W3C PROV Introduction

ESWC 2013

Paul Groth

Slides from Ivan Herman and Luc Moreau

Plan for today

- 09:30 - 10:00: Introduction
- 10:00 - 10:30: A Walk Through of PROV-O
- 10:30 - 11:00: Coffee!!
- 11:00 – 11:15: PROV-CONSTRAINTS
- 11:15 – 11:45: PROV-AQ
- 11:45 - 12:30: PROV Hands On

The goal is simple…

- We should be able to express special “meta” information on the data
  - who played what role in creating the data (author, reviewer, etc.)
  - view of the full revision chain of the data
  - in case of integrated data which part comes from which original data and under what process
  - what vocabularies/ontologies/rules were used to generate some portions of the data
  - etc.

…the solution is more complicated

- Requires a complete model describing the various constituents (actors, revisions, etc.)
- The model should be usable with RDF to be used on the Semantic Web
- Has to find a balance between
  - simple (“scruffy”) provenance: easily usable and editable
  - complex (“complete”) provenance: allows for a detailed reporting of origins, versions, etc.

Lots of application areas need provenance

- Open Information Systems
  - origin of the data, who was responsible for its creation
- Science applications
  - how the results were obtained
- News
  - origins and references of blogs, news items
- Law
  - licensing attribution of documents, data
  - privacy information
- Etc.

Definition of Provenance (by the Provenance WG)

Provenance is defined as a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing.

“Provenance” is not a new subject

- There has been lot of work around
  - workflow systems
  - databases
  - knowledge representation
  - information retrieval
- There are communities and vocabularies out there
  - Open Provenance Model (OPM)
  - Dublin Core
  - Provenir ontology
  - Provenance vocabulary
  - SWAN provenance ontology
  - etc.

W3C's Provenance Incubator Group

- Worked in 2009-2010 (Chaired by Yolanda Gil)
- Issued a final report
  - "Provenance XG Final Report"
    - http://www.w3.org/2005/Incubator/prov/XGR-prov/
  - provides an overview of the various existing approaches, vocabularies
  - proposes the creation of a dedicated W3C Working Group

W3C Provenance Working Group

- Set up in April 2011 (co-chaired by Paul Groth and Luc Moreau)
- Goal was to define a standard way to interchange provenance on the web.
- Specifically targets the semantic web
- This is what I will talk about in what follows…

Participants

- DERI Galway
- European Broadcasting Union
- FORTH
- Financial Services Technology Consortium
- DFKI
- IBBT
- IBM
- Library of Congress
- Mayo Clinic
- NASA
- OCLC
- 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 Aberdeen
- University of Edinburgh
- University of Manchester
- University of Oxford
- University of Southampton
- VU University Amsterdam
- Wright State University

