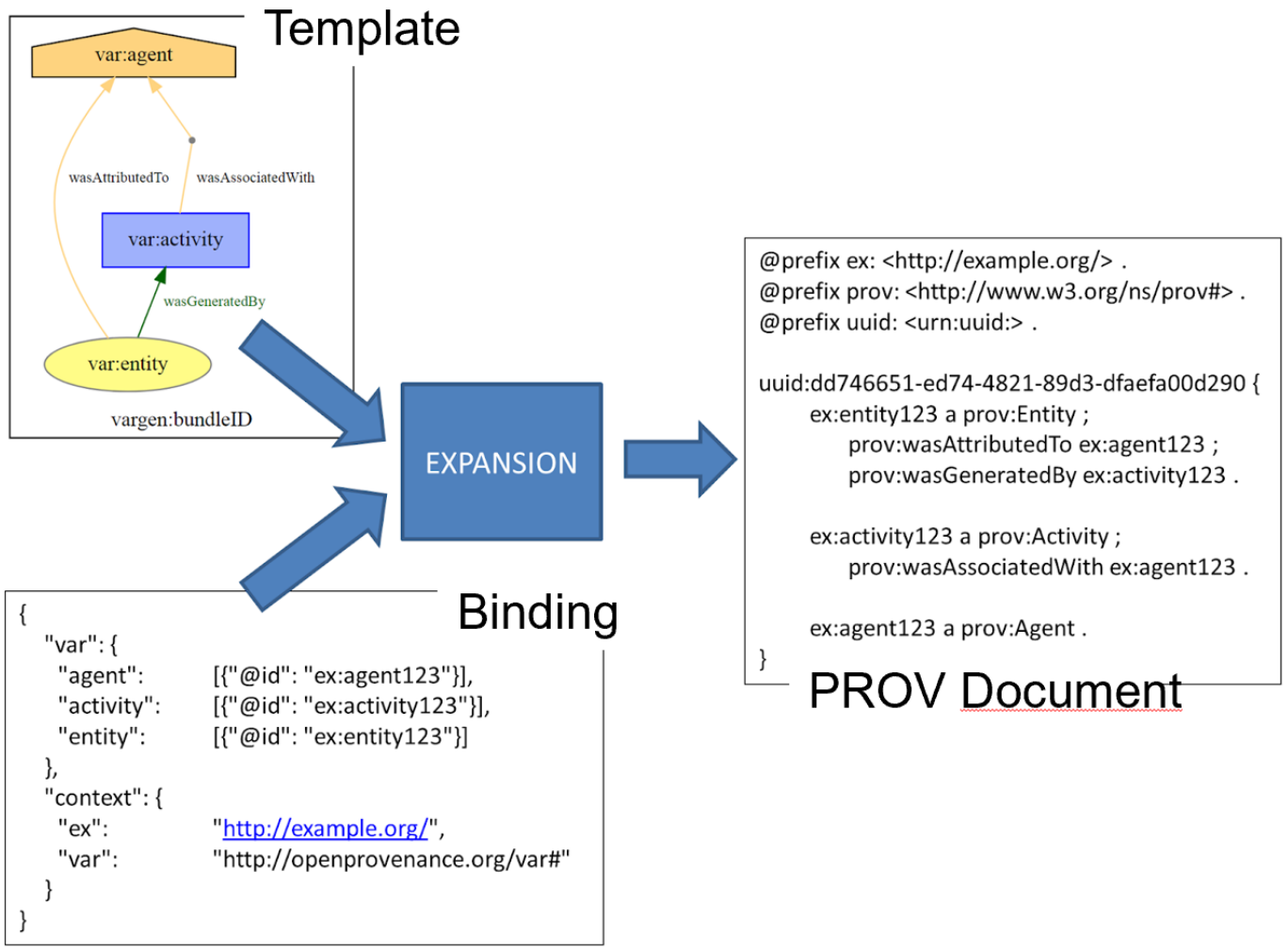


The service is accessible via
https://envri.eu/provenancetemplates

See YouTube video:
https://youtu.be/dS58qTfscSM

ENVRI wiki:
https://wiki.envri.eu/pages/viewpage.action?pageId=40861727

PROV Templates and Bindings



To be RE-USABLE

PROV Template Registry and Expansion Service

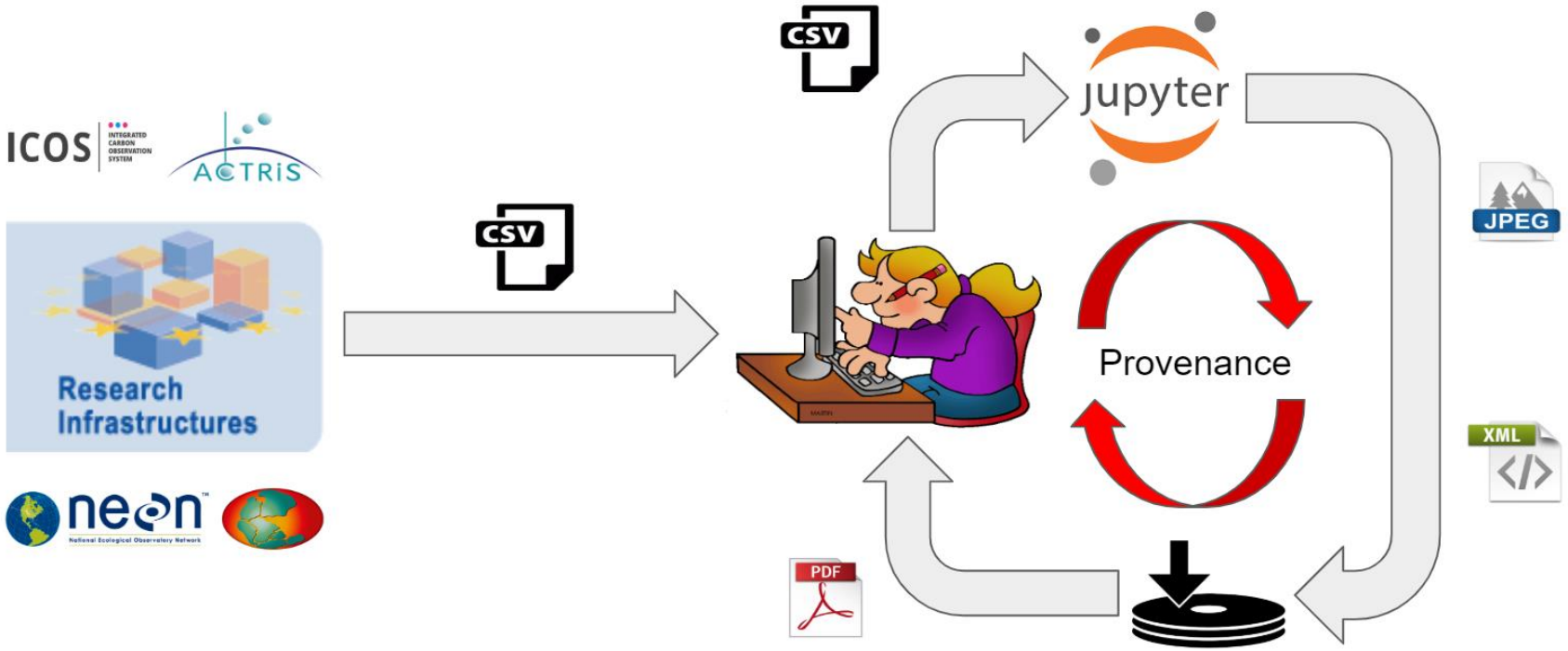
Test it yourself!

