Core Ontological Frameworks

Standing on the shoulders of giants

“Upper” vs. “Domain” vs. “Application” ontologies

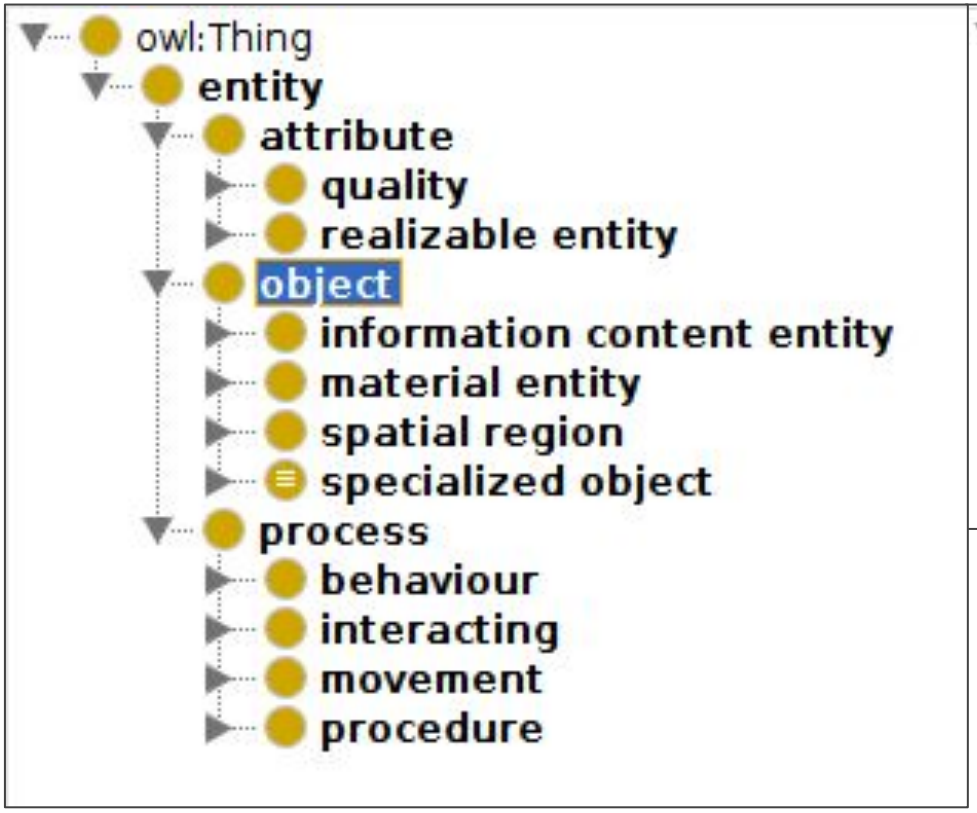
Ontologies fall into three general categories:

Those that model “the nature of existence in general” → Upper ontologies

Those that model the expert knowledge in a domain → Domain ontologies

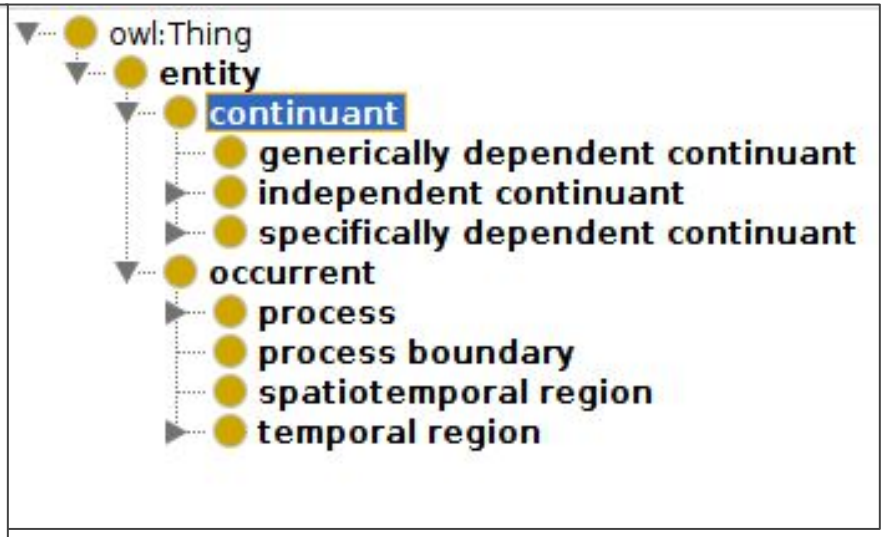
Those that model for a specific purpose → Application ontologies

Upper ontologies



SemanticScience Integrated Ontology (SIO)

