W3C PROV FAMILY OF SPECIFICATIONS: AN UPDATE

Paul Groth – W3C co-chair Provenance Working Group

Thanks to
Luc Moreau
Yolanda Gil
Michael Lang Jr.
The entire W3C Provenance Working Group

http://www.w3.org/2005/Incubator/prov/wiki/What_Is_Provenance

Our Working Definition of Provenance

Provenance of a resource is **a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource**.

Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility.

- Is provenance = metadata? Or = trust? Or = authentication?
  - Provenance can be seen as metadata, but not all metadata is provenance
  - Provenance provides a substrate for deriving different trust metrics
  - Provenance records can be used to verify and authenticate among other uses

- Notice:
  - Provenance assertions can have their own provenance
  - Inference is useful if provenance records are incomplete/erroneous
  - There may be alternative accounts of provenance of the same resource

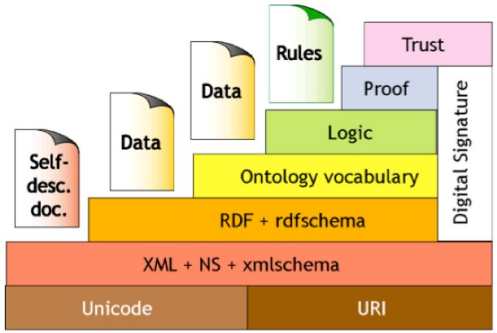
Broad Need for Provenance in Many Areas

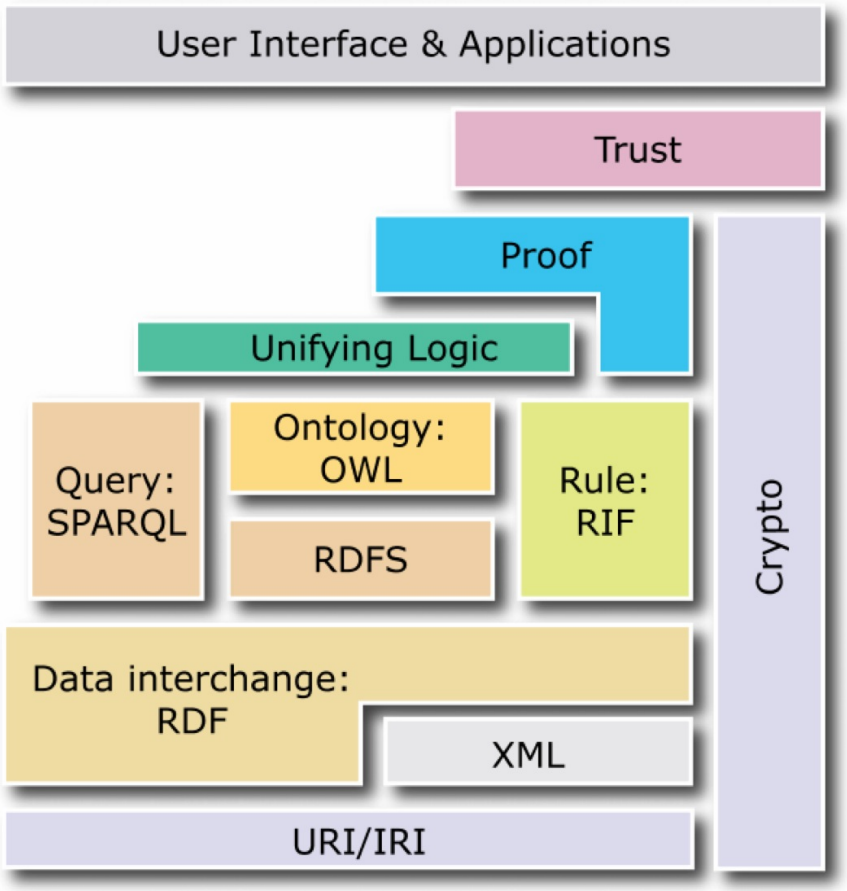
- Open information systems (such as the Web)
  - Making trust judgments on what web content to trust
- Business practices
  - Manufacturing processes and providers of a given product
- Science applications
  - How new results were obtained: from assumptions to conclusions and everything in between
- News-spheres
  - Blogosphere, twittosphere
- Laws for IP and privacy protection
  - Licensing and attribution of a document/software that combines permissions and rights of text, images, etc.
  - Privacy of information as well as of its provenance

Provenance and “Web Design Issues”

"At the toolbar (menu, whatever) associated with a document there is a button marked "Oh, yeah?". You press it when you lose that feeling of trust. It says to the Web, 'so how do I know I can trust this information?'. The software then goes directly or indirectly back to metainformation about the document, which suggests a number of reasons.”