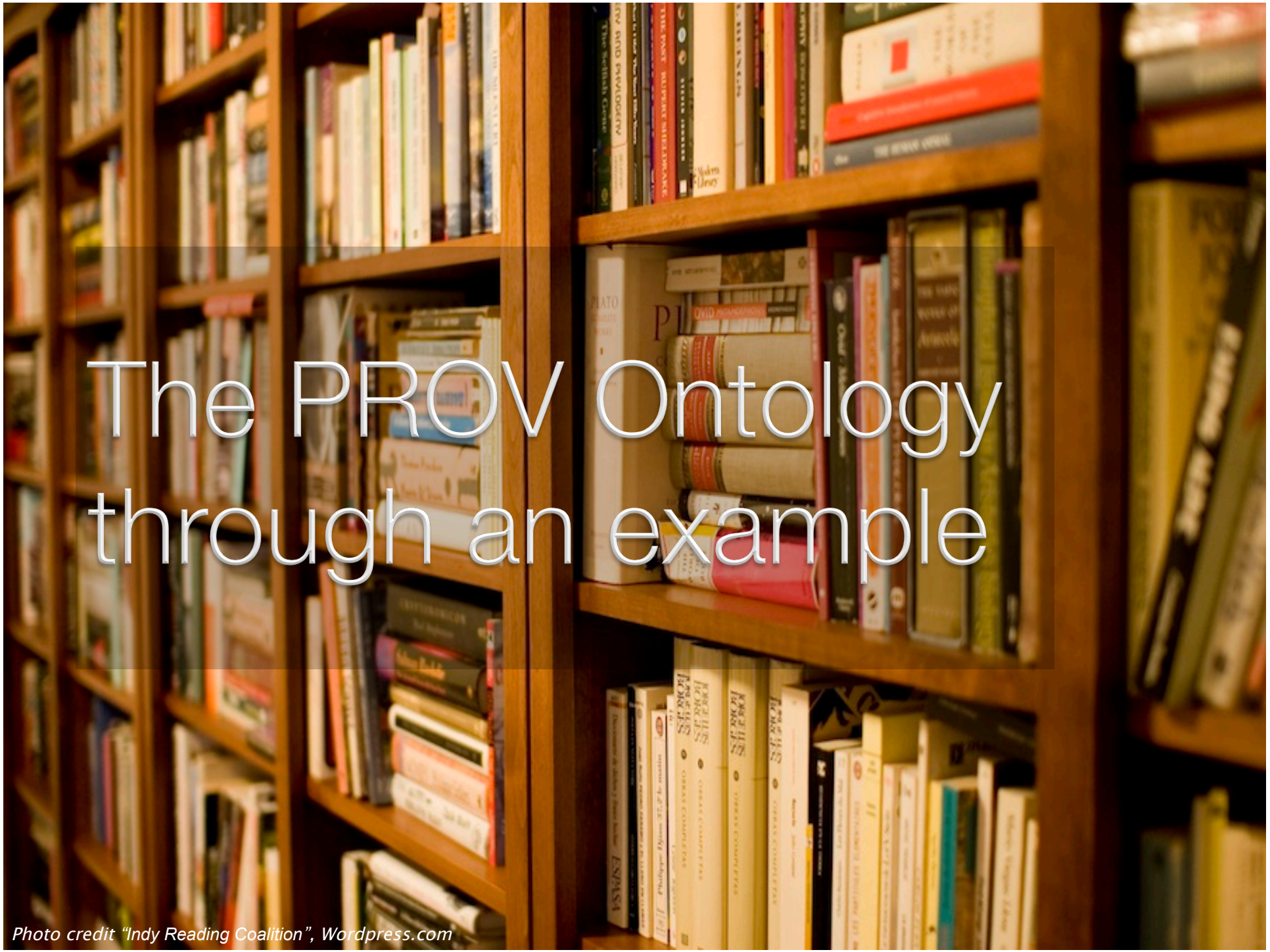
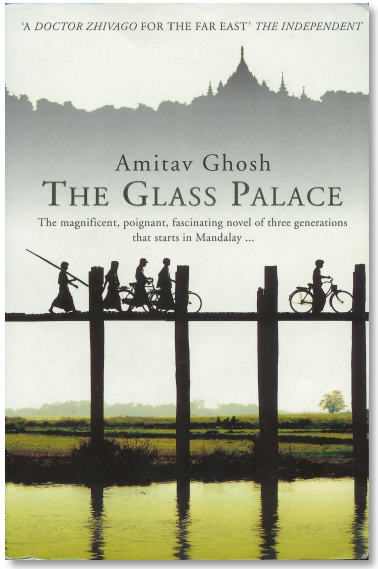
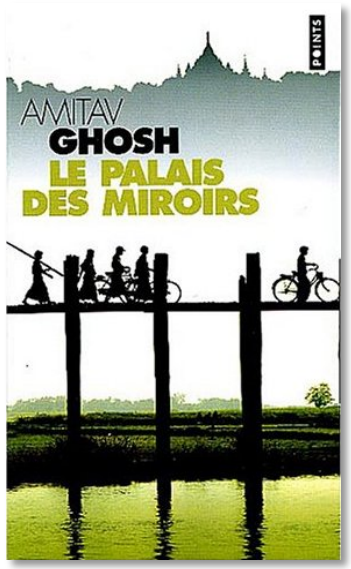


