Open Information Systems 2019-2020

Lecture 10: Dataset Metadata and Provenance Information

Christophe Debruyne

Introduction

So far, we have covered:

- What ontologies are and their role in addressing the problem of semantic heterogeneity
- Knowledge representation on the Web
  - RDF as a data model
  - RDFS, OWL, and Rule Languages as Ontology Languages
- Various ways of servicing meaningful RDF: SPARQL and Linked Data
- Uplift: transforming non-RDF into RDF
- Ontology matching, and ontology evaluation

In this lecture, we will cover some important vocabularies.

What is a vocabulary?

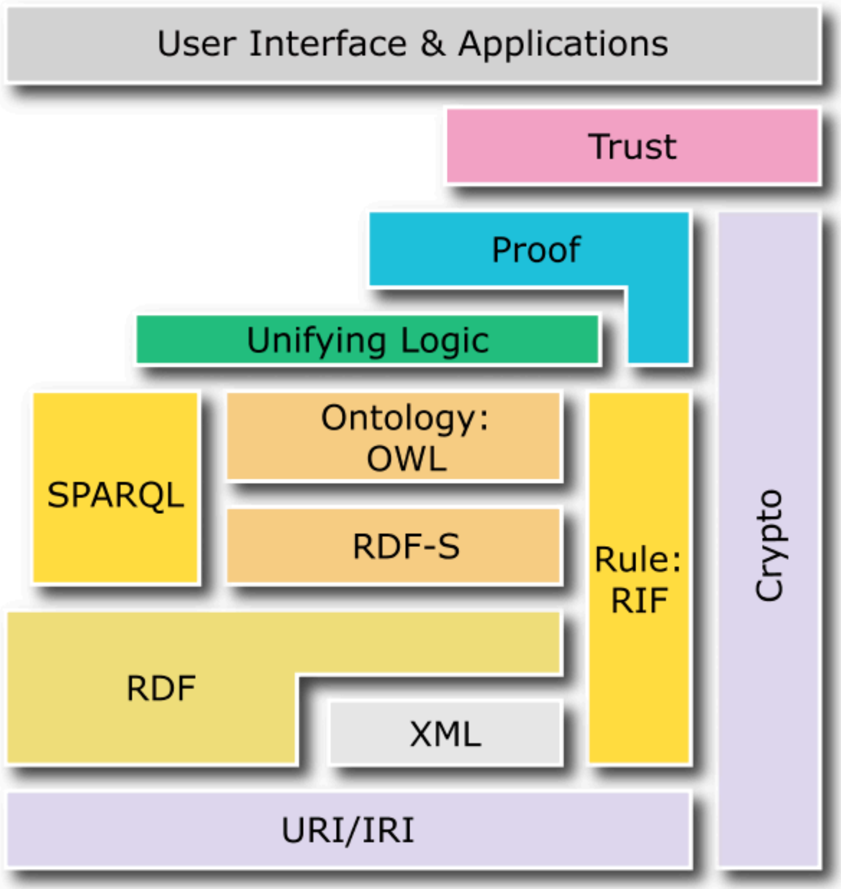
A vocabulary is a “lightweight” ontology. A vocabulary usually consists of a type hierarchy, a property hierarchy, and very few axioms. Domain- and range declarations are fine, inverse property declarations are fine, some disjoint classes are fine. The goal of a vocabulary is to support interoperability, not to support certain reasoning tasks.

Provenance

- **What is provenance?** Provenance information provides insights on a resource’s origin, such as who created that resource, when it was modified, or how it was created [1].
  - The title of a paper is metadata, but not provenance information.
- **Why provenance information?** Provenance is key in evaluating the quality of, and establishing trust in information on the Web [2].
- **Where is it used?** GLAM, (scientific) publications, tracking the activities and intermediate outcomes of ETL processes, provenance information of (news) articles, GDPR, etc.

PROV-O

The PROV Ontology PROV-O IS a W3C Recommendation (since 2013) for representing and exchanging provenance information as RDF.
https://www.w3.org/TR/prov-o/



https://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/#(24)

PROV-O

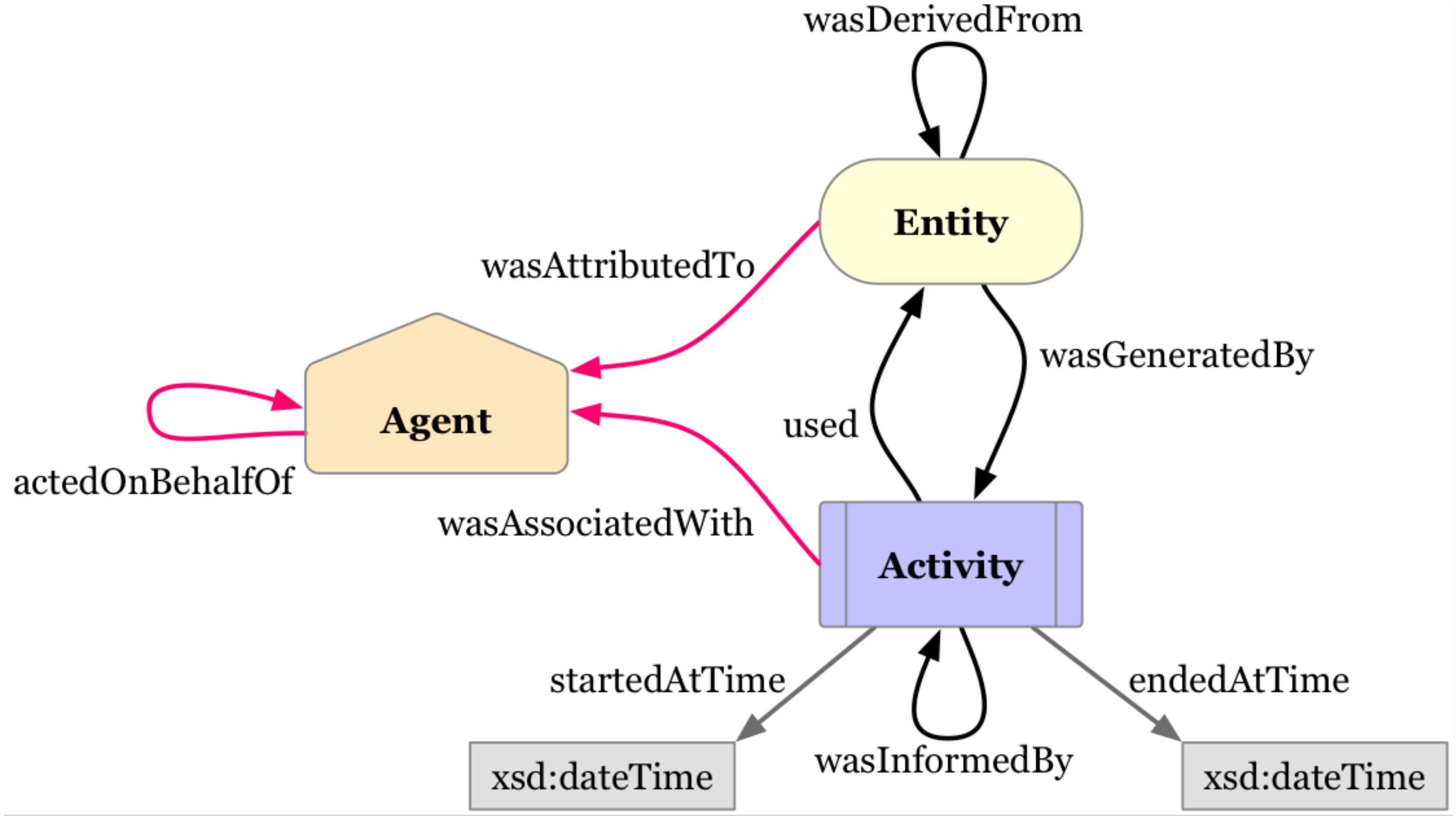
A lot of effort and thought has been put in developing PROV-O