- T. Berners-Lee, Web Design Issues, September 1997





W3C Chartered a New Provenance Group in Sept'09 (Chair: Y. Gil)

- Provenance is a pressing issue in many areas for W3C
  - Linked Data and the semantic web (linkedopendata.org)
  - Open government (data.gov, data.gov.uk)
  - HCLS
- Most people do not know how to approach provenance
  - Many are asking for a standard and methodology that they can use immediately
- Existing work scattered in many areas of computer science and library sciences research
  - *"The number of publications on provenance is [...] a total of 425 [...] with about half the papers published in the last two years." – Luc Moreau*

W3C Technology and Society domain

Provenance Interchange Working Group Charter

The **mission** of the Provenance Working Group, part of the Semantic Web Activity, is to support the widespread publication and use of provenance information of Web documents, data, and resources. The Working Group will publish W3C Recommendations that define a language for *exchanging* provenance information among applications.

Join the Provenance Working Group.

End date	1 October 2012
Confidentiality	Proceedings are public
Initial Chairs	Luc Moreau, University of Southampton Paul Groth, VU University Amsterdam
Initial Team Contacts (FTE %: 20)	Sandro Hawke
Usual Meeting Schedule	Teleconferences: Weekly Face-to-face: Once Annually

2. Scope

The Provenance Interchange Working Group has the following objective: to define a provenance interchange language and define methods to publish and access provenance using that language. For the purpose of this charter, we refer to the provenance interchange language to be defined as “PIL”. The Working Group will select an appropriate name for the language.

PIL should

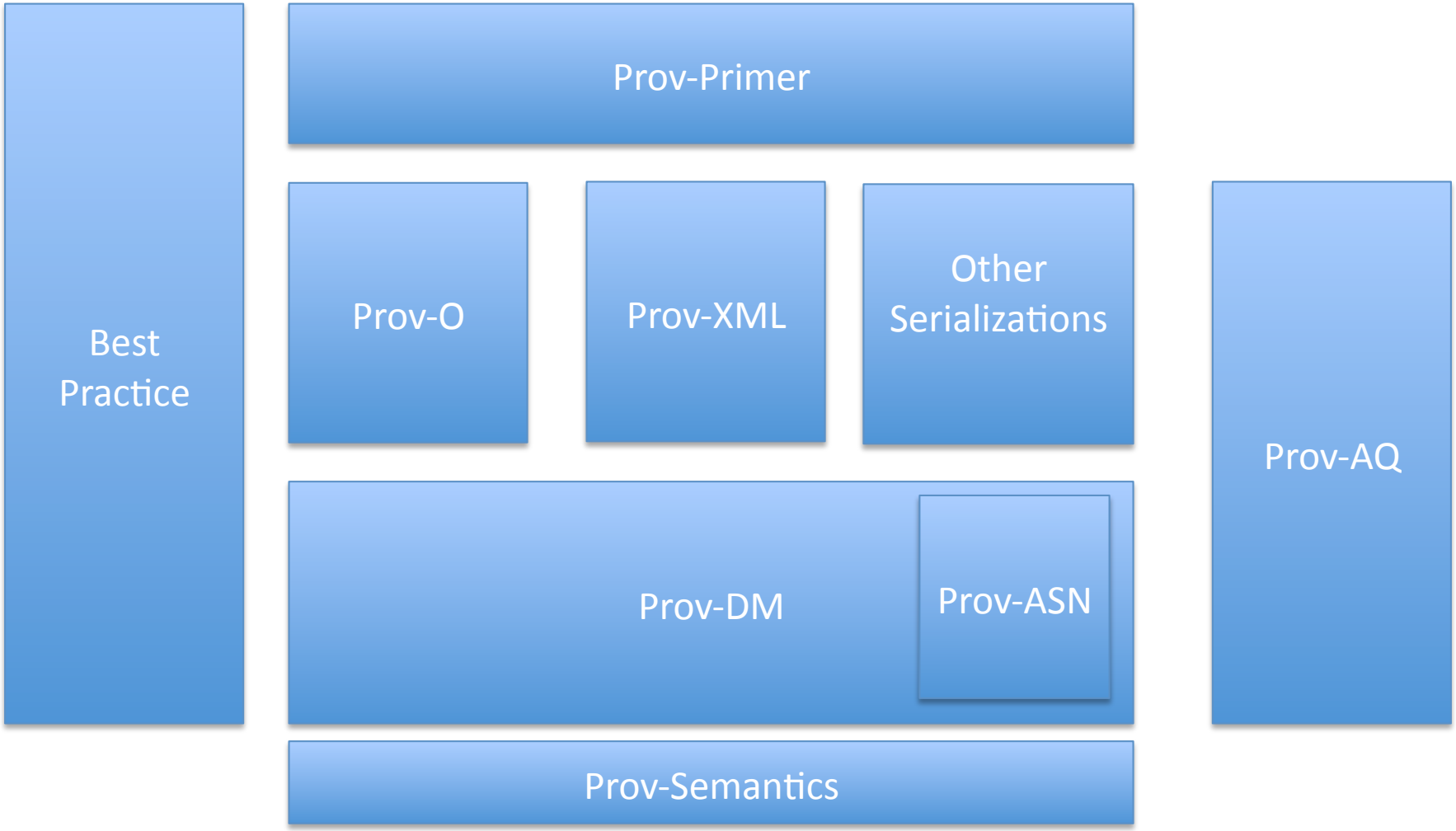
- be applicable to any resource, not just for Semantic Web objects;
- have a low entry point to facilitate widespread adoption, and makes it easy to do simple things;
- have a small core model and allow for extensions (ie, profiles, integration of other more expressive/complementary vocabularies/frameworks);