To be RE-USABLE

Provenance with Python and Semantic Technologies

An implemenation of Markus Stocker (TIB)

A Jupiter Notebook implementation



Use Case

- Aerosol Science
- Study of new particle formation events
- Atmospheric events
- Aerosol particles form and grow over time
- Events studied for effect on climate change
- As well as human respiratory health

Provenance tracking of use case

Primary data

Particle size distribution data

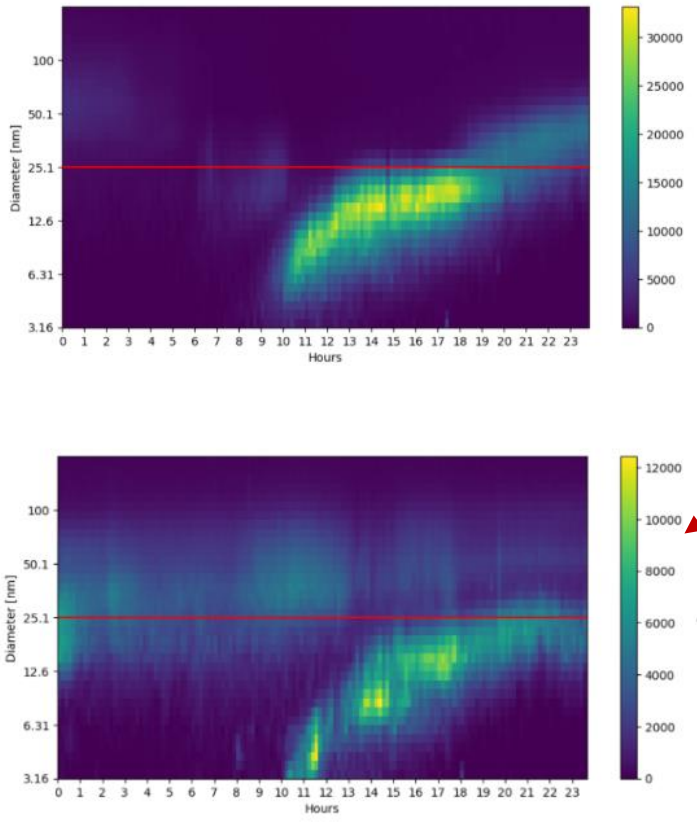
Secondary data

Tertiary data

Data about new particle formation events

Provenance

Create secondary data + provenance



Provenance

	day	beginning	end	duration
0	2007-04-15	10:30	12:30	02:00:00
1	2009-03-22	12:30	15:00	02:30:00
2	2008-02-19	12:00	13:00	01:00:00
3	2007-05-05	12:00	16:00	04:00:00
4	2007-05-18	12:30	13:30	01:00:00
5	2007-10-19	14:30	17:30	03:00:00
6	2009-03-19	11:30	15:00	03:30:00

Instructions:

- Go to https://notebooks.egi.eu
- Sign in with EGI Check-in
- In Jupyter Lab, open a Terminal
- Change to *work* directory
  cd work
- Clone the repository
  git clone https://github.com/markusstocker/lecce-summer-school.git
- In the menu select work/lecce-summer-school/provenance-lab.ipynb
- Follow the instructions in the notebook