Photo credit "Indy Reading Coalition", Wordpress.com

The example

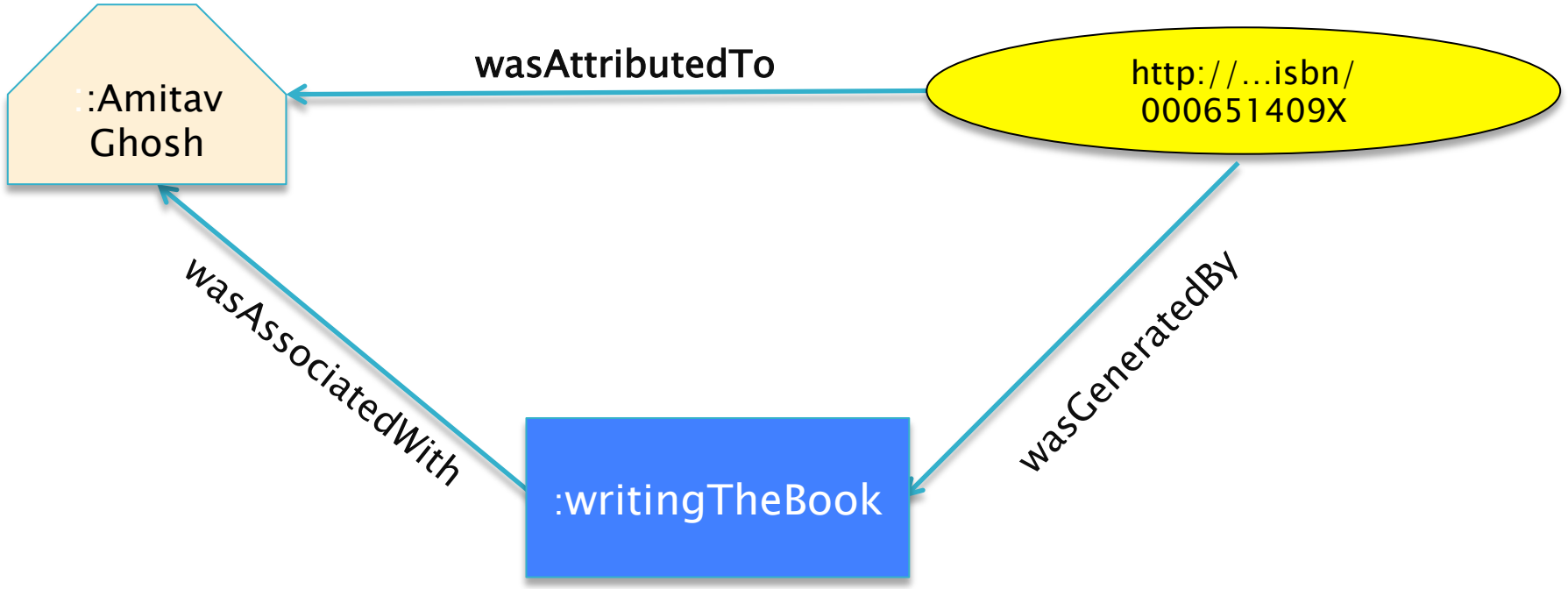
- We have data on two books
  - “The Glass Palace”, written by Amitav Ghosh
  - “Le palais des mirroirs”, the French translation, done by Christianne Besse, of the book of Amitav Ghosh
  - we want to describe some very basic facts on the provenance of these





A very simple attribution

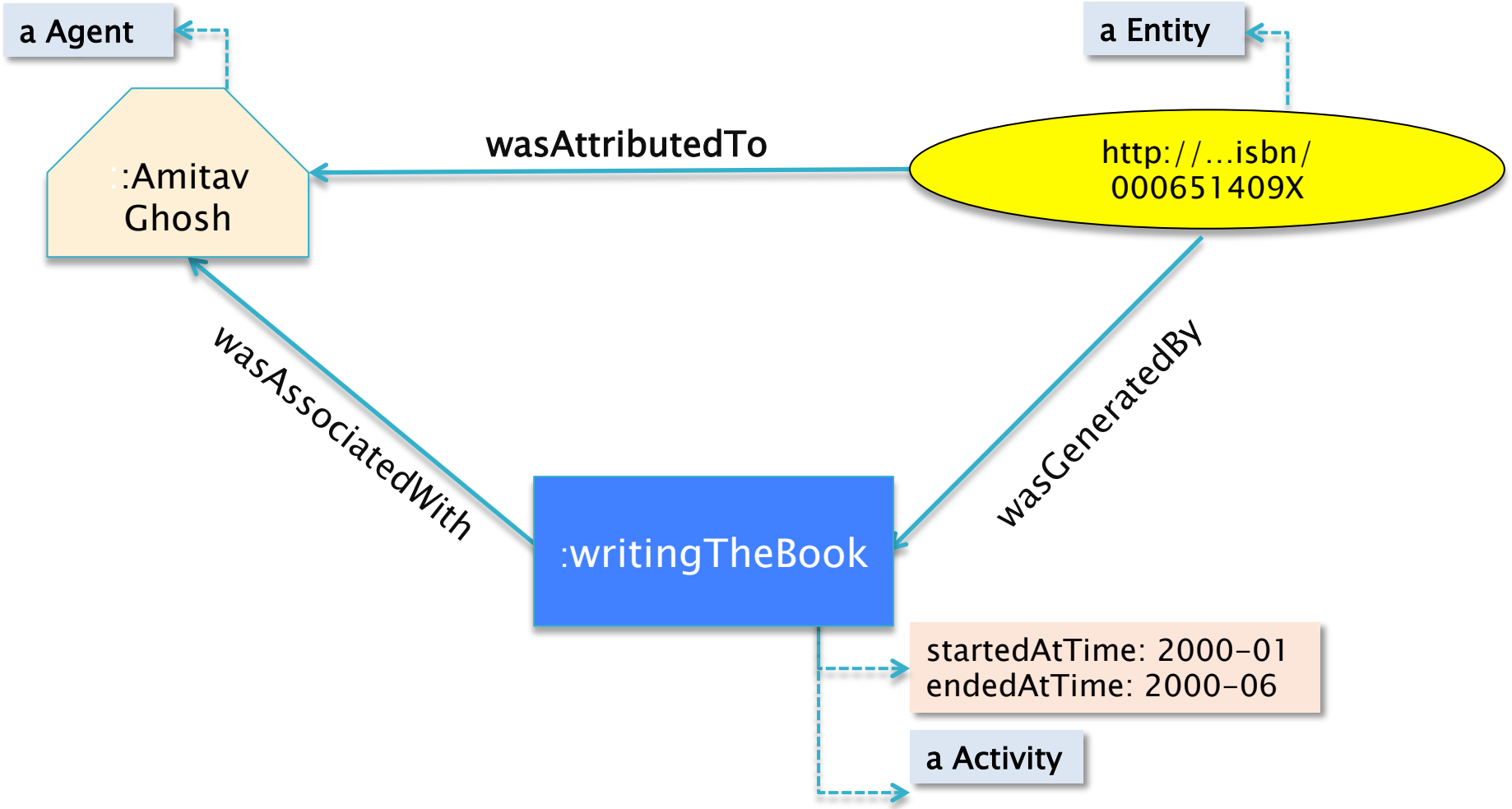
A bit more complicated: make the activity explicit