Drawing on existing vocabularies/ontologies (namely: Changeset Vocabulary, Dublin Core, Open Provenance Model (OPM), PREMIS, Proof Markup Language (PML), Provenance Vocabulary, Provenir ontology, SWAN Provenance Ontology, Semantic Web Publishing Vocabulary, WOT Schema; see also the Incubator Group Report for a more detailed description of those), the Incubator Group has identified a set of concepts that will constitute the core of PIL (see Section 8.1.4 of the Incubator Group report). The number of concepts is intentionally limited, so as to ensure a cohesive and tractable core. Other concepts can be relevant to provenance, but it is anticipated that they would be defined by means of the envisaged extension mechanism of PIL.

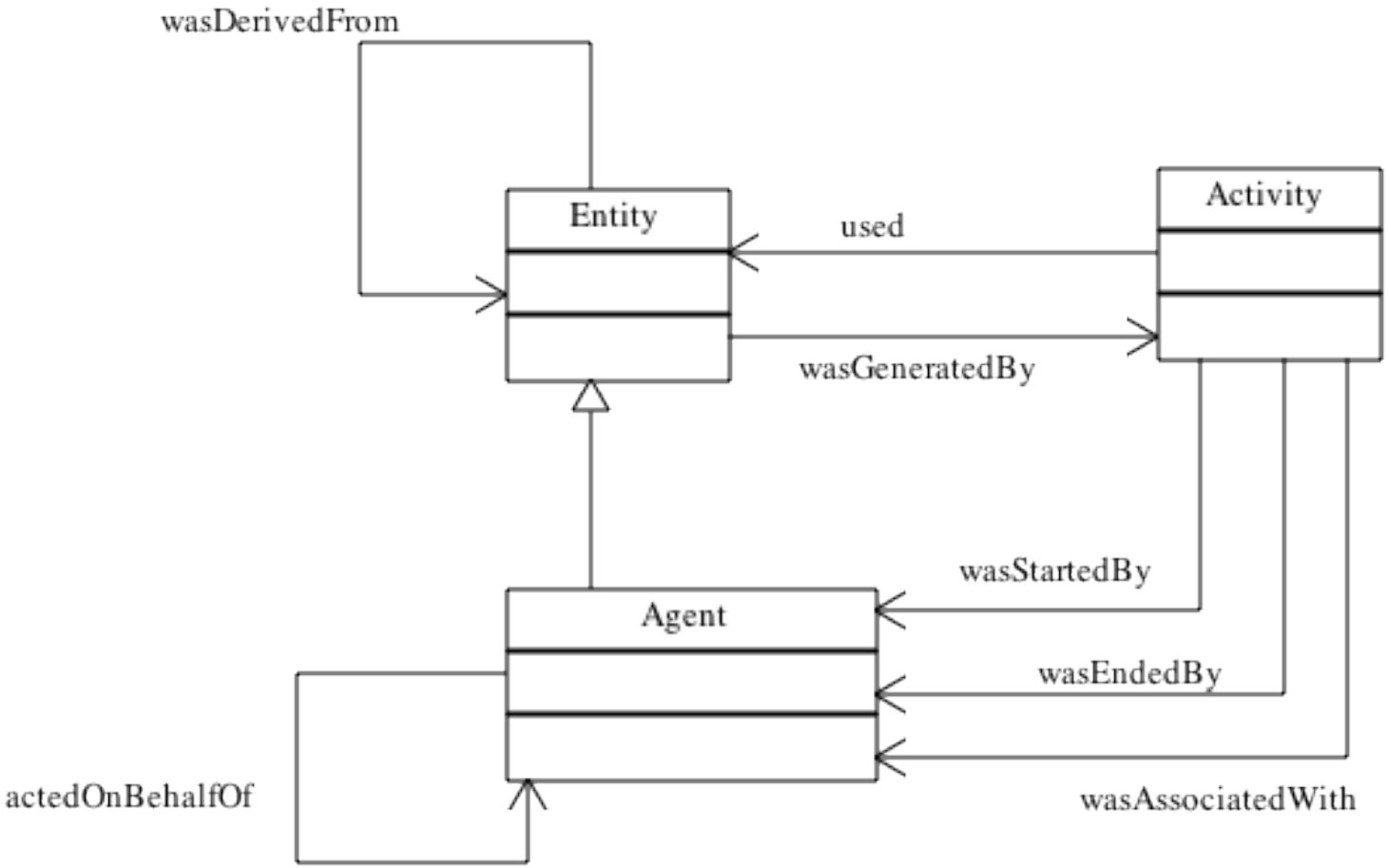
Who's involved

- **The Web Community**