- The ontology engineers wanted to remain in the OWL 2 RL profile to allow efficient and scalable reasoning, but: they realized that some axioms (anonymous concept unions) were not within that realm. They proposed changes to the ontology that one can apply (with some information loss) that fits the RL flavor. -> well documented
- The ontology engineers wanted to avoid inverse properties and only introduced two to maximize interoperability at graph level. To avoid the introduction of inverse properties, PROV-O also “reserved” the properties that could have been declared. In other words, you should not declare and use those non-inexistent inverse properties.

All decisions were thought out, well informed, and documented.

PROV-O Core Concepts and Relations

```
Namespace prov: <http://www.w3.org/ns/prov#> .
```



Core concepts and relations in PROV-O from, Copyright © 2011-2013 W3C® (MIT, ERCIM, Keio, Beihang).

PROV-O Core Concepts and Relations

```
Namespace prov: <http://www.w3.org/ns/prov#> .
```

	Activity	Entity	Agent
Activity	wasInformedBy	used	wasAssociatedWith
Entity	wasGeneratedBy	wasDerivedFrom	wasAttributedTo
Agent	--	--	actedOnBehalfOf

Accountability!

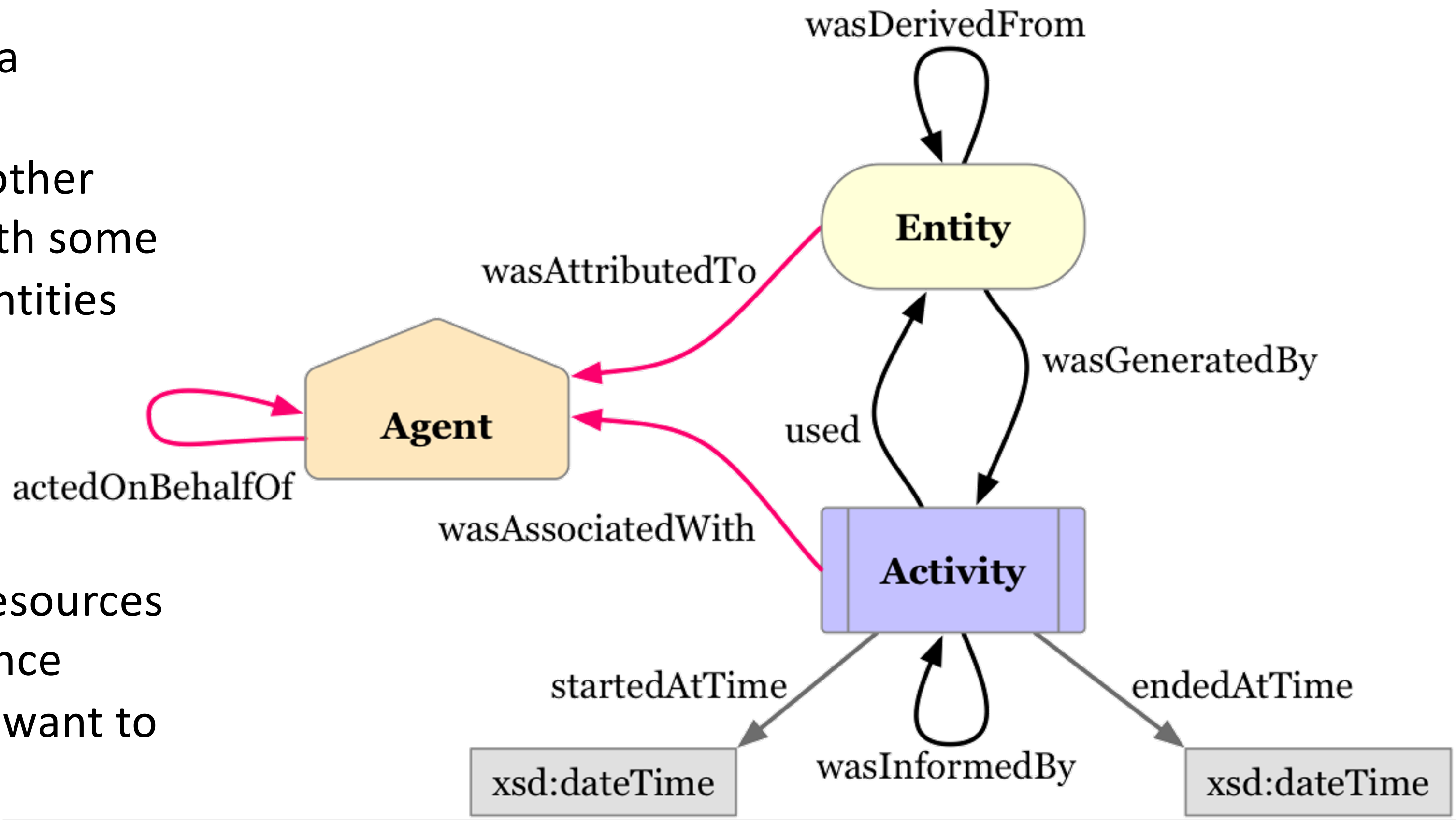
Core concepts and relations in PROV-O from, Copyright © 2011-2013 W3C® (MIT, ERCIM, Keio, Beihang).

PROV-O Core Concepts and Relations

```
Namespace prov: <http://www.w3.org/ns/prov#> .
```

A prov:Entity is a physical, digital, conceptual, or other kind of thing with some fixed aspects; entities may be real or imaginary.

These are the resources whose provenance information we want to represent.