Why?

To make some “metadata” explicit

A more complete attribution: make the activity explicit

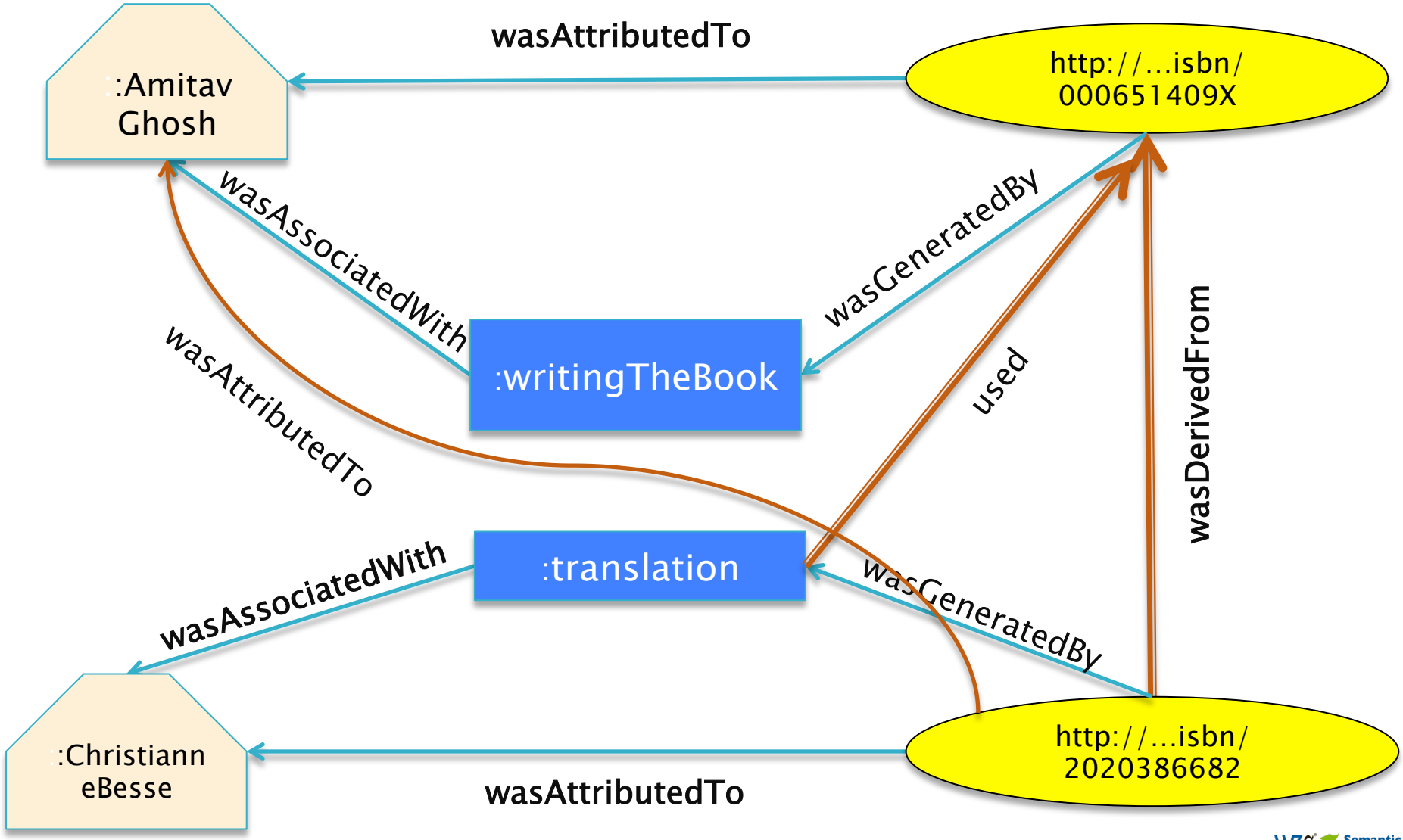


The fundamental notions of PROV

- This simple example shows the fundamental notions
  - Entity:
    - the “things” whose provenance we want to describe
  - Activity:
    - describes how entities are created, changed. The “dynamic” aspect of the world
  - Agent:
    - are responsible for the actions.
  - Usage, generation, derivation, attribution,..
    - connections describing how entities, agents, and activities interact