To be RE-USABLE

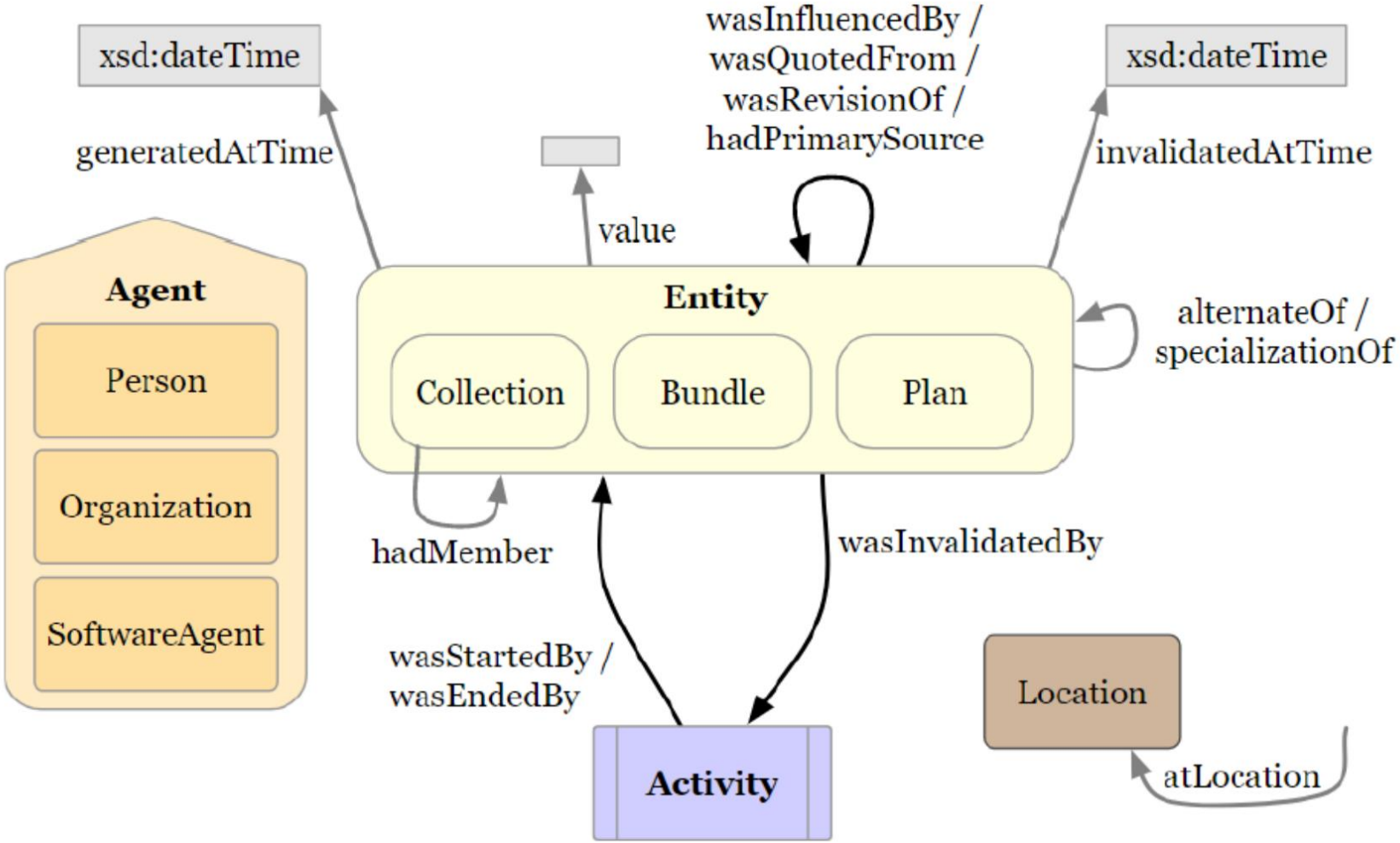
Provenance with Python and Semantic Technologies

Try it yourself!

To be RE-USABLE

More insights

Expanded terms



Expanded classes

Subclasses of Agent, which may overlap und thus are not disjoint:

- prov:Person
- prov:Organization
- prov:SoftwareAgent

Subclass of Entity:

- prov:Bundle is a named set of provenance descriptions to allow provenance of provenance
- prov:Collection: entities can be members of a collection
- prov:Plan can be a software program, a cooking recipe or anything else that describes how an activity was carried out.

