Defining and Integrating a CDIF

PROV Extension

September 9, 2025 – Arofan Gregory

# Overview

This document attempts to provide a framework for taking the candidate classes for describing the provenance of research data, as outlined in the working paper [*"CDIF Context and Provenance Workgroup Introduction"*](https://docs.google.com/document/d/1Vbo8TW2taprOrj5IWjHJM0nhSxfceueRK2yixF4m_bM/edit?tab=t.0)and providing more complete written definitions of them, along with examples. Further, it examines the existing PROV Data Model and identifies how the proposed classes fit best into the existing PROV standard. As part of this work, any additional properties which need to be added will be specified. It is hoped that this document can provide the basis for a candidate set of extensions to be written in RDF for possible incorporation into the CDIF Guidelines.

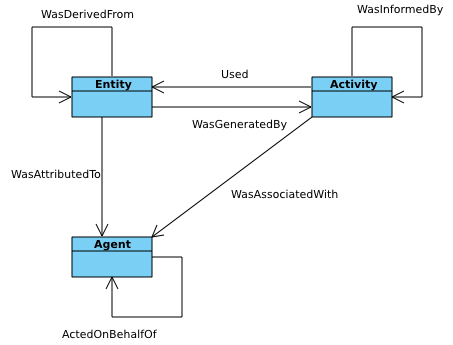
Like its predecessors, this is intended to be a working document. Its content will in part be the repetition of material from PROV which is relevant, to assist in the integration of the proposed classes with the existing specification. Due to variations in terminology across domains, it is important to refer to the formal definitions provided in PROV for the different constructs in the data model, as it is very easy to unintentionally re-create something which already exists. Such duplication is to be avoided when aiming to create an integrated set of standards (!)

# Reference Material from the PROV Data Model

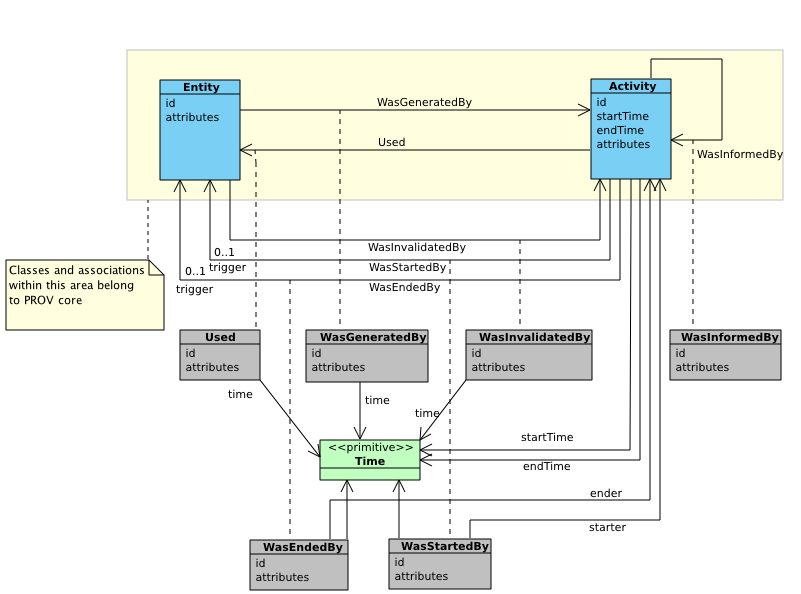
Below are some useful diagrams from the PROV Overview and the PROV Data Model, along with a chart which lists out the types and relations in all of its components. The intention here is to provide an easy reference and a set of links (in the chart) which will allow for questions to be answered as they arise. The terminology in PROV may not align with the terminology of any given domain, so attention must be paid to the definitions as provided in the PROV specifications.