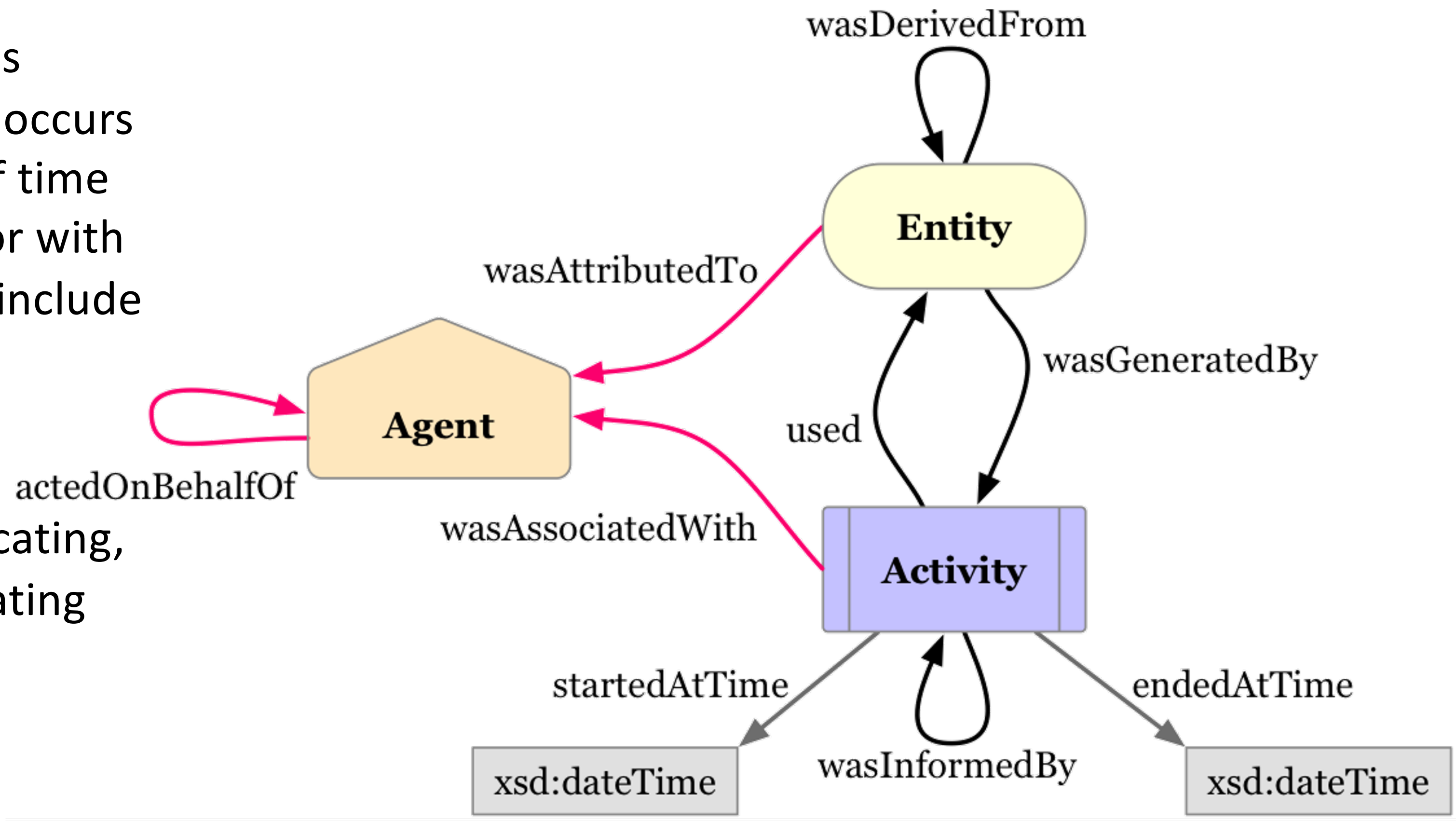


Core concepts and relations in PROV-O from, Copyright © 2011-2013 W3C® (MIT, ERCIM, Keio, Beihang).

PROV-O Core Concepts and Relations

`Namespace prov: <http://www.w3.org/ns/prov#> .`

A prov: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.

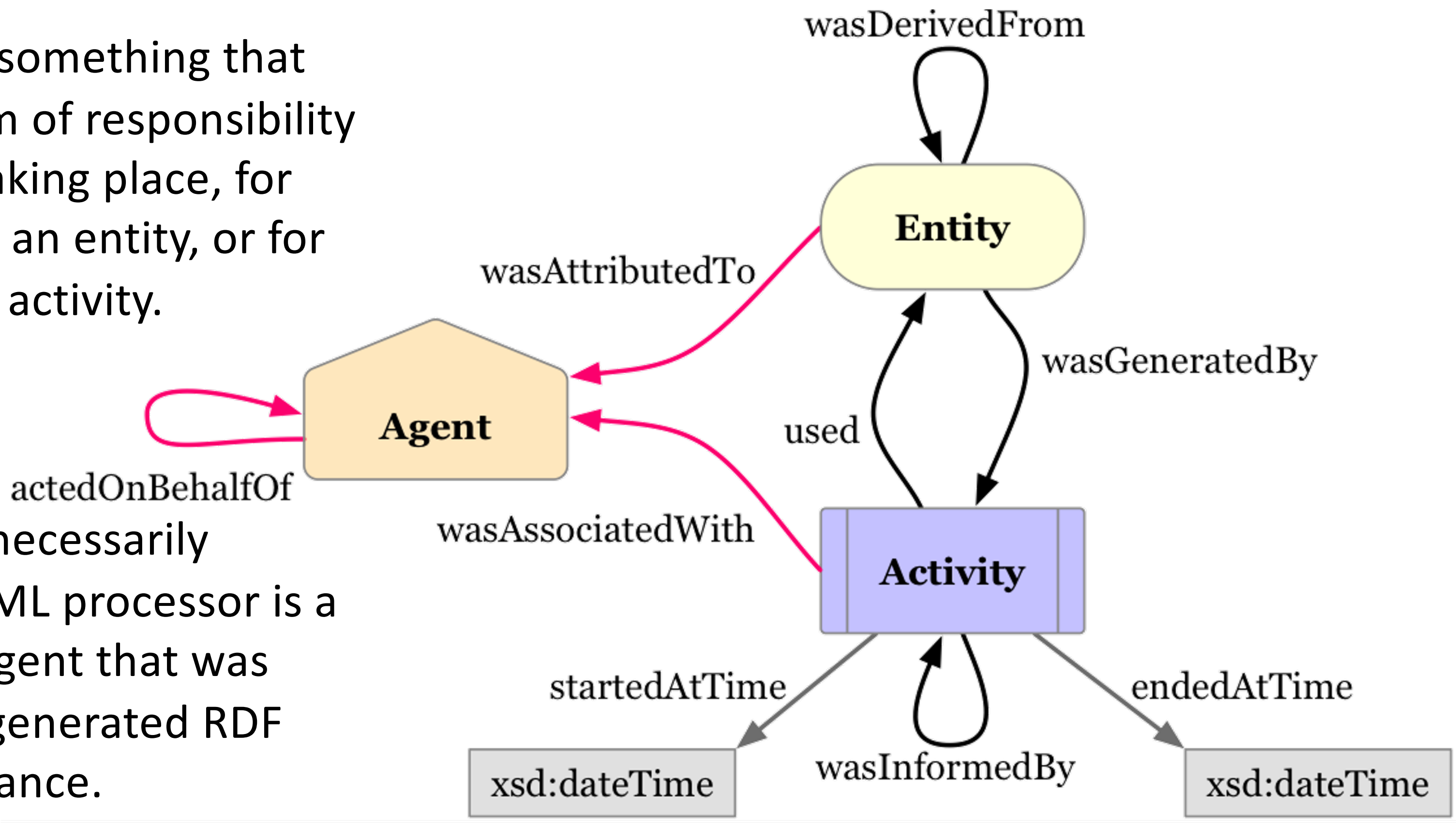


Core concepts and relations in PROV-O from, Copyright © 2011-2013 W3C® (MIT, ERCIM, Keio, Beihang).