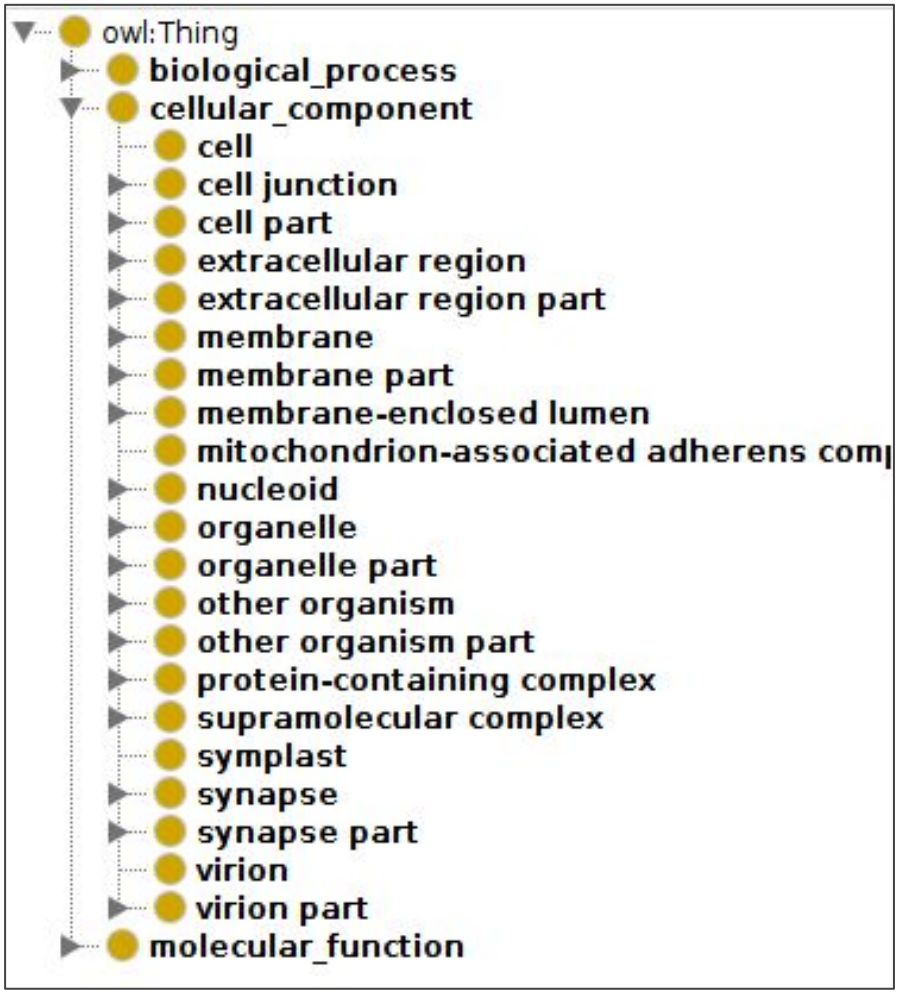
M. Dumontier, lead developer



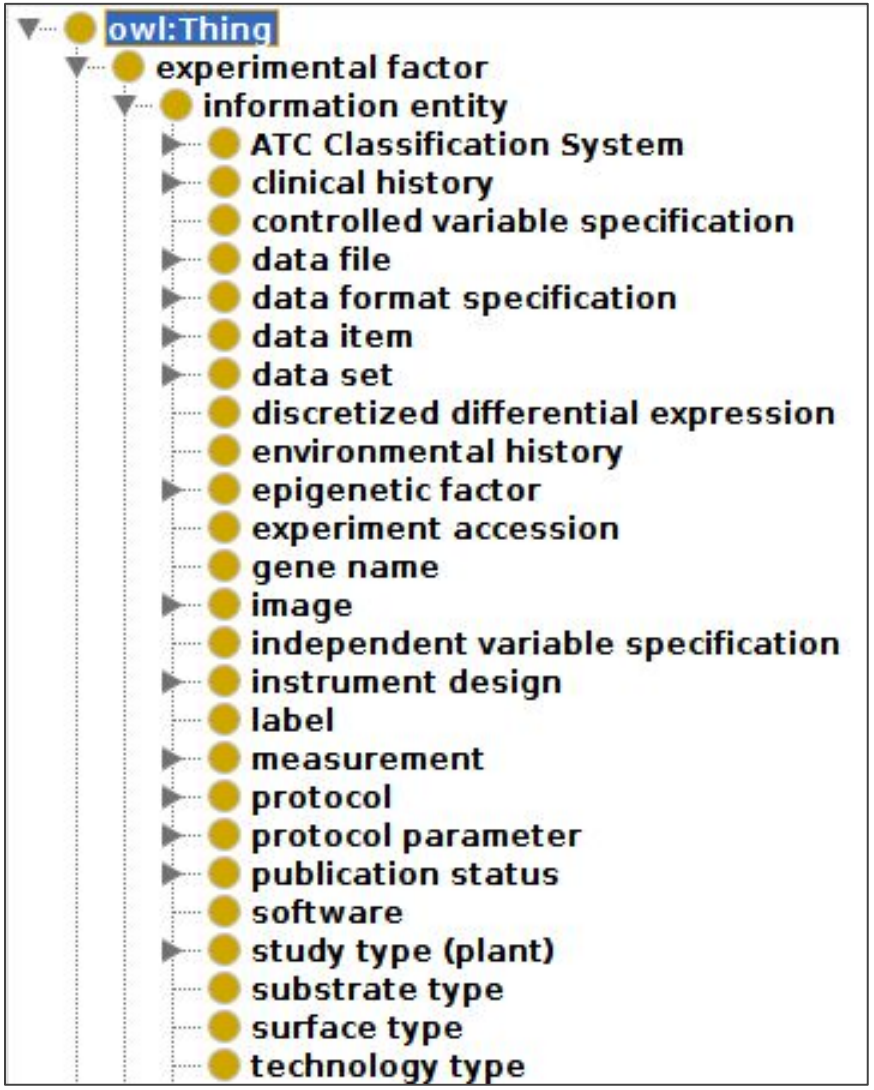
Basic Formal Ontology (BFO)

B. Smith, lead developer

Domain Ontologies

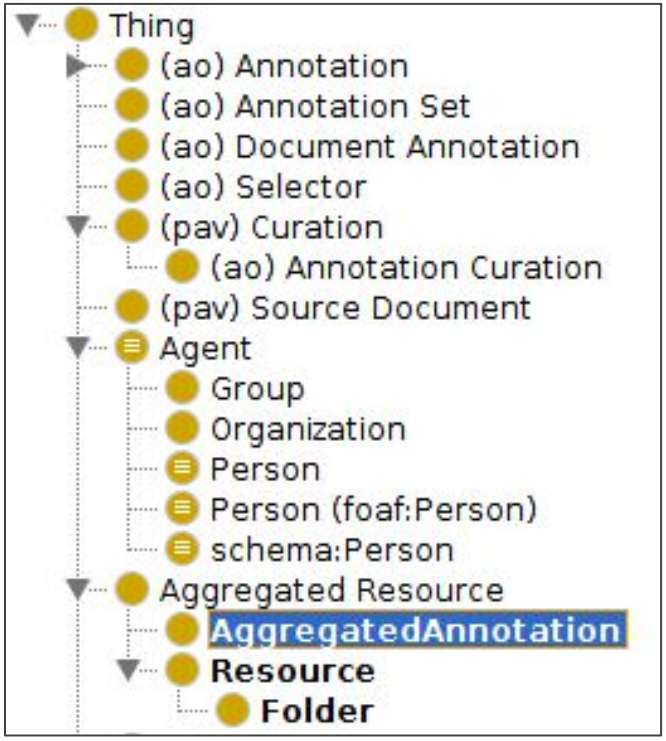


Gene Ontology



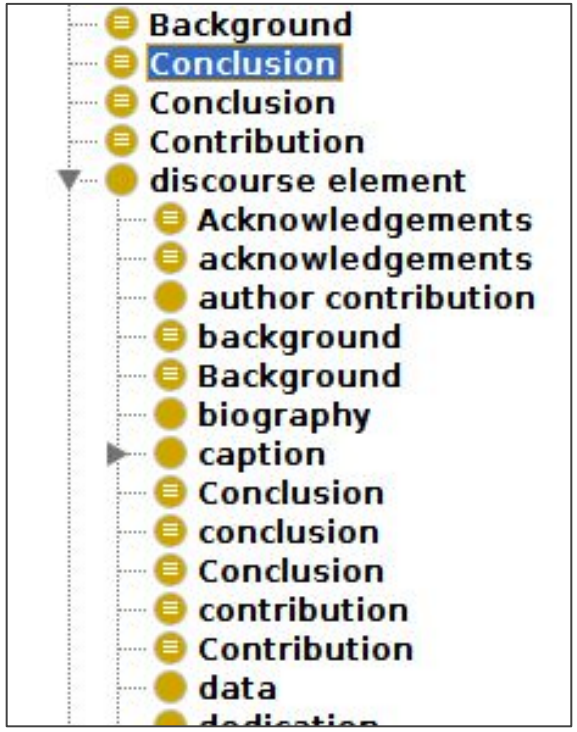
Experimental factor ontology

Application Ontologies



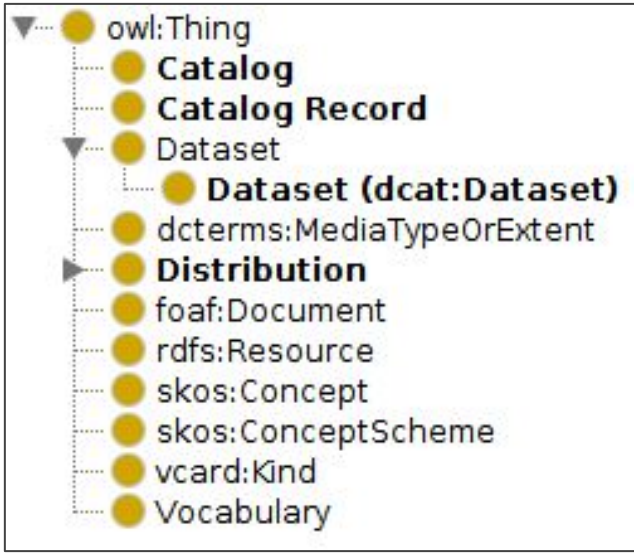
Research Object Ontology

used to support tools that work on aggregations of data + analytical workflows + publications, etc.