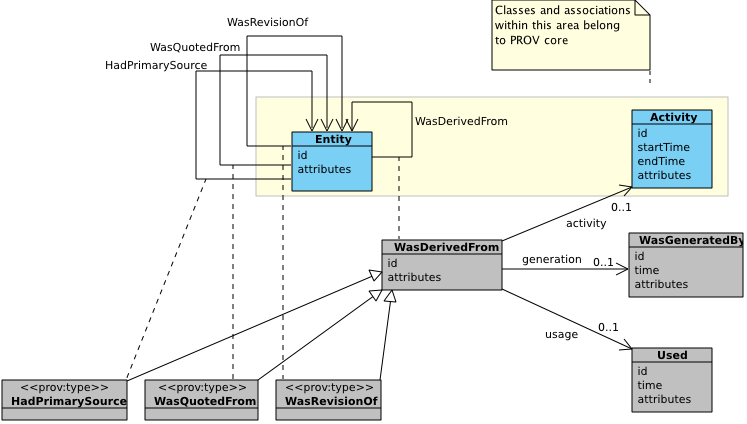
To assist with this, in the following section, relevant definitions from PROV have been cited for the basic constructs used to organize the extensions.

From the PROV Overview:



From the PROV Data Model:





|  |  |  |
| --- | --- | --- |
| Table 6 [◊:](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#prov-dm-types-and-relations) PROV-DM Types and Relations | | |
| Type or Relation Name | Representation in the PROV-N notation | Component |
|  |  |  |
| [**Entity**](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#concept-entity) | [**entity(id, [ attr1=val1, ...])**](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-entity) | [Component 1: Entities/Activities](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#component1) |
| [**Activity**](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#concept-activity) | [**activity(id, st, et, [ attr1=val1, ...])**](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-activity) |
| [**Generation**](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#concept-generation) | [**wasGeneratedBy(**id;**e,a**,t,attrs**)**](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-wasgeneratedby) |
| [**Usage**](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#concept-usage) | [**used(**id;**a,e**,t,attrs**)**](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-used) |
| [**Communication**](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#concept-communication) | [**wasInformedBy(**id;**a2,a1**,attrs**)**](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-wasinformedby) |
| [Start](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#concept-start) | [wasStartedBy(id;a2,e,a1,t,attrs)](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-wasstartedby) |
| [End](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#concept-end) | [wasEndedBy(id;a2,e,a1,t,attrs)](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-wasendedby) |
| [Invalidation](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#concept-invalidation) | [wasInvalidatedBy(id;e,a,t,attrs)](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-wasinvalidatedby) |
|  |  |  |
| [**Derivation**](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#concept-derivation) | [**wasDerivedFrom(**id; **e2, e1**, a, g2, u1, attrs**)**](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-wasderivedfrom) | [Component 2: Derivations](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#component2) |
| [Revision](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-revision) | [... prov:type='prov:Revision' ...](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-prov:revision) |
| [Quotation](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-quotation) | [... prov:type='prov:Quotation' ...](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-prov:quotation) |
| [Primary Source](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-primary-source) | [... prov:type='prov:PrimarySource' ...](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-prov:primary-source) |
|  |  |  |
| [**Agent**](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#concept-agent) | [**agent(id, [ attr1=val1, ...])**](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-agent) | [Component 3: Agents, Responsibility, Influence](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#component3) |
| [**Attribution**](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#concept-attribution) | [**wasAttributedTo(**id;**e,ag**,attr**)**](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-wasattributedto) |
| [**Association**](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#concept-association) | [**wasAssociatedWith(**id;**a,ag**,pl,attrs**)**](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-wasassociatedwith) |
| [**Delegation**](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#concept-delegation) | [**actedOnBehalfOf(**id;**ag2,ag1**,a,attrs**)**](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-actedonbehalfof) |
| [Plan](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#concept-plan) | [... prov:type='prov:Plan' ...](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-prov:plan) |
| [Person](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#concept-person) | [... prov:type='prov:Person' ...](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-prov:person) |
| [Organization](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#concept-organization) | [... prov:type='prov:Organization' ...](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-prov:organization) |
| [SoftwareAgent](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-prov:softwareAgent) | [... prov:type='prov:SoftwareAgent' ...](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#concept-software-agent) |
| [Influence](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#concept-influence) | [wasInfluencedBy(id;e2,e1,attrs)](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-wasinfluencedby) |
|  |  |  |
| [Bundle constructor](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#concept-bundle) | [bundle id description\_1 ... description\_n endBundle](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-bundle-constructor) | [Component 4: Bundles](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#component4) |
| [Bundle type](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-bundle-type) | [... prov:type='prov:Bundle' ...](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-prov:bundle) |
|  |  |  |
| [Alternate](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#concept-alternate) | [alternateOf(alt1, alt2)](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-alternateof) | [Component 5: Alternate](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#component5) |
| [Specialization](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#concept-specialization) | [specializationOf(infra, supra)](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-specializationof) |
|  |  |  |
| [Collection](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#concept-collection) | [... prov:type='prov:Collection' ...](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-prov:collection) | [Component 6: Collections](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#component6) |
| [EmptyCollection](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#concept-empty-collection) | [... prov:type='prov:EmptyCollection' ...](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-prov:emptyCollection) |
| [Membership](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#concept-membership) | [hadMember(c,e)](https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-hadmember) |