PROV-O Core Concepts and Relations

`Namespace prov: <http://www.w3.org/ns/prov#> .`

A prov: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.



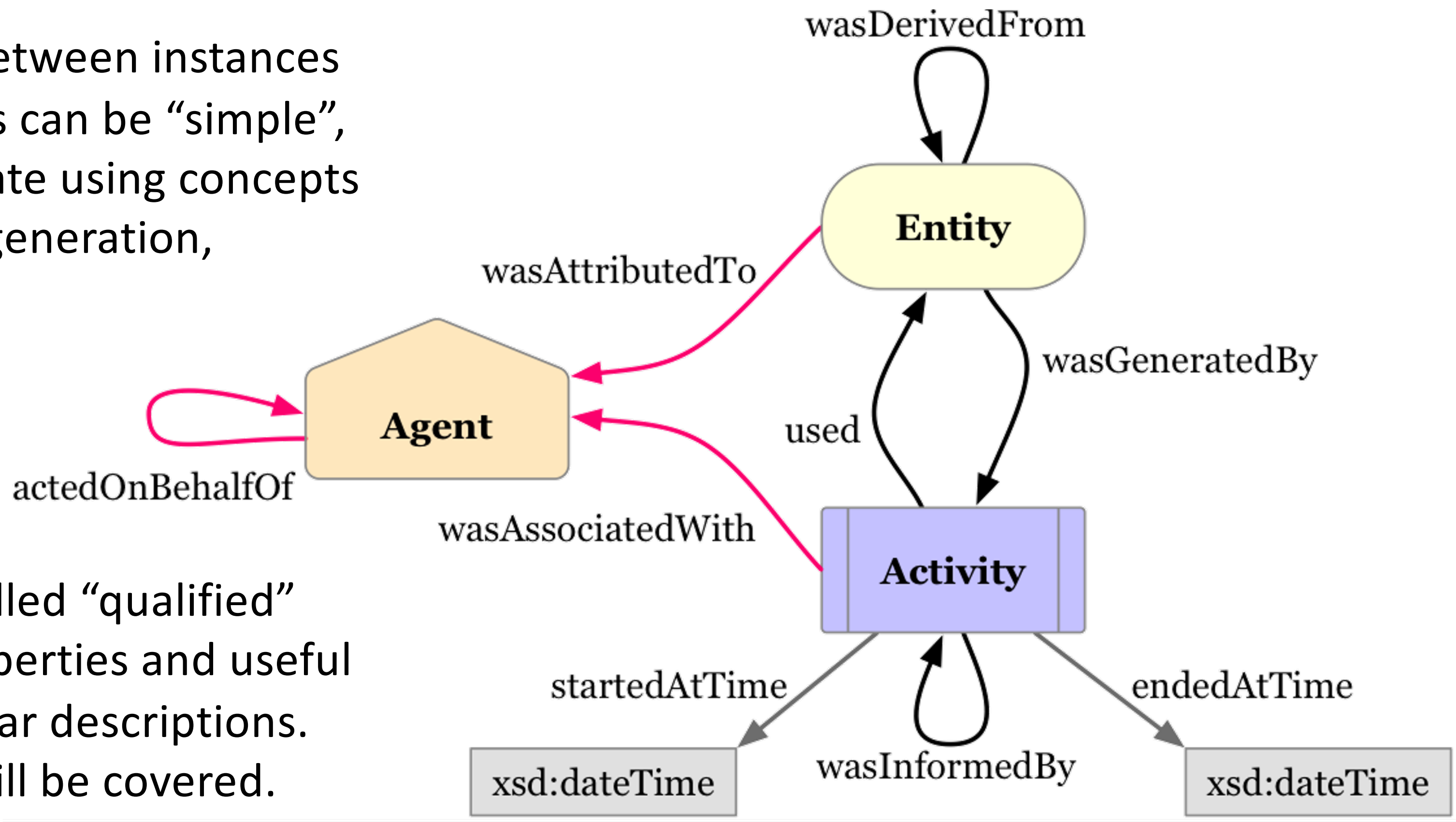
Agents are not necessarily human; an R2RML processor is a prov:SoftwareAgent that was attributed to a generated RDF dataset, for instance.

Core concepts and relations in PROV-O from, Copyright © 2011-2013 W3C® (MIT, ERCIM, Keio, Beihang).

PROV-O Core Concepts and Relations

```
Namespace prov: <http://www.w3.org/ns/prov#> .
```