Discourse Ontology

used to support tools that can explore scholarly discourse in the literature



Data Catalogue Vocabulary (DCAT)

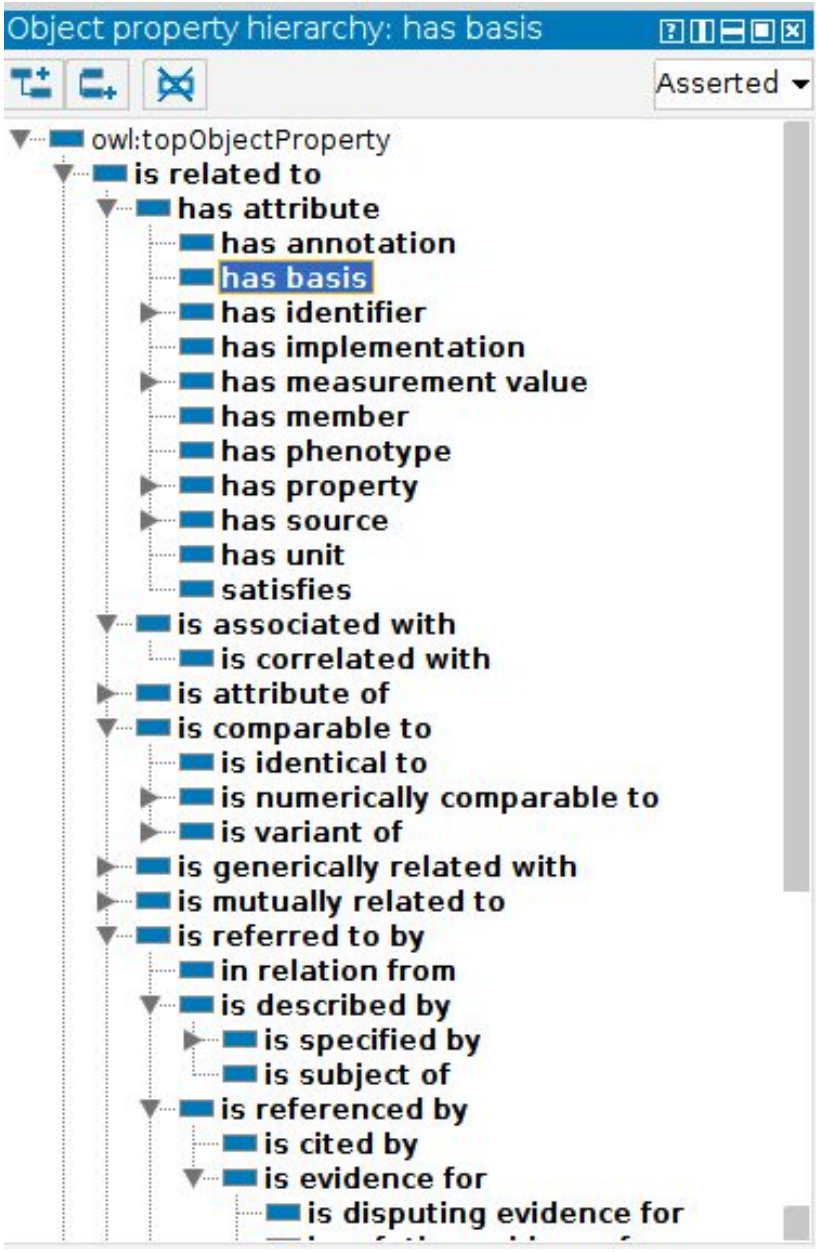
Used to support applications for data curation and cataloguing