- DERI Galway
- European Broadcasting Union
- FORTH
- Financial Services Technology Consortium
- DFKI
- IBBT
- Information Sciences Institute
- Library of Congress
- NASA
- National Cancer Institute
- Open Geospatial Consortium
- OpenLink Software
- Oracle
- Pacific Northwest National Laboratory
- Rensselaer Polytechnic Institute
- Revelytix, Inc
- Newcastle University
- The National Archives
- TopQuadrant
- Universidad Politecnica de Madrid
- University of Edinburgh
- University of Manchester
- University of Oxford
- University of Southampton
- VU University Amsterdam
- Wright State University

Layered Approach



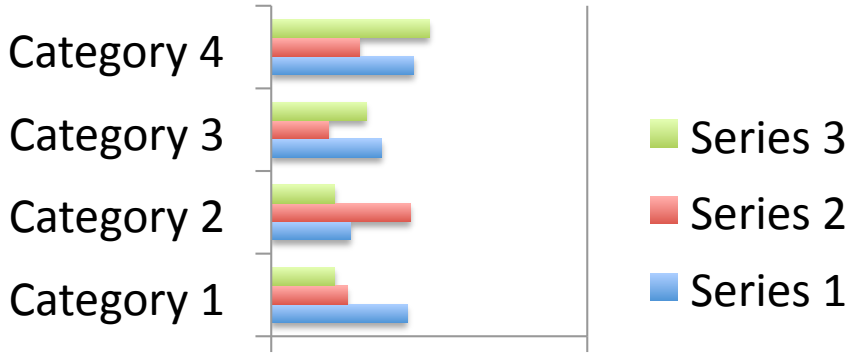
PROV-DM: Data Model



Running Example

Using excel, Alice created a chart from a data set.

What is the provenance of the chart?