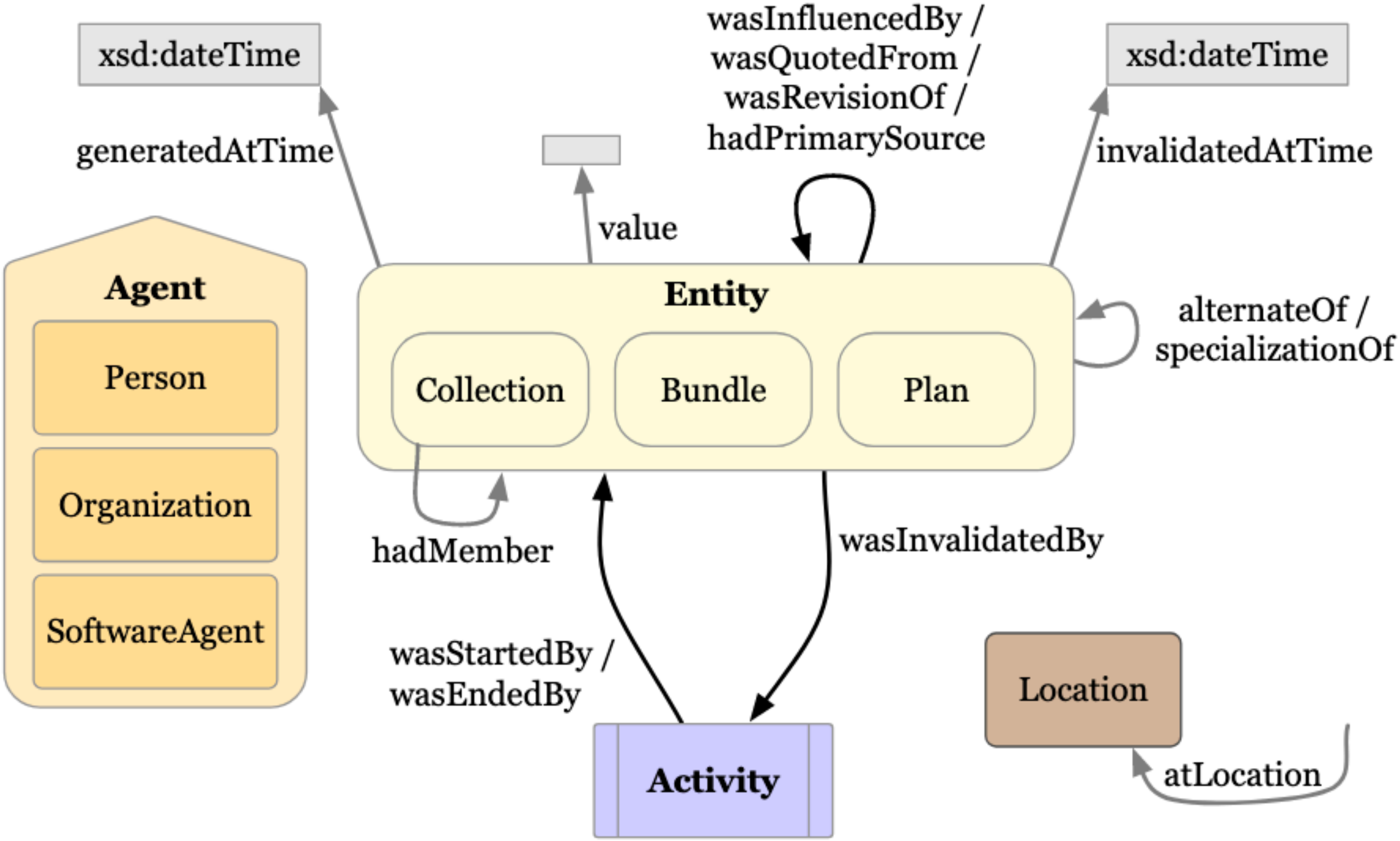
Relationships between instances of these entities can be "simple", or more elaborate using concepts such as usage, generation, attribution, etc.



These are so-called "qualified" classes and properties and useful for more granular descriptions. One example will be covered.

Core concepts and relations in PROV-O from, Copyright © 2011-2013 W3C® (MIT, ERCIM, Keio, Beihang).

PROV-O Core Concepts and Relations



Copyright © 2011-2013 W3C® (MIT, ERCIM, Keio, Beihang).

PROV-O

- PROV-O seems (is?) convoluted, but why? Let’s start with a simple example using Dublin Core, which is a simple vocabulary providing you with terms for describing resources.

```
ex:mypaper
  dct:title "GConsent - A Consent Ontology Based on the GDPR" ;
  dct:creator ex:harsh, ex:christophe, ex:declan, ex:dave ;
  dct:created "2019-05-27"^^xsd:date ;
  dct:publisher ex:springer ;
  dct:issued "2019-05-29"^^xsd:date ;
  dct:replaces ex:mypapercameraready .
```

- Who created the paper? How long did the process take? How was the camera ready paper used?
- PROV-O allows for a more fine-grained, and extensible representation of such information.

PROV-O

- The Universe of Discourse of provenance information is quite complex. You have activities, actors, revisions, roles, etc.
- While more granular than other vocabularies, PROV-O did aim to provide a range of granularity: from ”simple” statements using the concepts and relations to more complex and truthful representations using qualified relations.

PROV-O

```
ex:mypaper
  dct:title "GConsent - A Consent Ontology Based on the GDPR" ;
  dct:creator ex:harsh, ex:christophe, ex:declan, ex:dave ;
  dct:created "2019-05-27"^^xsd:date ;
  dct:publisher ex:springer ;
  dct:issued "2019-05-29"^^xsd:date ;
  dct:replaces ex:mypapercameraready .
```

Let’s assume that Springer uses the camera ready, submitted on the 18th of March for the creation of the publication. That process takes a couple of weeks and ends on the 29th of May. The actual publication, however, was generated on the 27th of May.

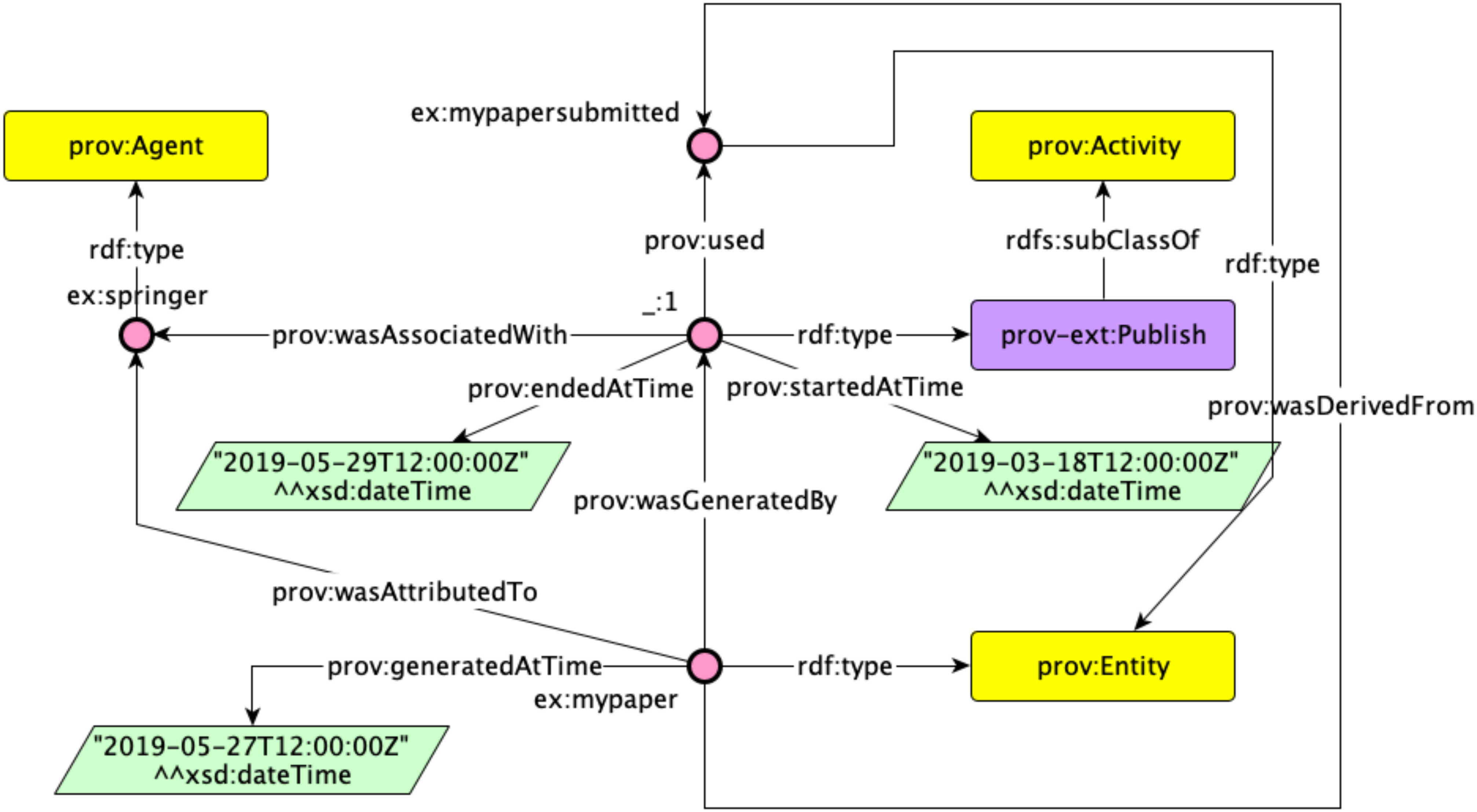
PROV-O

```
ex:mypaper
  dct:title "GConsent - A Consent Ontology Based on the GDPR" ;
  dct:creator ex:harsh, ex:christophe, ex:declan, ex:dave ;
  dct:created "2019-05-27"^^xsd:date ;
  dct:publisher ex:springer ;
  dct:issued "2019-05-29"^^xsd:date ;
  dct:replaces ex:mypapercameraready .
```