Let’s make it a bit more complex

Adding the translation…

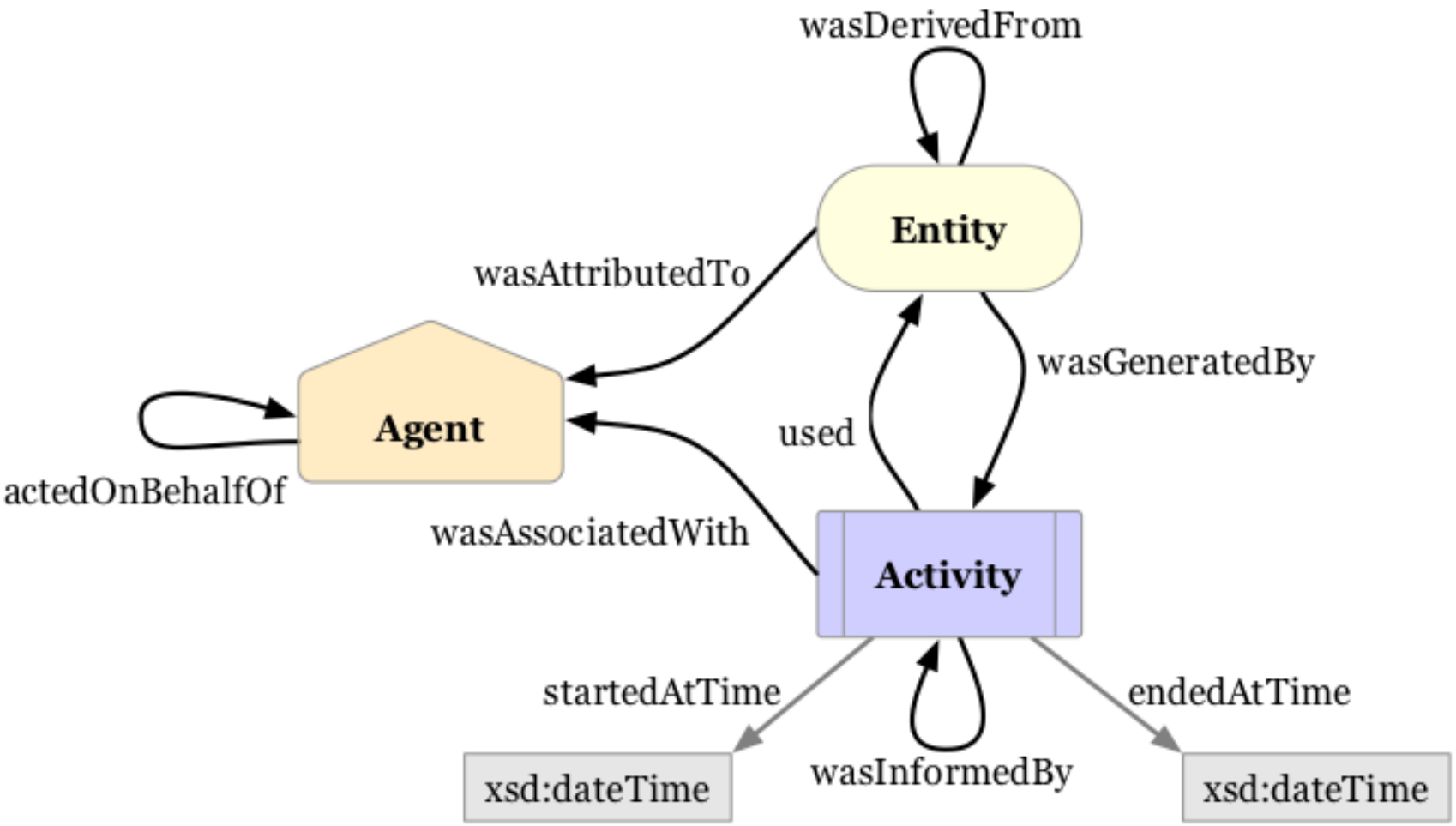


Categories of PROV Terms

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 provenance geeks ☺

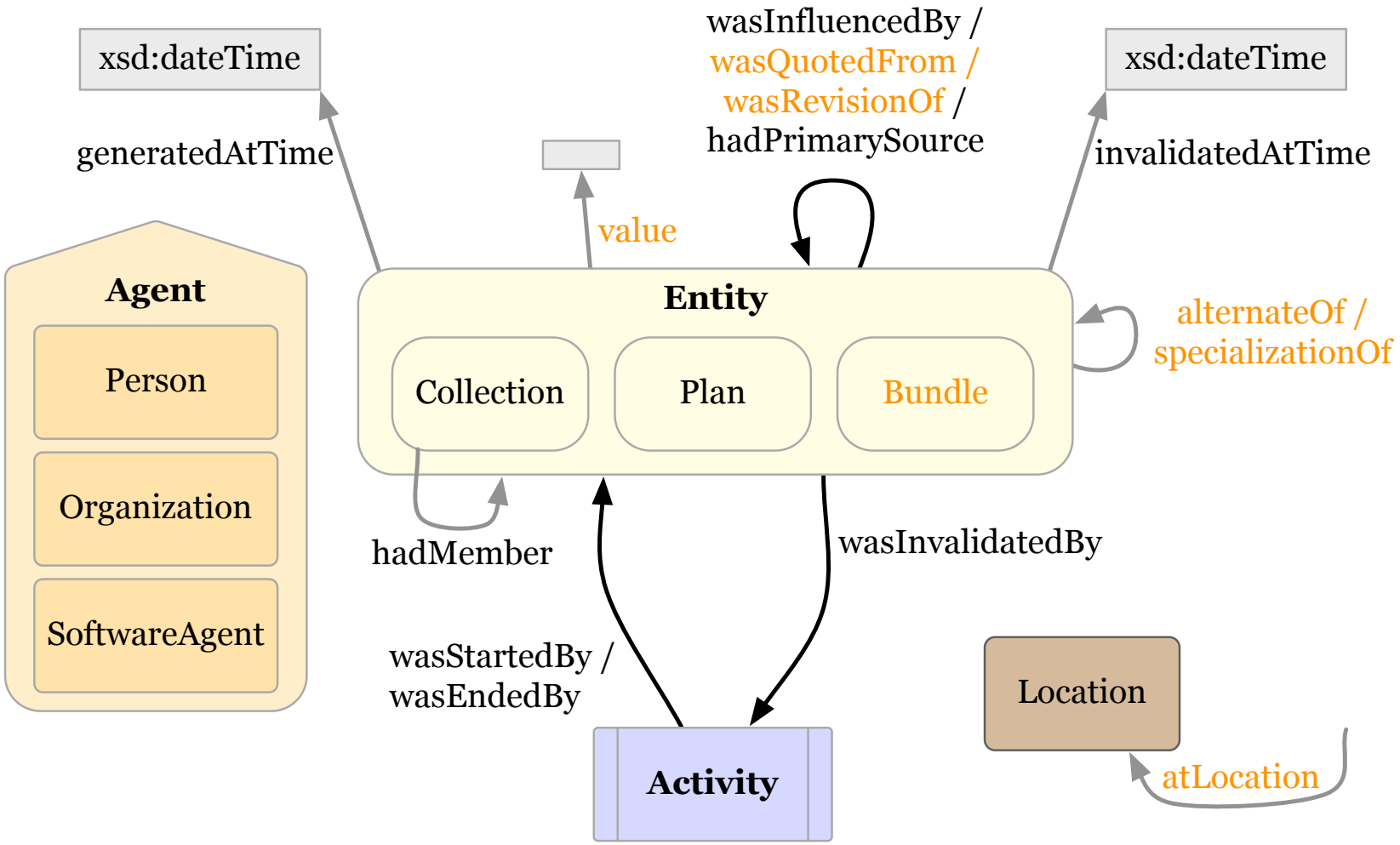
Starting point classes and properties



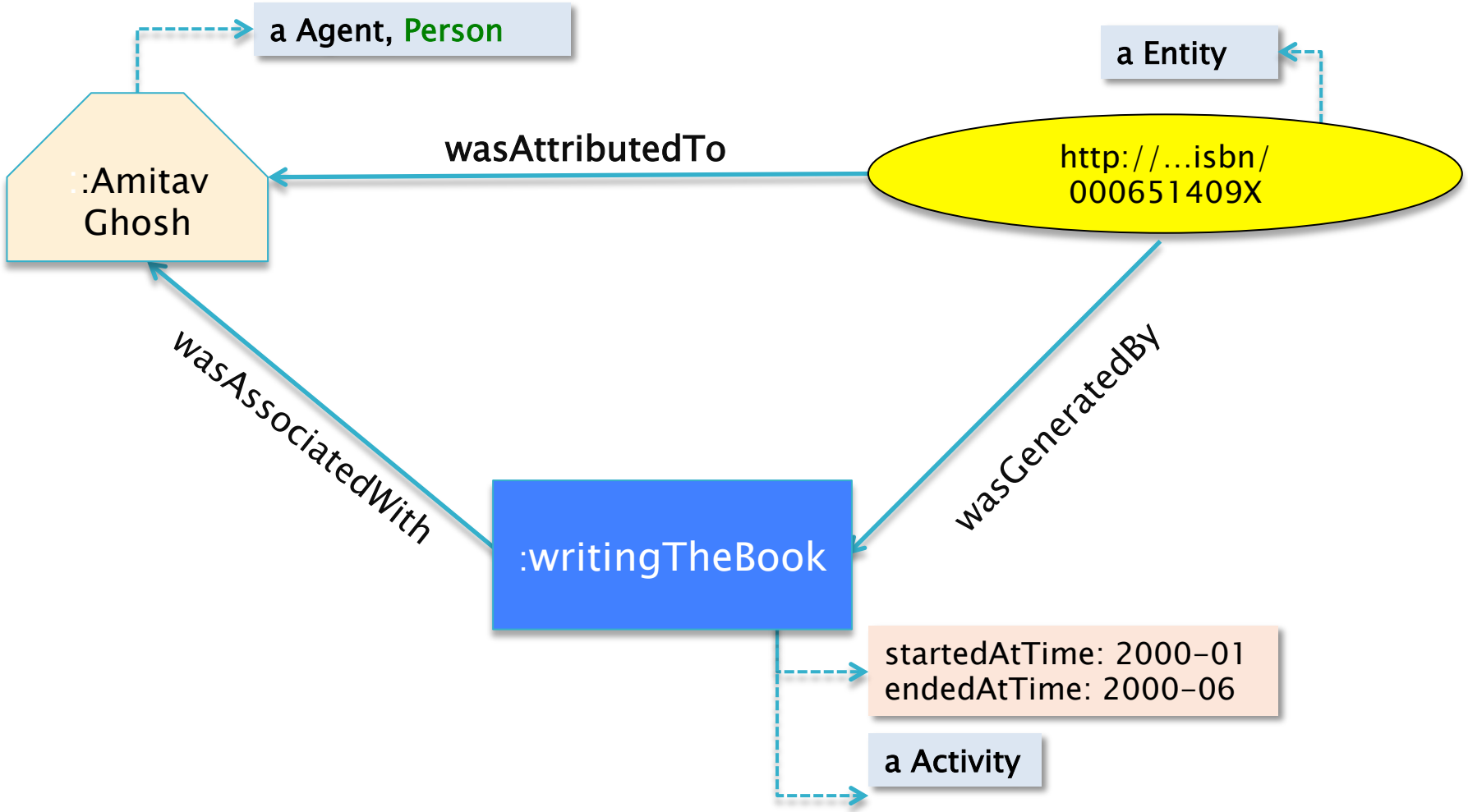
Expanded classes and properties

- Some extra classes, defined as subclasses of agents:
  - Organization, Person, SoftwareAgent
- Some extra properties describing versioning, influencing, invalidation, or creation of entities, etc.
- Nothing structurally different, just extended
  - applications are of course welcome to add their own specializations

Some examples for extra properties



Adding some extra properties




Relationship to
Dublin Core
courtesy to "analogue kid"

Dublin Core

- Complementary with PROV
  - some terms have direct mappings
  - some need a slightly more complex relationship

Some simple Dublin Core relationship examples

DC Term	Relation	PROV Term	Rationale
dct:created	rdfs:subPropertyOf	prov:generatedAtTime	Property used to describe the time of creation of a resource (i.e., the time of its generation). We map it as a subproperty of `prov:generatedAtTime` because "creation" is one of the many activities that generate an entity (for example, generation includes modification, issue, acceptance, etc.).
dct:creator	rdfs:subPropertyOf	prov:wasAttributedTo	A creator is one of the agents who participated in the creation of a resource. They have the attribution for the outcome of that activity.
dct:contributor	rdfs:subPropertyOf	prov:wasAttributedTo	A contributor is associated with either the creation activity or the updating of the resource. Therefore, he/she has attribution over the outcome of those activities.
dct:dateAccepted	rdfs:subPropertyOf	prov:generatedAtTime	Property used to describe the date when the resource was accepted. `dct:dateAccepted` is mapped as a subproperty of `prov:generatedAtTime` because the accepted resource was generated by an "Accept" activity which may have changed it from its previous state.