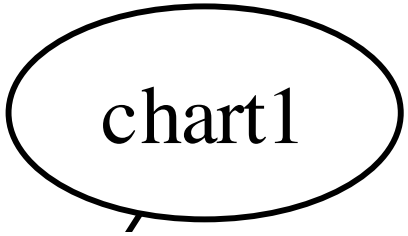


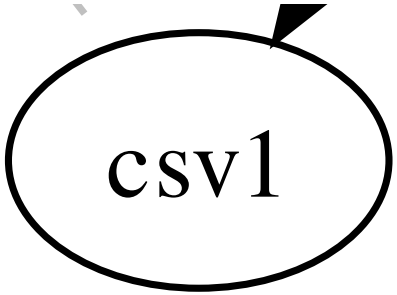
	Series 1	Series 2	Series 3
Category 1	4.3	2.4	2
Category 2	2.5	4.4	2
Category 3	3.5	1.8	3
Category 4	4.5	2.8	5

Entities

- Entities are things in the world one wants to provide provenance for.

- Examples:
  - document at URI http://www.w3.org/TR/prov-dm/,
  - a file in a file system
  - a car
  - an idea.

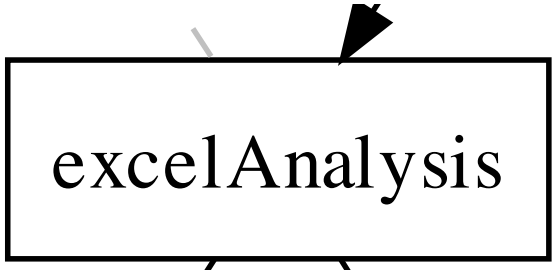




Activities

- An activity is anything that acts upon or with entities; this action can take multiple forms: consuming, processing, transforming, modifying, relocating, using, generating, or being associated with entities, etc.

- Examples:
  - publishing of a document on the web,
  - sending a twitter message,
  - driving a car from Boston to Cambridge,
  - assembling a data set based on a set of measurements,
  - performing a statistical analysis over a data set,
  - sorting news items according to some criteria, running a SPARQL query over a triple store, and editing a file.



Agent

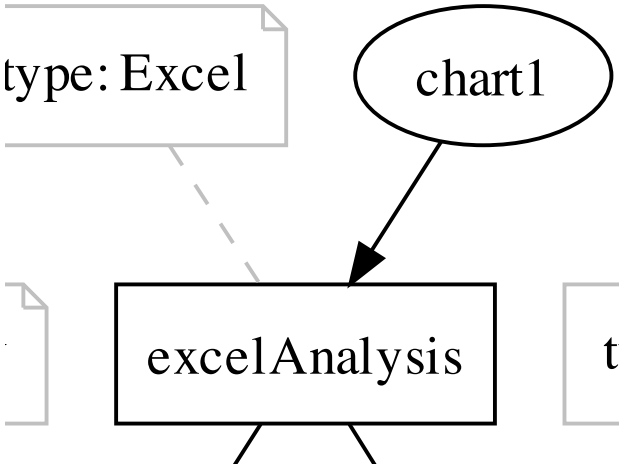
- An agent is a type of entity that takes an active role in an activity such that it can be assigned some degree of responsibility for the activity taking place
- Examples
  - People
  - Organizations
  - Software

Linking them together

Generation

- Generation is the completed production of a new entity by an activity.

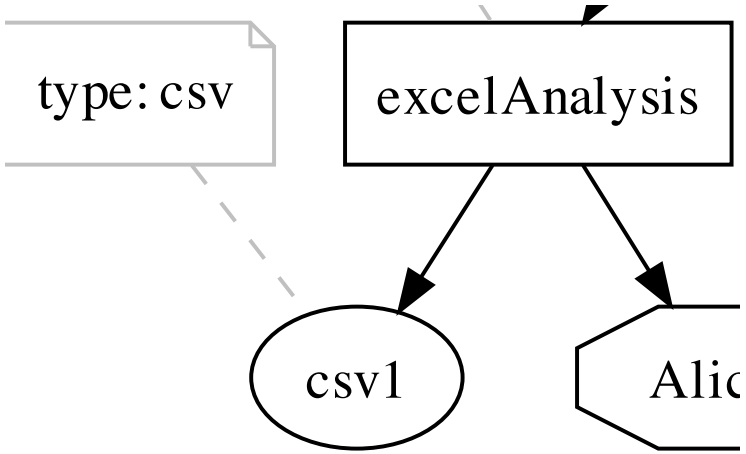
- Examples:
  - the creation of a file by a program
  - creation of a linked data set
  - publication of a new version of a document.