Expanded properties

Datatype properties: to allow time validity descriptions for activities

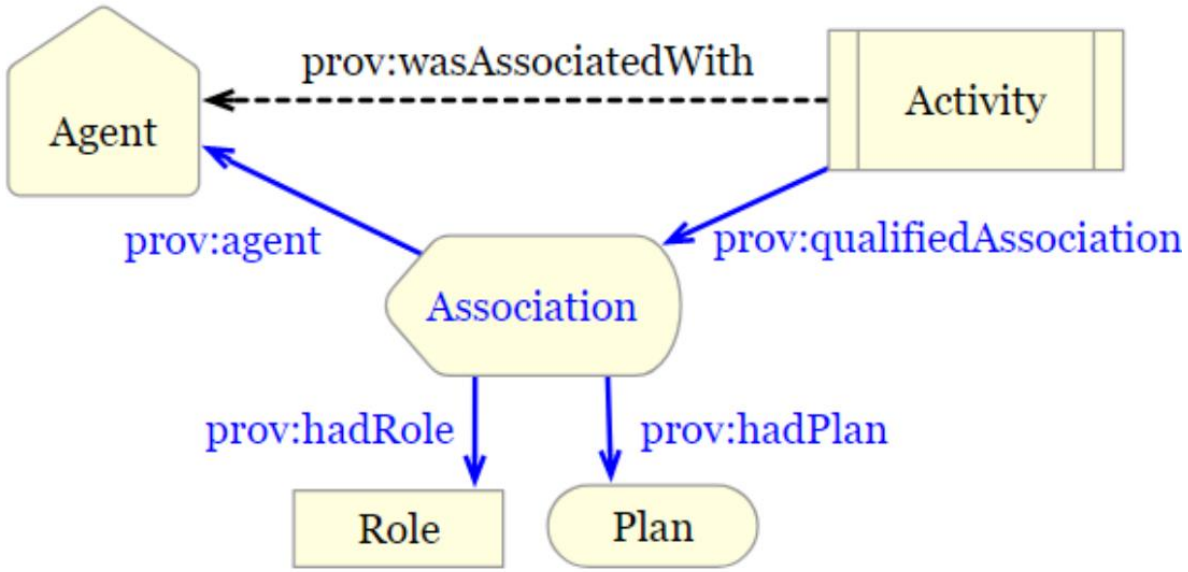
- prov:generatedAtTime
- prov:invalidatedAtTime

Subproperties of prov:wasDerivedFrom:

- prov:wasQuotedFrom cites a source such as a journal or website
- prov:wasRevisionOf refers to an older version of an Entity

Qualified terms

Qualified terms are used to provide additional attributes of the binary relations (object properties).



To model prov:Activity prov:wasAssociatedWith a particular prov:Entity, one creates an instance of prov:Association that indicates the influencing entity with the prov:entity property.
Meanwhile, the influenced prov:Activity indicates the prov:Usage with the property prov:qualifiedAssociation.
Now the plan of actions and steps that the Agent used to achieve its goals is provided by adding the object property prov:hadPlan to the Association qualified class and an instance of prov:Plan

Qualified influence classes

Unqualified influence properties (Starting Point relations) can be turned into qualified influence classes to be further described:

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

Qualified classes

- **Generation** is the completion of production of a new entity by an activity. This entity did not exist before generation and becomes available for usage after this generation.
- A **derivation** is a transformation of an entity into another, an update of an entity resulting in a new one, or the construction of a new entity based on a pre-existing entity.
- **Attribution** is the ascribing of an entity to an agent.
- **Usage** is the beginning of utilizing an entity by an activity. Before usage, the activity had not begun to utilize this entity and could not have been affected by the entity.