Let’s assume that

- Springer uses the camera ready, submitted on the 18th of March for the creation of the publication.
- That the publication process takes a couple of weeks, starts on the 18th of March, and ends on the 29th of May.
- The actual publication, however, was generated on the 27th of May.

Note that “Publish” is an activity that does not provided by PROV-O, so we had to declare our own class (in our own namespace).



PROV-O

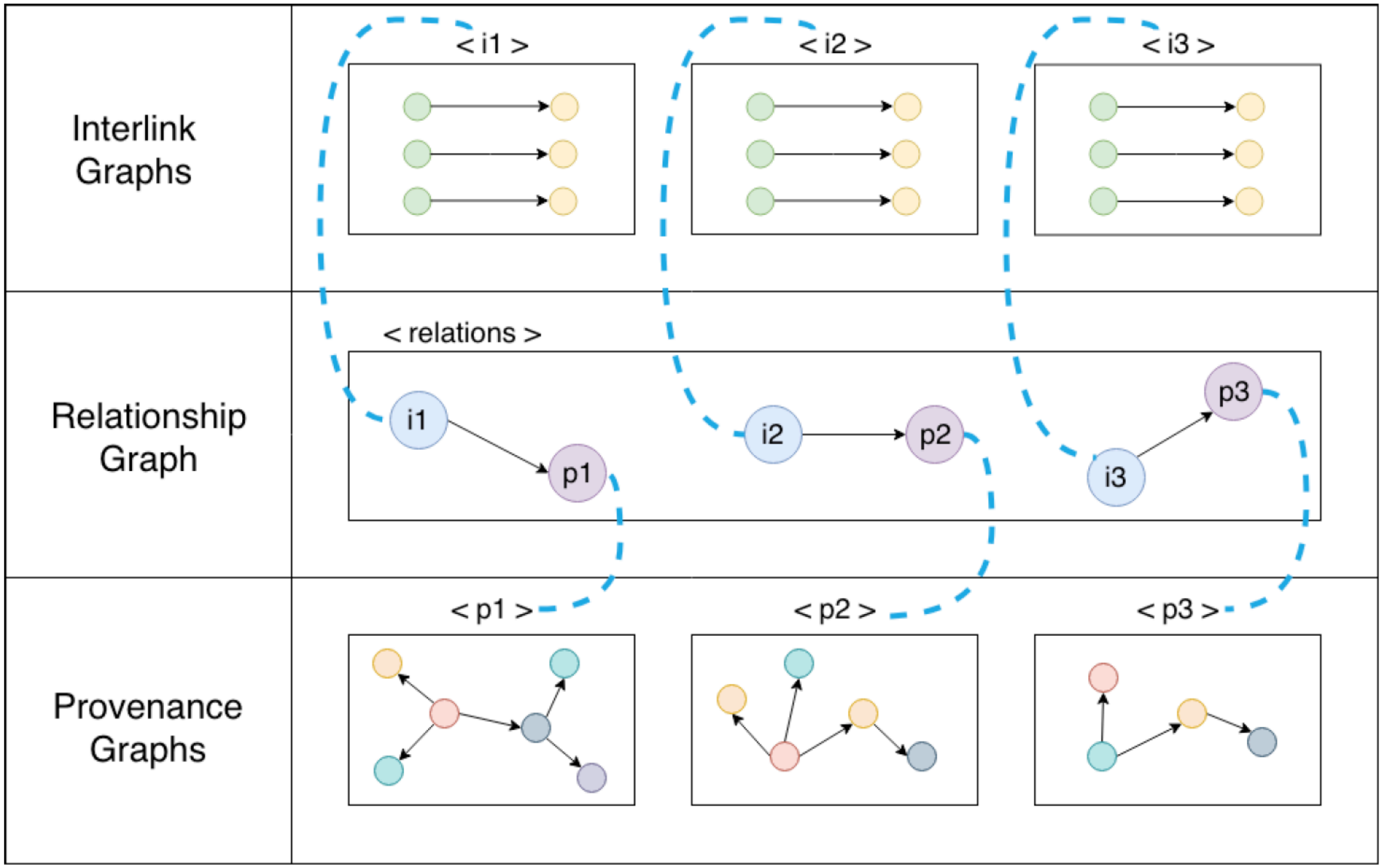
PROV-O provides you with an abstract framework suitable for extension. PROV-O also recognized the need to qualify relations. We turn relationships into instances of classes. PROV-O provides predicates to avoid reification. For instance:

```
ex:mypaper prov:wasAttributedTo ex:christophe .
```

How do state that Christophe is the second author?

```
ex:mypapercameraready
  prov:qualifiedAttribution [
    a prov:Attribution ;
    prov:agent ex:christophe ;
    prov:hadRole ex:author ;
    ex:order "2"^^xsd:integer ;
  ] .
```