Application Ontologies

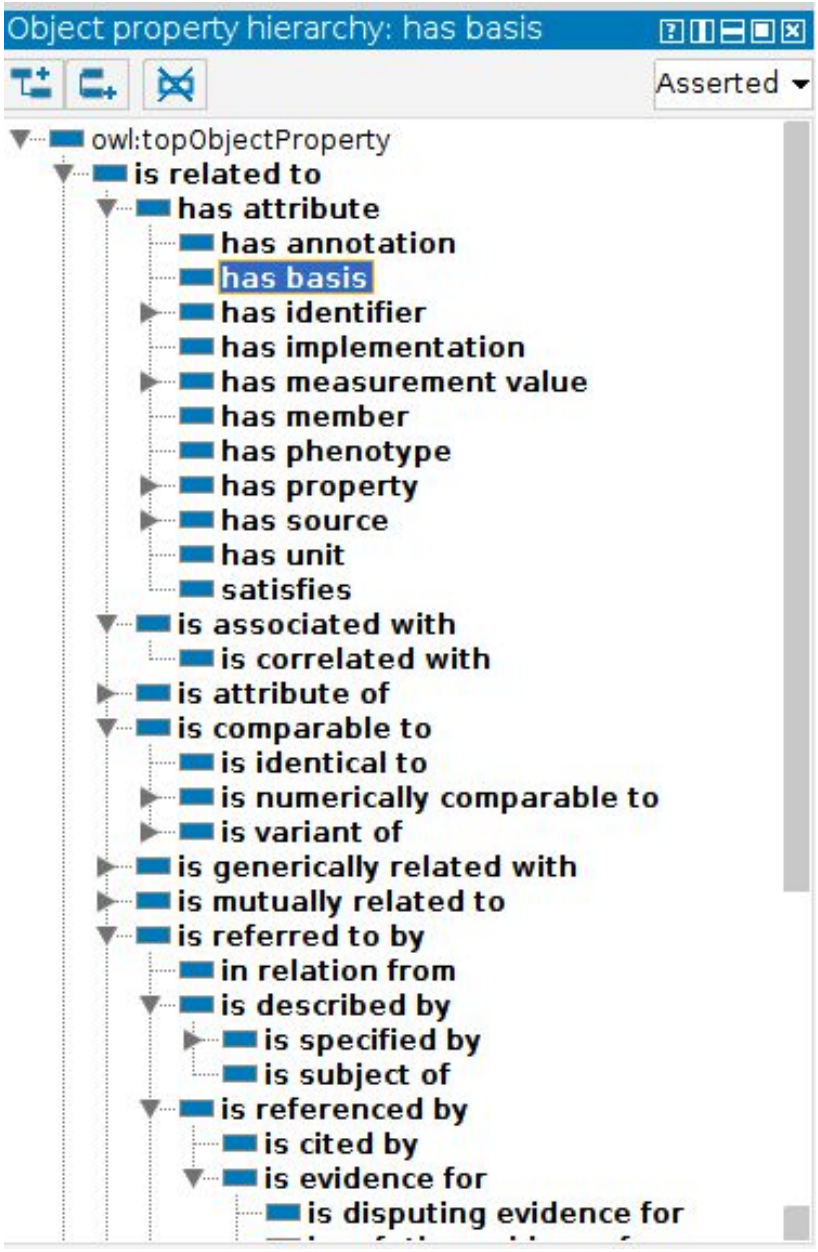
Provenance Ontology

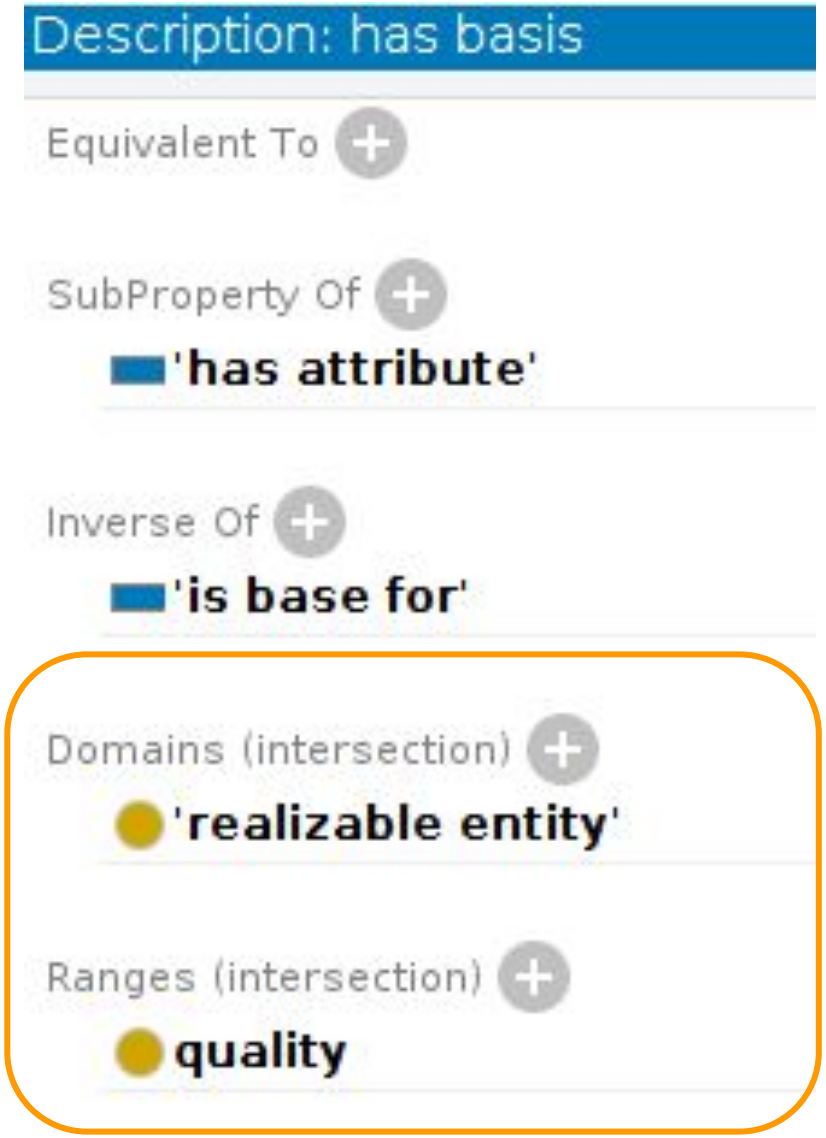
Used to annotate an activity (like the running of a workflow)