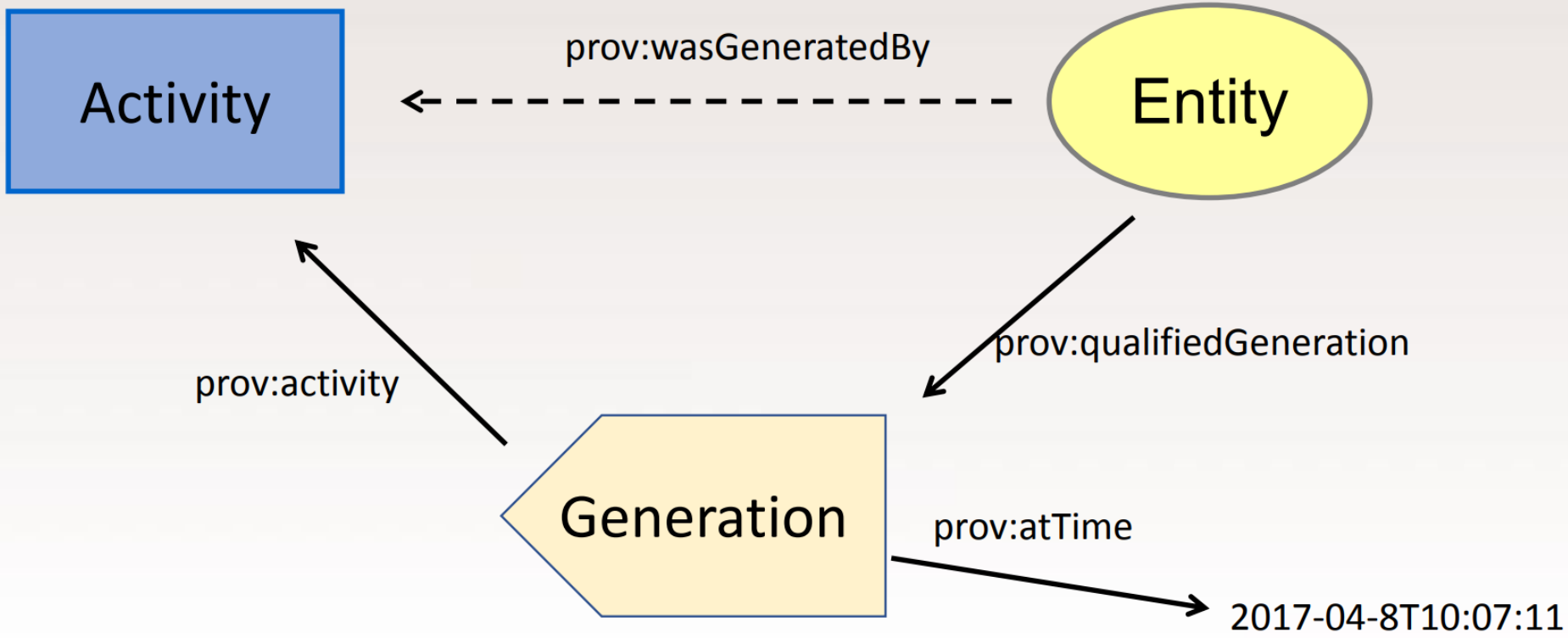
Qualified classes

- **Communication** is the exchange of some unspecified entity by two activities, one activity using some entity generated by the other.
- **Delegation** is the assignment of authority and responsibility to an agent to carry out a specific activity as a delegate or representative.
- A **derivation** is a transformation of an entity into another, an update of an entity resulting in a new one, or the construction of a new entity based on a pre-existing entity.

Qualified classes

- **Invalidation** is the start of the destruction, cessation, or expiry of an existing entity by an activity. The entity is no longer available for use (or further invalidation) after invalidation.
- **Start** is when an activity is deemed to have been started by an entity, known as trigger. The activity did not exist before its start.
- **End** is when an activity is deemed to have been ended by an entity, known as trigger. The activity no longer exists after its end.
- An **association** is an assignment of responsibility to an agent for an activity, indicating that the agent had a role in the activity. It further allows for a plan to be specified.

Generation


Activity
prov:wasGeneratedBy
Entity
prov:activity
prov:qualifiedGeneration
Generation
prov:atTime
2017-04-8T10:07:11

Provenance Tools:

- https://openprovenance.org/

Tools & Libraries

- ProvToolbox - a Java toolbox for handling PROV
- Prov Python - a Python implementation of the PROV data model
- ProvJS - a JavaScript implementation of the PROV data model
- ProvExtract - for dealing with PROV embedded in web pages
- ProvVis - experimental visualizations of PROV
- PROV-N Editor - a text editor with PROV-N syntax highlighted

ProvStore

A provenance repository that allows storing, browsing, and managing provenance documents via a Web interface or a REST API.

Translator

Translates between different representations of PROV. Supports PROV-N, PROV-XML, PROV-O and PROV-JSON.

Validator

A RESTful web service that validates PROV descriptions against the PROV Constraints specification. Supports uploading PROV by URL, file upload or inline statements.