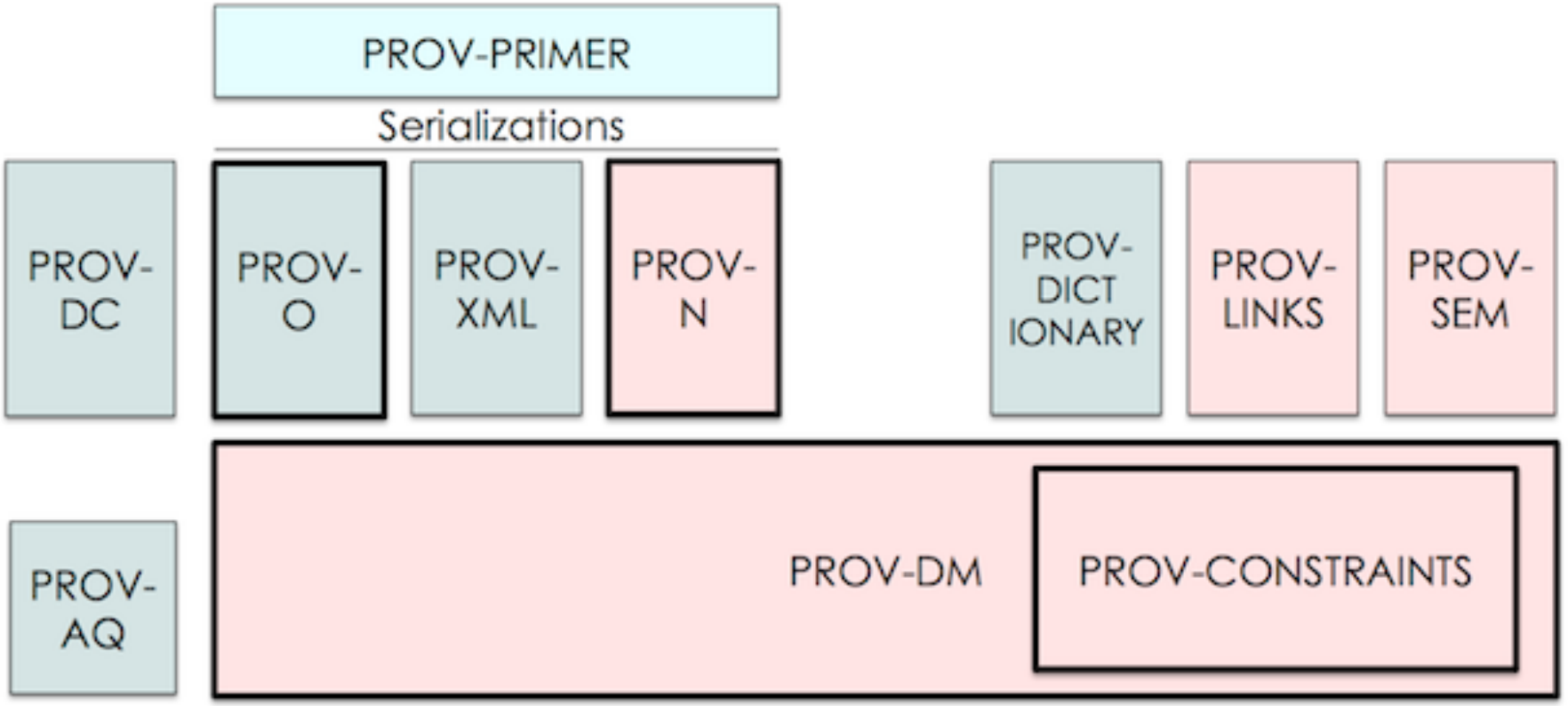
Some cases are more complicated

- For example, Dublin Core's "creator" has more to it than simply an agent. The correspondence is something like:
  - "If an entity is attributed to an agent, and the agent's role matches Dublin Core's definition of a creator, then the agent is the creator of the entity in the Dublin Core sense"
- These (few) cases are described in terms SPARQL CONSTRUCT rules

Available documents

Photo credit "Abizern", Flickr

Documents published by the Group



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

Namespace: http://www.w3.org/ns/prov#

Implementations

- 66 implementations
  - 41 systems
  - 22 vocab/datasets
  - 3 validators

Table 2: Coverage of PROV-DM terms in implementations o

PROV Component	Term	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	#18	#19	#20	#21	#22
C1: Entities/Activities	Entity	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		●	●	●
	Activity	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	
	Generation	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	
	Usage	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	
	Communication				●			●			●	●			●			●		●			
	Start	●	●		●			●	●		●	●						●	●			●	
	End	●	●		●			●	●		●	●	●			●	●	●		●	●	●	
	Invalidation							●			●	●						●		●			

Thank you for your attention

prov:wasDerivedFrom
https://dvcs.w3.org/hg/prov/file/tip/presentations/iswc-2012/prov-intro-iswc2012.pptx
prov:wasDerivedFrom
http://www.w3.org/2012/Talks/1009-MIT-IH/
prov:wasAttributedTo
Ivan Herman