Properties, domains, and ranges

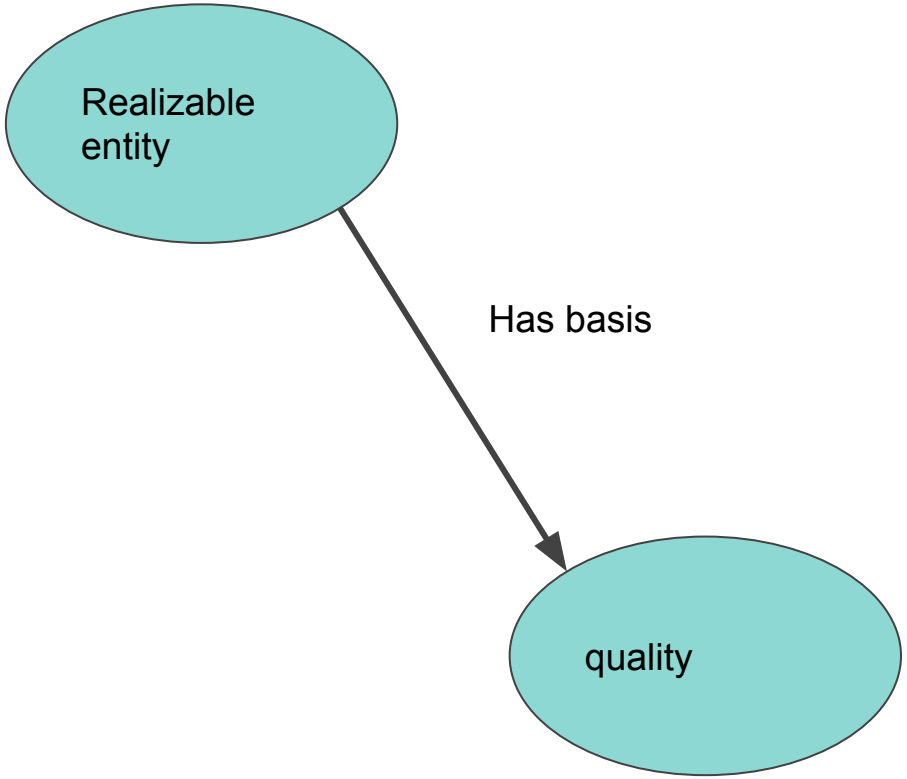


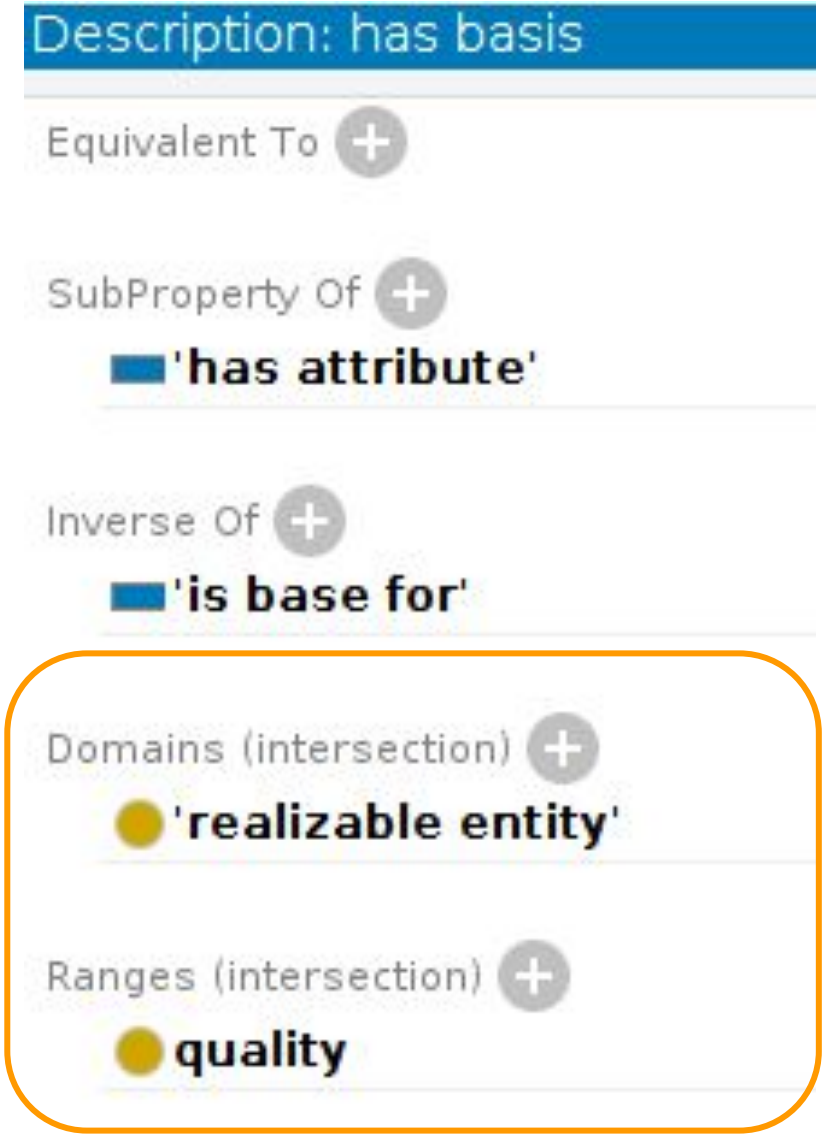
Properties, domains, and ranges





Properties, domains, and ranges





Diagrams from the Entities and Relationships

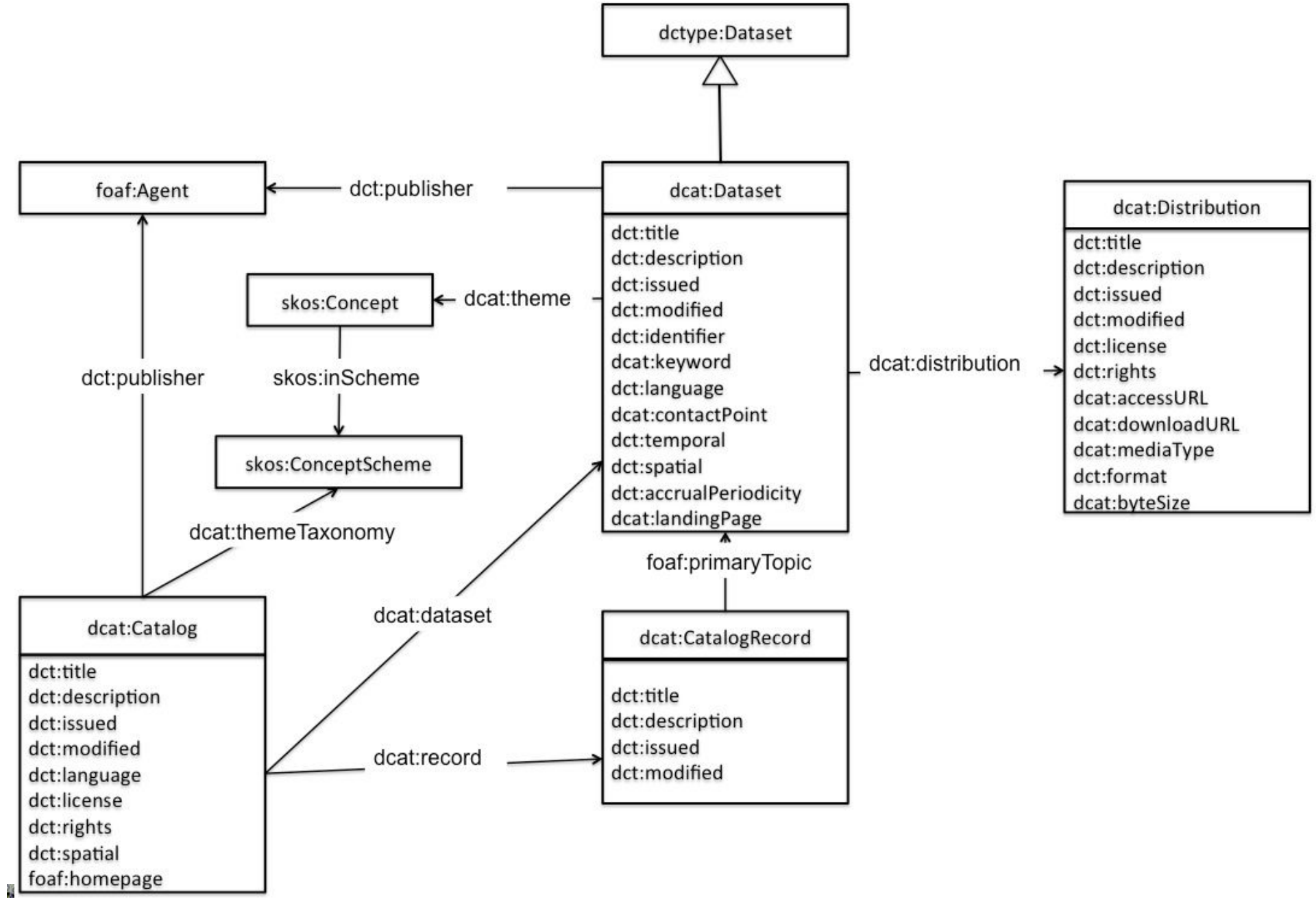
Many ontologies place restrictions on what kinds of entities can be connected by what kinds of properties, and in what direction

For example: Person → hasMedicalRecord → Medical Record

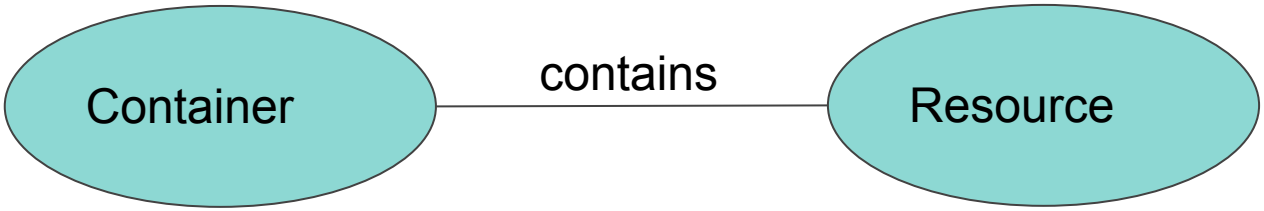
Not: Medical Record → hasMedicalRecord → Person

This allows us to make diagrams that reveal the overall structure of the idea that the ontology is trying to model. Here are some examples:

DCAT

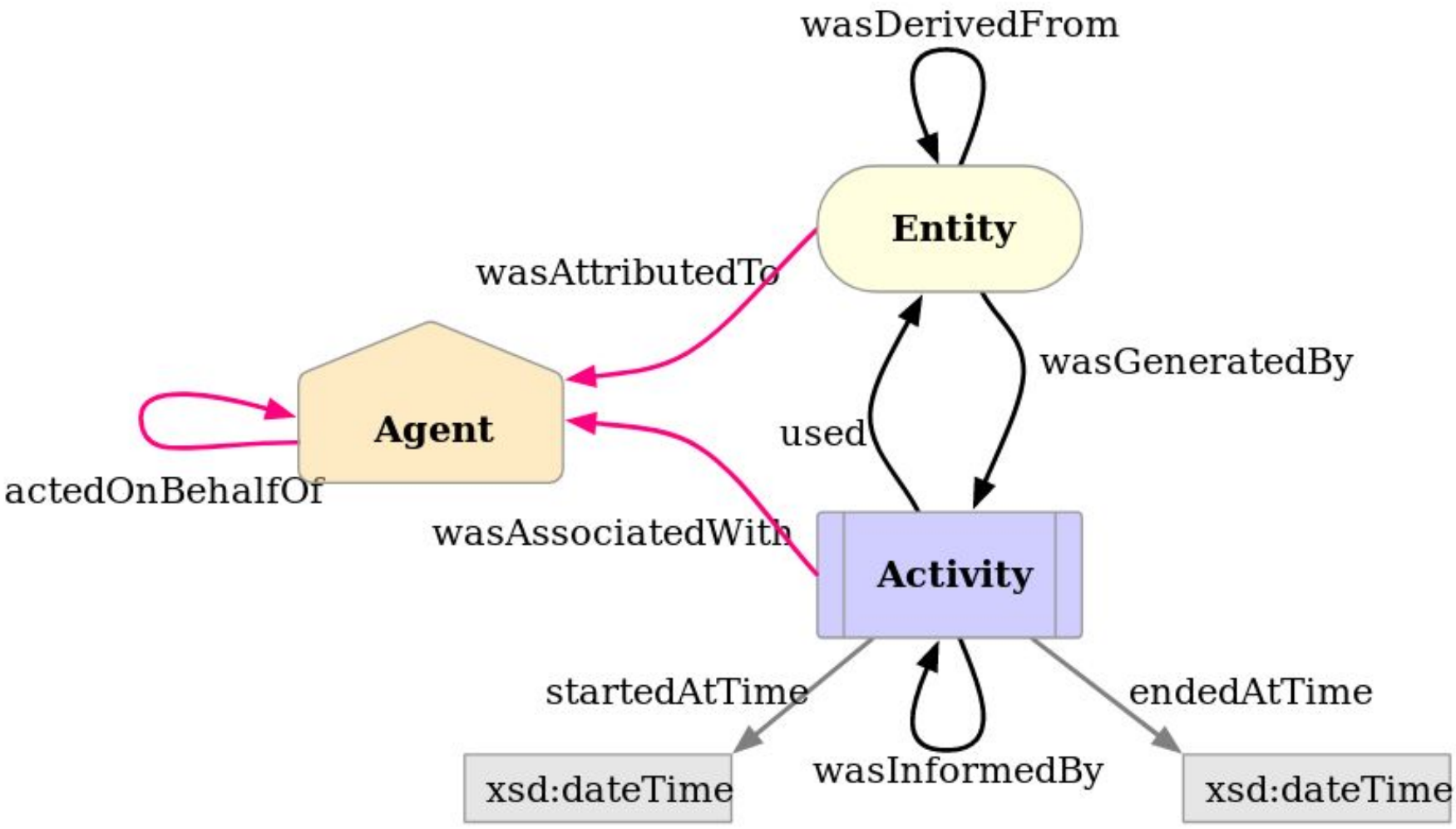

dctype:Dataset
foaf:Agent
dct:publisher
dcat:Dataset
dct:title
dct:description
dct:issued
dct:modified
dct:identifier
dcat:keyword
dct:language
dcat:contactPoint
dct:temporal
dct:spatial
dct:accrualPeriodicity
dcat:landingPage
dcat:Distribution
dct:title
dct:description
dct:issued
dct:modified
dct:license
dct:rights
dcat:accessURL
dcat:downloadURL
dcat:mediaType
dct:format
dcat:byteSize
skos:Concept
dcat:theme
dct:publisher
skos:inScheme
dcat:distribution
skos:ConceptScheme
dcat:themeTaxonomy
foaf:primaryTopic
dcat:Catalog
dct:title
dct:description
dct:issued
dct:modified
dct:language
dct:license
dct:rights
dct:spatial
foaf:homepage
dcat:dataset
dcat:CatalogRecord
dct:title
dct:description
dct:issued
dct:modified
dcat:record

Linked Data Platform

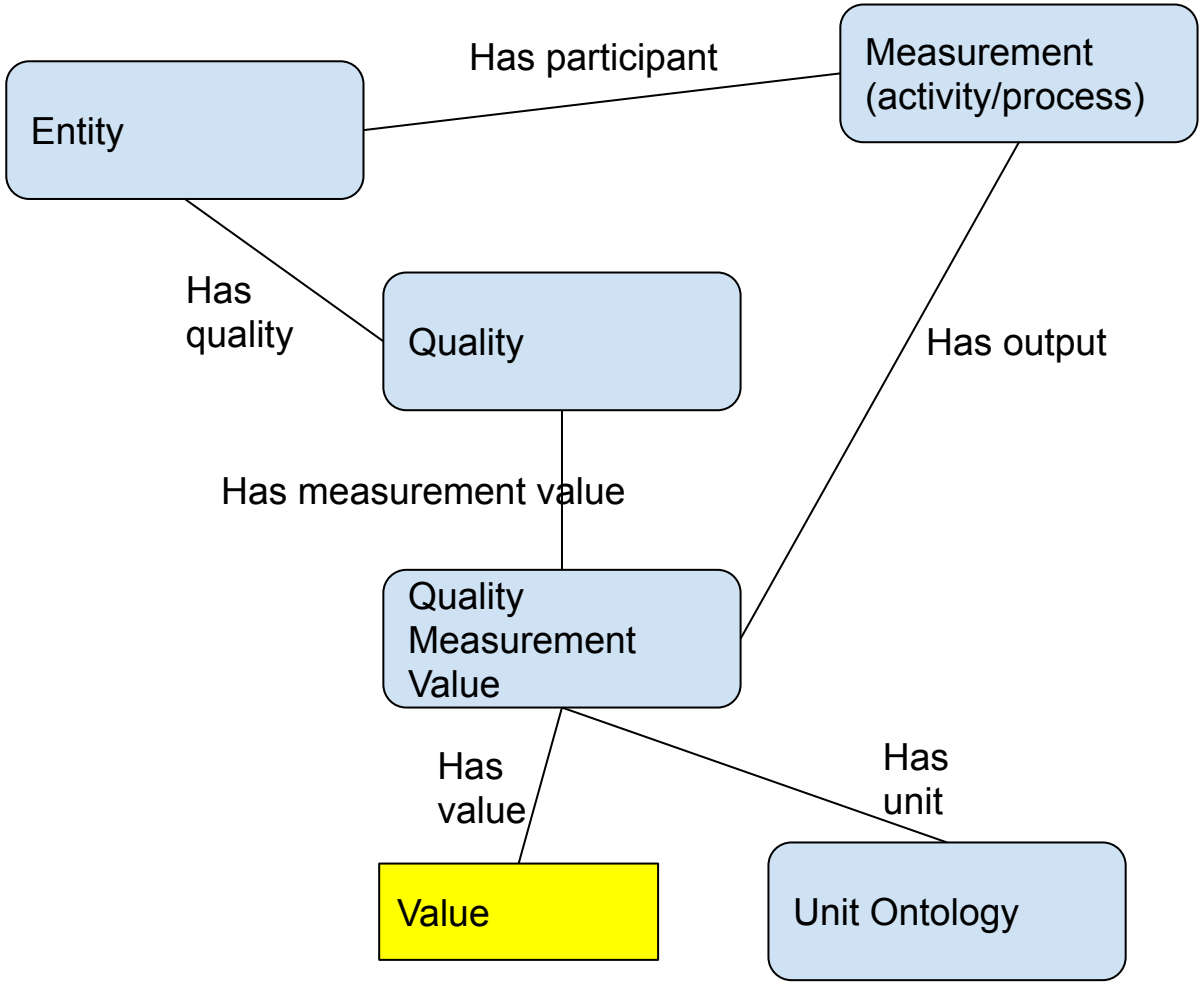


(this is a dramatic over-simplification of LDP! But this is all we need to know for later in this workshop)