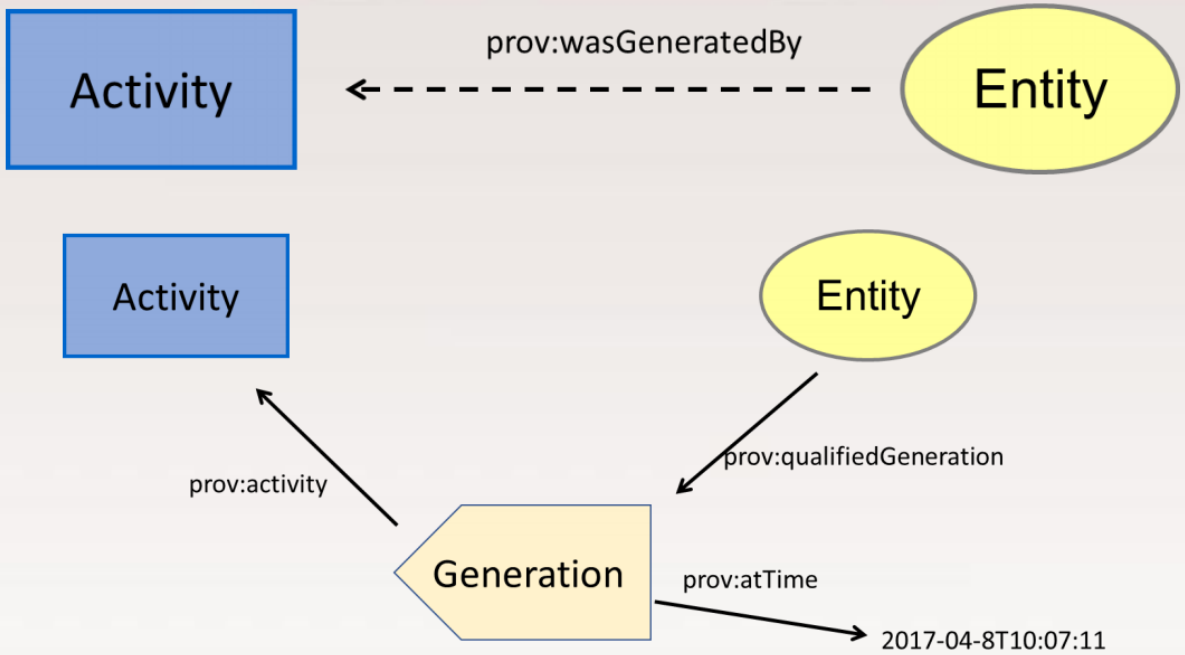
Recipies and patterns

- See L. Moreau and P. Groth, ‘Provenance: An Introduction to PROV’. Morgan & Claypool Publishers, 2013.
- RDA WG Provenance patterns: http://patterns.promsns.org
- A. Spinuso, ‘Active Provenance for Data-Intensive Research’, PhD thesis, School of Informatics, University of Edinburgh, 2018.

Initiatives dealing with provenance

- DataONE ProvWG (https://www.dataone.org/)
- ESIP (Earth Science Information Partners) http://www.esipfed.org
- ENVRIplus www.envriplus.eu
- RDA https://www.rd-alliance.org:
  - Research Data Provenance IG
  - Provenance Pattern WG (PPWG)
  - Publishing Data Workflows WG (already completed)
  - Dynamic Data Citation WG (already completed)
  - PID Kernel Information Types WG (already completed)
  - Reproducibility IG (active)
  - PID IG (active) □ Archives and records professionals for research data IG (active)
  - Data discovery paradigms IG (active)
  - Preservation e-Infrastructure IG (active)
  - From Observational Data to Information IG (active) □ Metadata IG (active)
  - Data in Context IG (active)

Generation expressed in PROV-O Turtle



Unqualified

```
:e a prov:Entity.
:a a prov:Activity.
:e prov:wasGeneratedBy :a.
```

Qualified

```
:g a prov:Generation.
:e prov:qualifiedGeneration :g.
:g prov:activity :a.
:g prov:atTime "2012-04-
01T12:01:01"^^xsd:dateTime.
```

Class	Parent	
prov:Generation	prov:ActivityInfluence, prov:InstatenousEvent	
Property	**Domain**	**Range**
prov:wasGeneratedBy	prov:Entity	prov:Activity
prov:qualifiedGeneration	prov:Entity	prov:Generation
prov:activity	prov:ActivityInfluence	prov:Activity
prov:atTime	prov:InstantaneousEvent	xsd:dateTime

To be RE-USABLE

Thank you!

Barbara.Magagna@umweltbundesamt.at