Usage

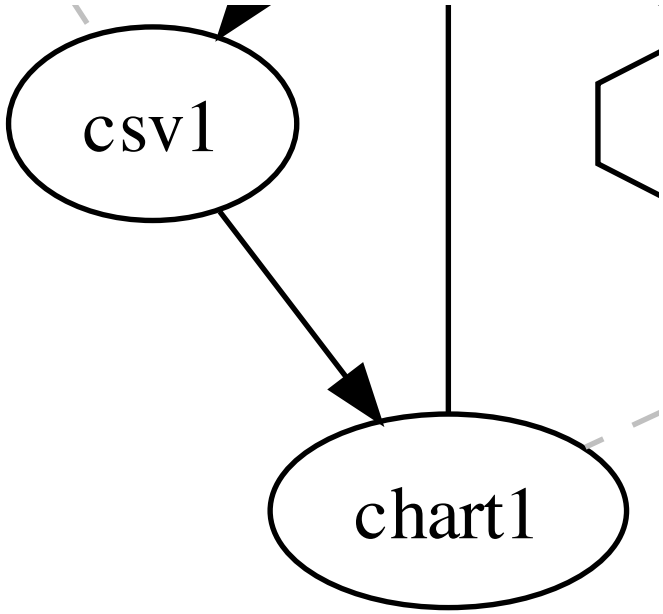
- Usage is the beginning of an entity being consumed by an activity.

- Examples:
  - program beginning to read a configuration file
  - a document used in a legal proceeding
  - A dataset used in a statistical analysis



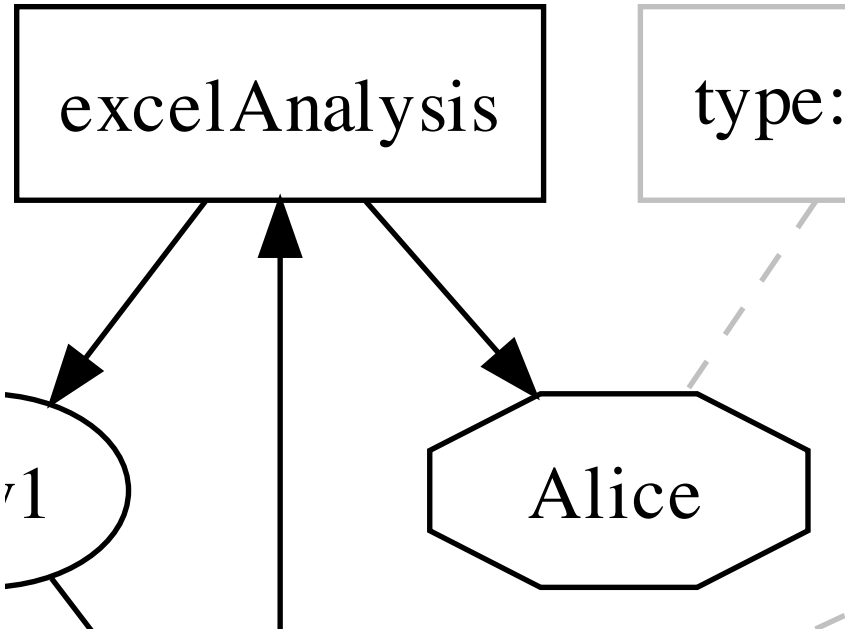
Derivation

- Derivation of an entity from an another is a relation that denotes that the derived entity is transformed from, created from, or affected by the deriving entity.
- Examples:
  - the transformation of a relational table into a linked data set
  - the transformation of a canvas into a painting
  - the transportation of a work of art from London to New York