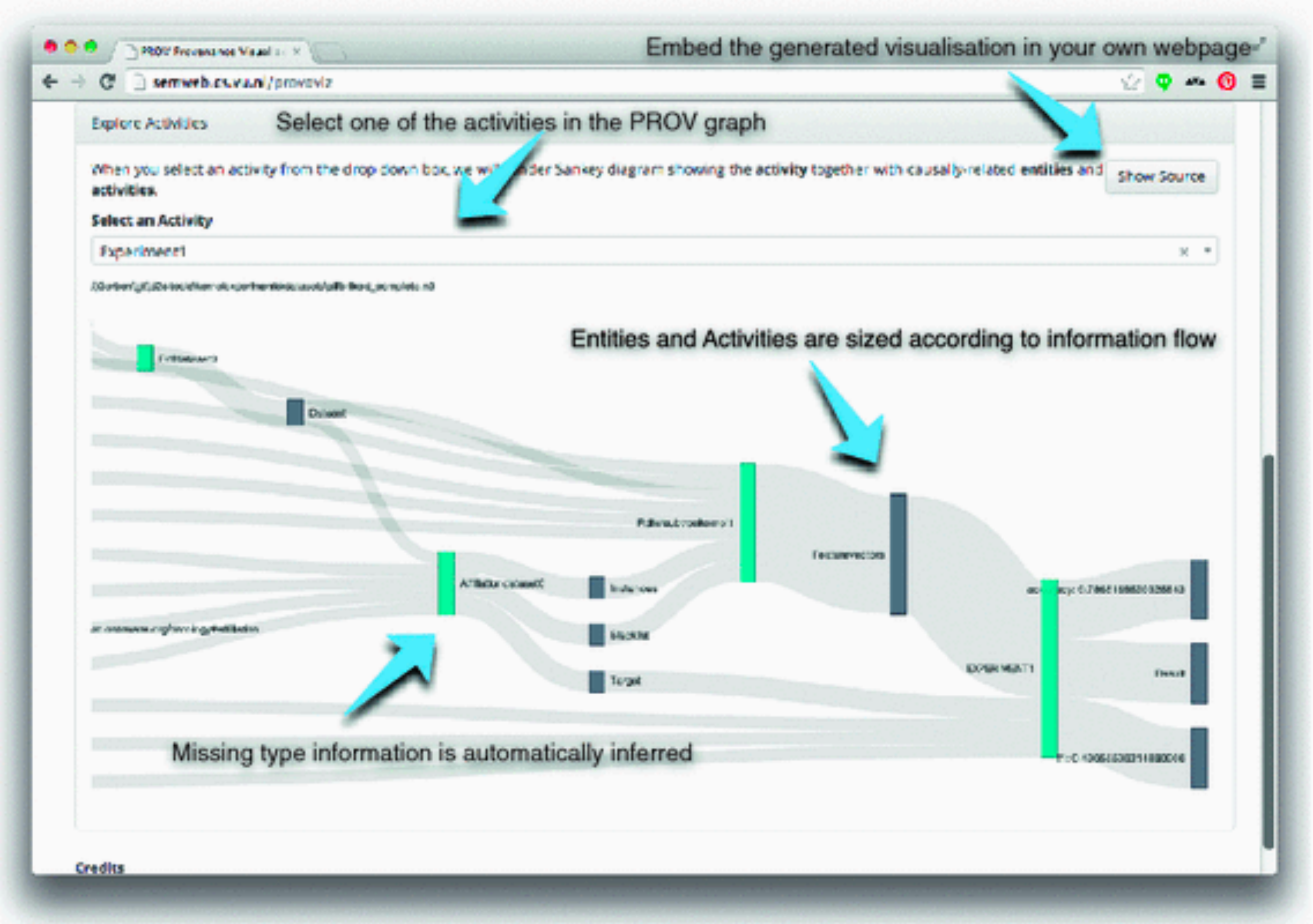
PROV-O in the wild: NAISC



Capturing the provenance of Linked Data interlinks created by librarians and archivists. Using named graphs to bundle the links with their bundles.

L. McKenna, C. Debruyne, D. O'Sullivan: NAISC: An Authoritative Linked Data Interlinking Approach for the Library Domain. JCDL 2019: 11-20

Tools: PROV-O-Viz



Embed the generated visualisation in your own webpage

Select one of the activities in the PROV graph

Entities and Activities are sized according to information flow

Missing type information is automatically inferred

semweb.cs.vu.nl/provoviz

Explore Activities

When you select an activity from the drop down box, we will render Sankey diagram showing the **activity** together with causally-related **entities** and **activities**.

Select an Activity

Experiment1

Hoekstra R., Groth P. (2015) PROV-O-Viz - Understanding the Role of Activities in Provenance. In: Ludäscher B., Plale B. (eds) Provenance and Annotation of Data and Processes. IPAW 2014. Lecture Notes in Computer Science, vol 8628. Springer, Cham