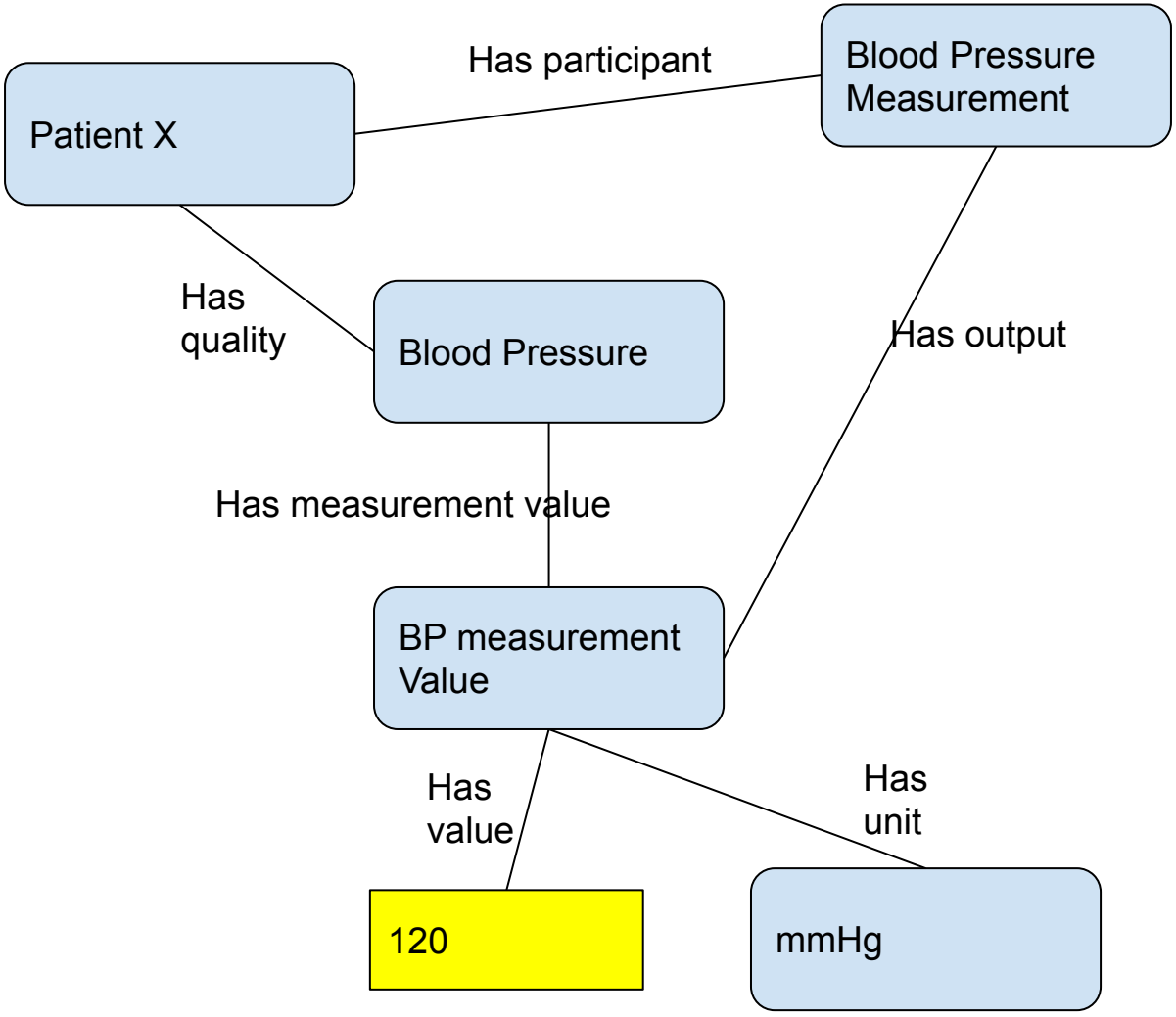
PROV-O


wasDerivedFrom
Entity
wasAttributedTo
wasGeneratedBy
Agent
used
actedOnBehalfOf
wasAssociatedWith
Activity
startedAtTime
endedAtTime
wasInformedBy
xsd:dateTime
xsd:dateTime

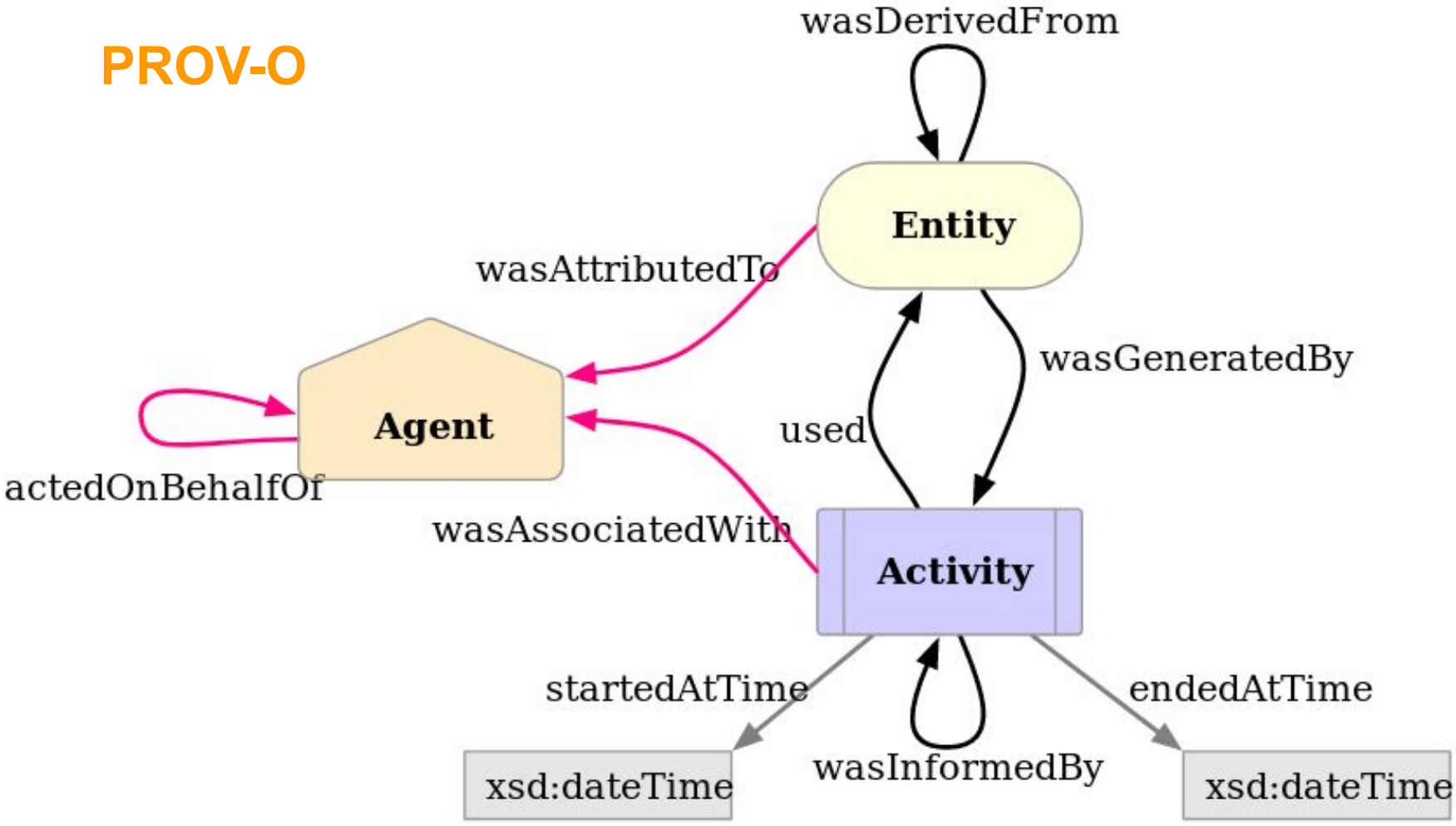
SIO (conceptual model)