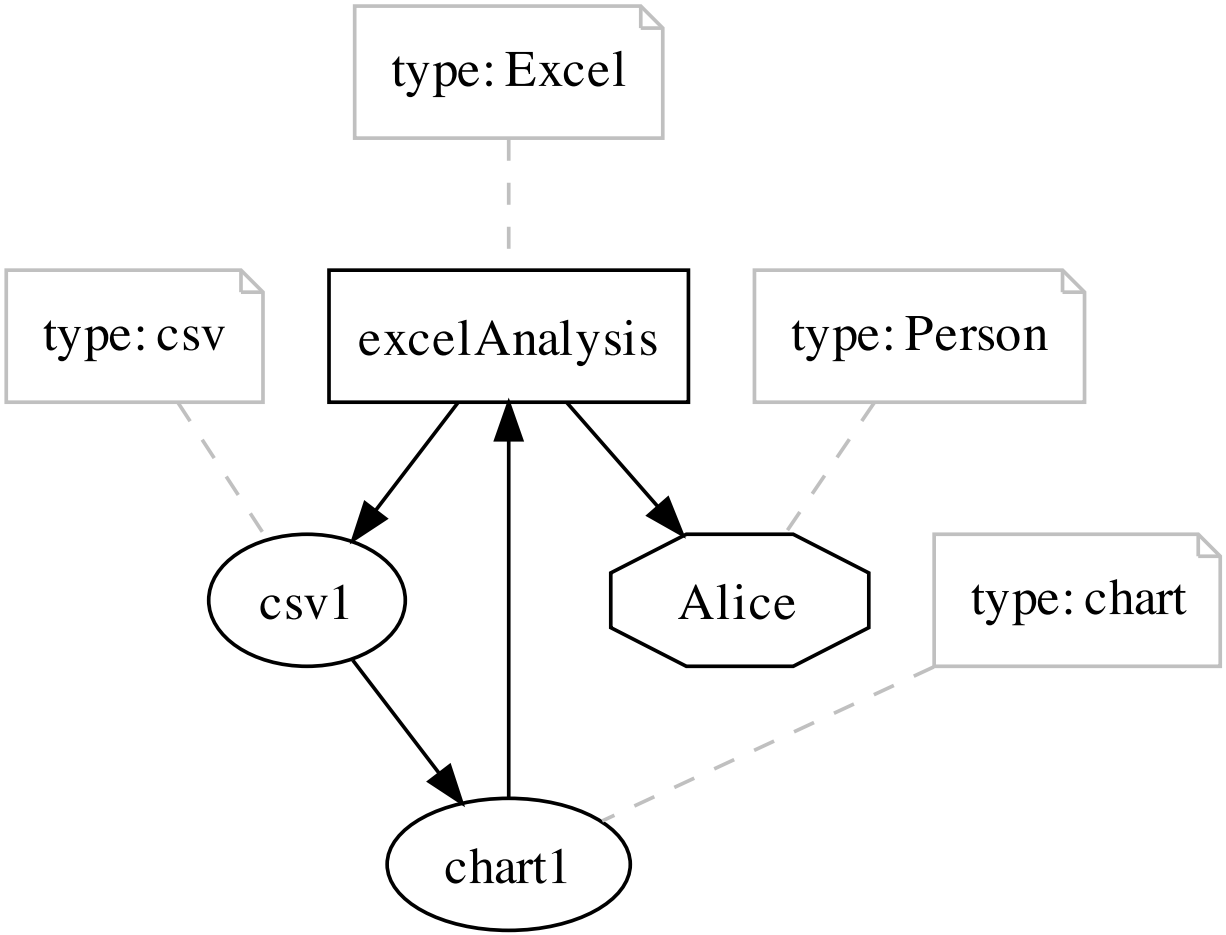


Activity Association

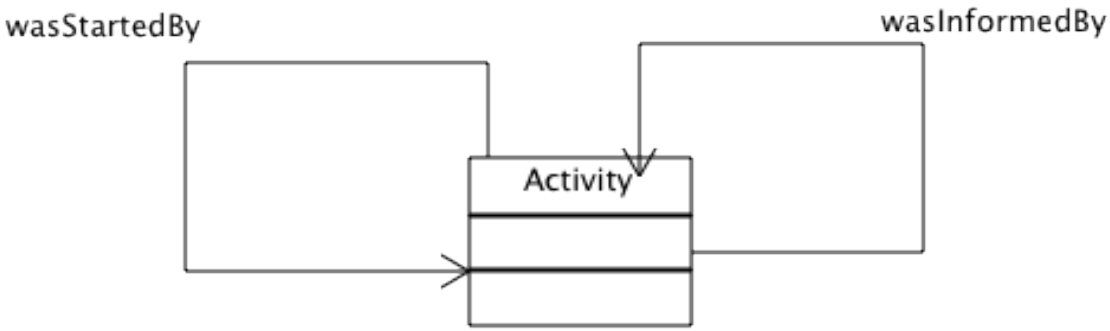
- activity association is an assignment of responsibility to an agent for an activity, indicating that the agent had an active role in the activity
- Examples:
  - creation of a web page under the guidance of a designer;
  - various forms of participation in a panel discussion, including audience member, panelist, or panel chair;
  - a public event, sponsored by a company, and hosted by a museum;
  - an XSLT transform initiated by a user;

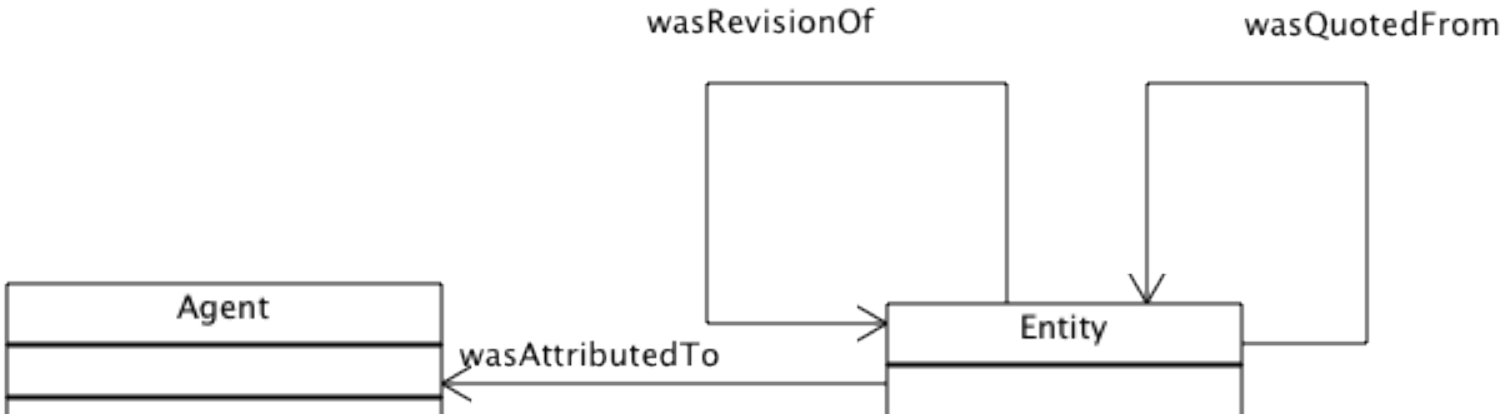


The Full Example



Other Handy Statements





In RDF

```
@prefix prov: <http://www.w3.org/ns/prov-o/> .
@prefix ex: <http://example.org/>.

ex:csv1 a prov:Entity .
ex:chart1 a prov:Entity .

ex:excelAnalysis a prov:Activity ;
        prov:used ex:csv1 .

ex:chart prov:wasGeneratedBy ex:excelAnalysis .

ex:Alice a prov:Agent .
ex:Alice prov:wasAssociatedWith ex:excelAnalysis.
ex:csv1 prov:wasDerivedFrom ex:chart1 .
```

Other Specs

- All drafts are available:
  - http://www.w3.org/2011/prov/wiki/Main_Page
- PROV-O
  - The OWL-RL Ontology of the data model
- PROV-PRIMER
- PROV-AQ
  - How to access provenance for a web resource
- PROV-SEM
  - A formal semantics of the data model

Status

- Core constructs are stable
- Working hard on simplifying explanations
- OWL-RL ontology available but under revision
- Start work on best practices with respect to Dublin Core
- Time table:
  - synchronous release of prov-primer, prov-o and prov-dm within the next month
  - Looking for a solid base in the next 2 months or so

How the community can help

- Looking for a wide range of users
- Feedback is appreciated and welcome
  - Best practices for extension
  - Implementations
  - Clarity of document
- Let us know what you think
  - public-prov-wg@w3.org