# Extensions to the PROV Model to Support CDIF

This section is organized according to the three entry classes in the PROV Ontology: Activity, Entity, and Agent. In each area, the definition from the PROV Data Model is given to clarify what is intended in the W3C spec.

## Activities

### From the PROV-O Data Model

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.

An *activity*, written activity (id, st, et, [attr1=val1, ...]) in PROV-N, has:

* *id*: an identifier for an activity;
* *startTime*: an OPTIONAL time (st) for the start of the activity;
* *endTime*: an OPTIONAL time (et) for the end of the activity;
* *attributes*: an OPTIONAL set of attribute-value pairs ((attr1, val1), ...) representing additional information about this activity.

### Proposed CDIF Extensions

#### **Exposure event / pre-event**

#### Definition:Establishment of the conditions that will subsequently be monitored / measured. E.g., dosing in a toxicology study or clinical trial, a meteor hitting earth that will then be sampled.

*Example:*

*Explanatory Notes:*

#### **Event**

#### Definition: Collection or definition of representative samples, material or otherwise for use as subjects in Observation/Measurement Events.

*Example:*

*Explanatory Notes:*

**Observation/Measurement Event:**

*Definition:* Direct measurement or observation of the subject which produces data. ALTERNATE: This is an event which produces data regarding a phenomenon or object of study. This includes samples. Data is a necessary output of such an event.

*Example:*

*Explanatory Notes:*

**Generation Event:**

*Definition:* An event where records or data are used to produce new data about a subject.

*Example:*

*Explanatory Notes:*

**Data Processing Event:**

*Definition:* Event where data is edited or transformed.

*Example:*

*Explanatory Notes:*

**Data Validation Event:**

*Definition:* Event where data is checked against some criteria, with some form of metrics or quality report as a result.

*Example:*

*Explanatory Notes:*

**Analysis Event**:

*Definition:* Event where data is interpreted to answer a research question.

*Example:*

*Explanatory Notes:*

**[QUERY: The following are part of the data life cycle. Are they important events for lineage/provenance understanding?]**

**Dissemination Event**:

*Definition:* Event where data is delivered to an end user.

*Example:*

*Explanatory Notes:*

**Documentation Event:**

*Definition:* Event where metadata is enhanced in order to support other functions.

*Example:*

*Explanatory Notes:*

**Data Access Event:**

*Definition:* Event where access is provided to data for a specific user.

*Example:*

*Explanatory Notes:*

**Exchange/Reporting Event**:

*Definition:* Event where data is transmitted to a counterparty in accord with a prior agreement (not to end users). [for provenance, significant if it is part of a multi-agent provenance chain]

*Example:*

*Explanatory Notes:*

**Registration/Cataloguing Event:**

*Definition:* An event where an entity is registered or catalogued.

*Example:*

*Explanatory Notes:*

**Archival Event:**

*Definition:* An event where an entity is stored for future use.

*Example:*

*Explanatory Notes:*

**Preservation Event:**

*Definition:* An event where an entity is stored in order to guarantee its continued existence into the future. [significant if an archived/preserved resource is ‘resurrected’ as part of a prov chain, and the archive/preservation process impacts some aspect of data quality/reliability/reusability]

*Example:*

*Explanatory Notes:*

**Comments:**

Do we use intermediate subclasses? As an example, do we have a “Data Origination Event” for which Observation/Measurement and Generation Events are sub-classes? What does this buy us?