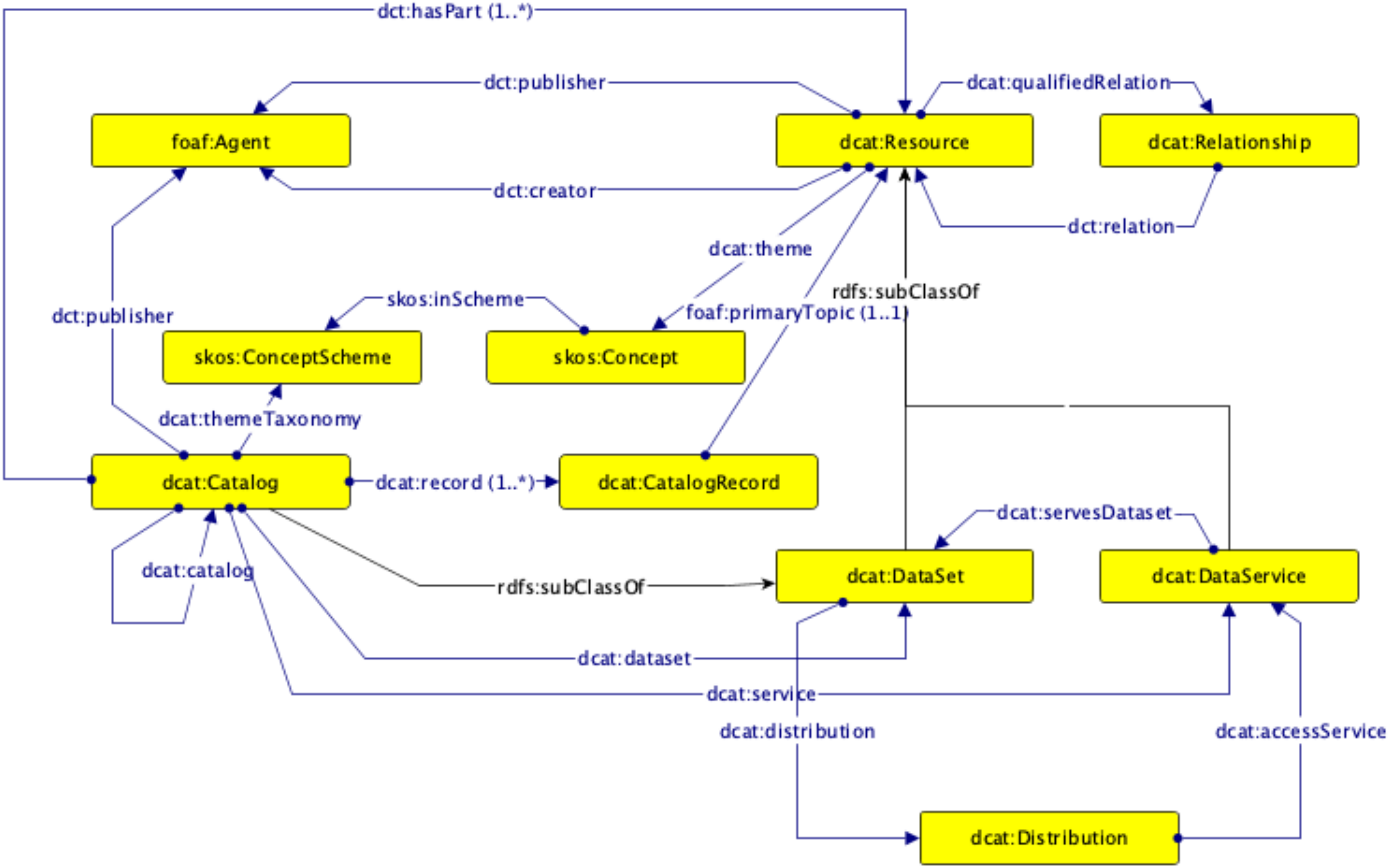
DCAT

The Data Catalog Vocabulary (DCAT) is a W3C Recommendation published in 2014 and a revision is currently a Proposed Recommendation.

DCAT is a vocabulary for describing and exchanging information about data catalogs.

DCAT is used to describe, in RDF, any dataset and how they are distributed. Datasets can exist on their own, or be part of a catalog.

Allows you to query catalogs for their formats, metadata, contents, and even... provenance ☺


dct:hasPart (1..*)
dct:publisher
dcat:qualifiedRelation
foaf:Agent
dcat:Resource
dcat:Relationship
dct:creator
dct:relation
dcat:theme
rdfs:subClassOf
skos:inScheme
dct:publisher
foaf:primaryTopic (1..1)
skos:ConceptScheme
skos:Concept
dcat:themeTaxonomy
dcat:Catalog
dcat:record (1..*)
dcat:CatalogRecord
dcat:servesDataset
dcat:catalog
dcat:DataSet
dcat:DataService
rdfs:subClassOf
dcat:dataset
dcat:service
dcat:distribution
dcat:accessService
dcat:Distribution

DCAT

- dcat:Dataset: “A collection of data, published or curated by a single agent, and available for access or download in one or more representations.”
- dcat:Distribution: “A specific representation of a dataset. A dataset might be available in multiple serializations that may differ in various ways […].”
- The same dataset can thus have multiple distributions.

DCAT

Notice that DCAT prescribes the use of other vocabularies (SKOS, DCT, FOAF,…), but one can (or sometimes needs) to extend the vocabulary for their own purposes.

“Complementary vocabularies can be used together with DCAT to provide more detailed format-specific information. For example, properties from the VoID vocabulary can be used within DCAT to express various statistics about a dataset if that dataset is in RDF format.”

DCAT

```
:ds1 a dcat:Dataset ;
  dcat:distribution :dist1, :dist2 .

:dist1 a dcat:Download ;
  dcat:accessURL  <http://example.org/file.csv>;
  dcat:format [ rdfs:label "CSV" ].

:dist2 a dcat:Distribution ;
  dcat:accessURL <http://example.org/index.html> .
  dct:format [ rdfs:label "CSV" ].
```

VoID

Describing Linked Datasets with the VoID Vocabulary: https://www.w3.org/TR/void/

Describing metadata about RDF datasets. This metadata includes predicates for describing their statistics. Not (yet?) a W3C Recommendation, but important to be part of the Linked Data Web.

VoID

Describing Linked Datasets with the VoID Vocabulary: https://www.w3.org/TR/void/

Describing metadata about RDF datasets. This metadata includes

- Classical metadata: author, creators, license,…
- Access mechanisms: SPARQL endpoint, data dumps,…
- Structural information: partitions and number

VoID thus provides predicates for describing their statistics. Not only does it allow you to describe datasets, but also partitions of datasets, and links between RDF datasets (called Link Sets).

Not (yet?) a W3C Recommendation, but important to be part of the Linked Data Web.

VoID Example inspired by data.geohive.ie

```
:geohive a void:Dataset ;
  foaf:homepage <http://data.geohive.ie/> ;
  dcterms:title "GeoHive" ;
  dcterms:source <http://www.osi.ie/> ;
  dcterms:license
     <https://creativecommons.org/licenses/by/4.0> ;
  dcterms:modified "2016-11-17"^^xsd:date ;
  # void:sparqlEndpoint <http://.../sparql> ;
  void:subset :counties ;
.
```

VoID Example inspired by data.geohive.ie

```
:counties a void:Dataset;
  dcterms:title "County Boundary Dataset" ;
  void:dataDump <http://.../boundary/county_20M.n3> ;
  # Number of geo:Feature in the dataset
  void:classPartition [
    void:class geo:Feature ;
    void:entities 26 ;
  ];
  # Number of rdfs:label assertions in the dataset
  void:propertyPartition [
    void:property rdfs:label ;
    void:triples 78 ;
  ];
.
```

VoID Example inspired by data.geohive.ie

```
# Describing linksets
# void:subjectsTarget and void:objectsTarget
# are both subproperties of void:target
:geohive_dbpedia a void:Linkset ;
    # the subject are from data.geohive.ie
    void:subjectsTarget :counties ;
    # and point to resources in DBpedia
    void:objectsTarget :dbpedia ;
    void:subset :counties ;
    void:linkPredicate voc:similarTo ;
    void:triples 26 ;
    .
```

Conclusions

- In this lecture we covered vocabularies for provenance (PROV-O) and dataset descriptions (DCAT and VoID).

- DCAT is a W3C Recommendation and meant for any dataset. VoID is for RDF datasets, but not yet standardized. Both DCAT and VoID can avail of other vocabularies to add additional information (including provenance information).

References

1. J. Zhao and O. Hartig. Towards interoperable provenance publication on the linked data web. In C. Bizer, T. Heath, T. Berners-Lee, and M. Hausenblas, eds,, WWW2012 Workshop on Linked Data on the Web, Lyon, France, 16 April, 2012, volume 937 of CEUR Workshop Proceedings. CEUR-WS.org, 2012.
2. O. Hartig and J. Zhao. Publishing and consuming provenance metadata on the web of linked data. In D. L. McGuinness, J. Michaelis, and L. Moreau, eds., *Provenance and Annotation of Data and Processes - Third International Provenance and Annotation Workshop, IPAW 2010, Troy, NY, USA, June 15-16, 2010. Revised Selected Papers*, volume 6378 of *Lecture Notes in Computer Science*, pages 78–90. Springer, 2010.