“Actuation Events” are discussed as a super-class of some of the events listed here (see SOSA model) – the value of including them in the model should be considered. This may be more GBIF-specific than we want.

## Entities

From the PROV-O Data Model:

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

An *entity*, written entity (id, [attr1=val1, ...]) in PROV-N, has:

* *id*: an identifier for an entity;
* *attributes*: an OPTIONAL set of attribute-value pairs ((attr1, val1), ...) representing additional information about the fixed aspects of this entity.

### Proposed CDIF Extensions

Study or project level entities:

**Experiment/Study Description**

*Definition:* The overall process to be used to validate the Hypothesis, involving various types of resources and events as indicated by the Methods used.

*Example:*

*Explanatory Notes:*

**Hypothesis**

*Definition:* The question or theory which is being pursued in the observational or analysis activity

*Example:*

*Explanatory Notes:*

Directly related to data acquisition and lineage:

**Sample**

*Definition:* A physical or representative thing (of the intended subject population) which can be involved in an observation event (or others, such as a Preservation Event).

*Example:*

*Explanatory Notes:*

**Subject of Observation**

*Definition:* An Observed Event, Actor, or Sample being observed and measured in an Observation/Measurement or Generation Event. This is the “proximate unit of study/object of interest”.)

*Example:*

*Explanatory Notes:*

**Observed Event**

*Definition:* An event which is the subject of an Observation/Measurement or Generation Event.

*Example:*

*Explanatory Notes:*

**Method**

*Definition:* A technique for performing some operation, such as data collection, generation, processing, etc.

*Example:*

*Explanatory Notes:*

**Configuration**

*Definition:* The settings used on an instrument which collects or generates data. ALTERNATE: Settings employed in the observation, measurement, analysis, processing, or calculation of data. These may refer to instrument configurations, constant values employed by specific methodologies, or other supporting information.

*Example:*

*Explanatory Notes:*

**Instrument**

*Definition:* The mechanism employed in observation/measurement. Does not possess agency.

*Example:*

*Explanatory Notes:*

**Observation Resource**

*Definition:* A resource used in an Observation Event, not otherwise specified in the model (e.g., not an Instrument, Configuration, etc.)

*Example:*

*Explanatory Notes:*

**Sampling Resource**

*Definition:* A resource used in a Sampling Event.

*Example:*

*Explanatory Notes:*

**Data Resource**

*Definition:* A resource used in a Data Processing (or similar) Event: metadata.

*Example:*

*Explanatory Notes:*

**Data**

*Definition:* Measurements or observations collected during an observation event, or produced through derivation in a generation event. ALTERNATE: A set of observations or measurements which describe the phenomenon or object (including samples) being studied. These are primary inputs to analysis.

*Example:*

*Explanatory Notes:*

Supplementary information entities:

**Documentation**

*Definition:* A description of another resource, event, or entity provided for the purpose of understanding or employing it. **[QUERY: IS THIS DIFFERENT FROM METADATA?]**

*Example:*

*Explanatory Notes:*

**Quality Criteria**

*Definition:* The thresholds used in testing the quality or validity of something.

*Example:*

*Explanatory Notes:*

**Quality Metrics/Report**

*Definition:* The results of a quality test or validation, in enumerated or written form.

*Example:*

*Explanatory Notes:*

## Agents

**Observer**

*Definition:* An entity capable of volition which conducts an Observation/Measurement Event.

*Example:*

*Explanatory Notes:*

**Generator**

*Definition:* An entity capable of volition which conducts a Generation Event.

*Example:*

*Explanatory Notes:*

**Participant**

*Definition:* An entity which is involved in an event, whose role is otherwise unspecified in the model.

*Example:*

*Explanatory Notes:*

**Software Agent**

*Definition:* An entity that acts on behalf of a human to effect some change in the data.

*Example:*

*Explanatory Notes:*

**[TO DO: Add Agent roles from XAS example!]**

# Worked Examples

## Describing an Instrument

[Take from the XAS use case]

## Describing Survey Data Collection

## Describing Samples

## Describing Data Aggregation and Harmonization

## Others?