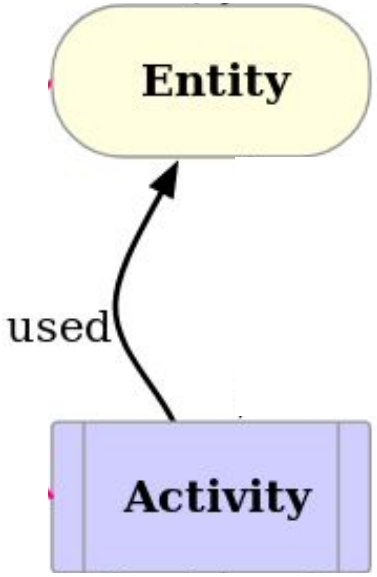
SIO (for example)



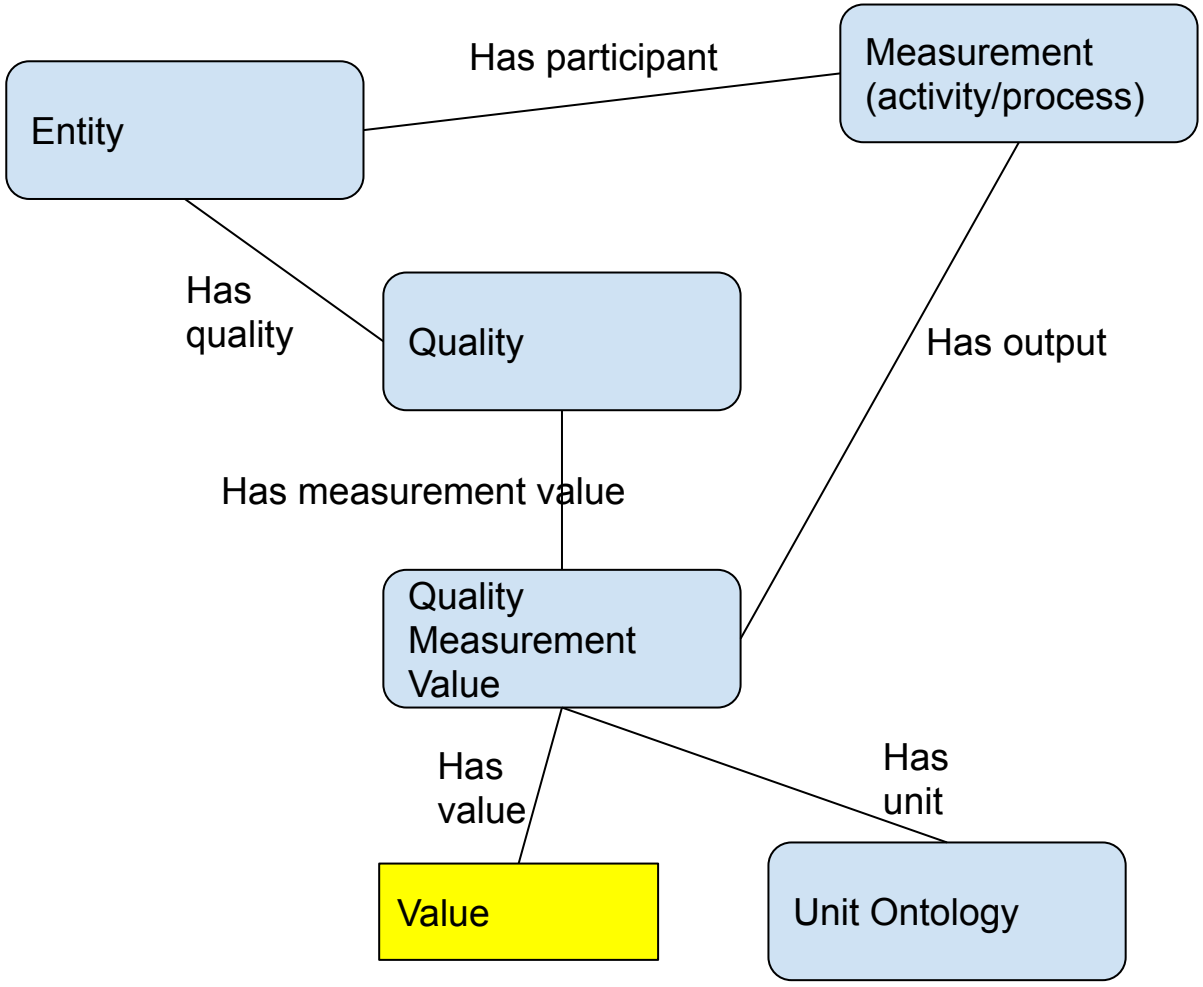
Conceptual overlap between core ontologies



Conceptual overlap between core ontologies

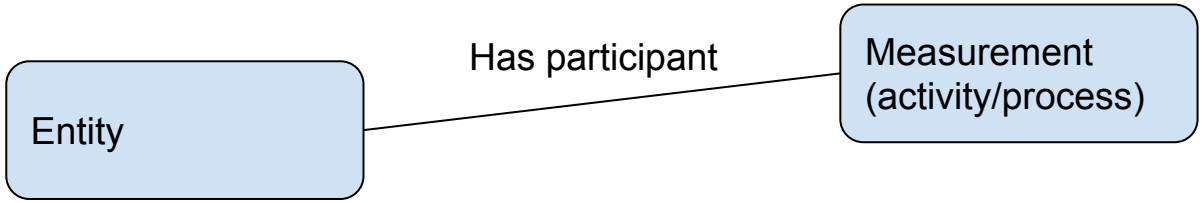


Conceptual overlap between core ontologies

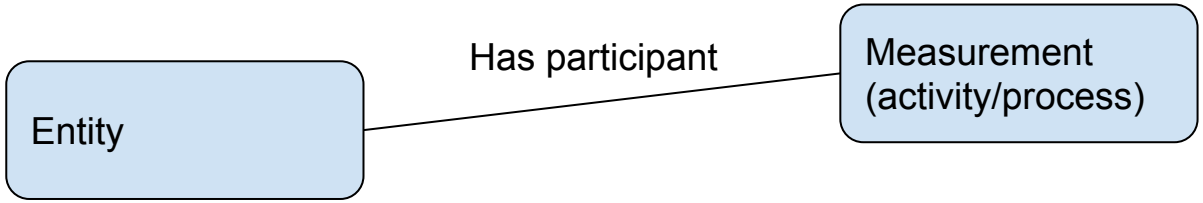


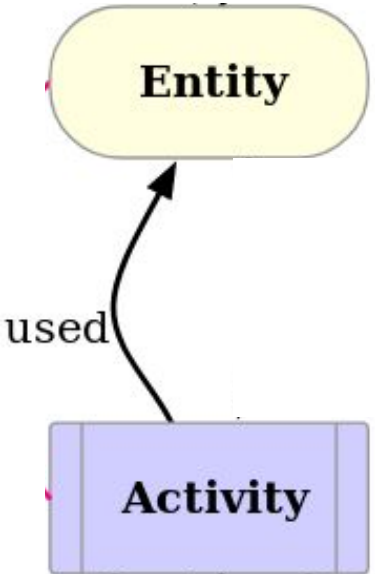
SIO

Conceptual overlap between core ontologies

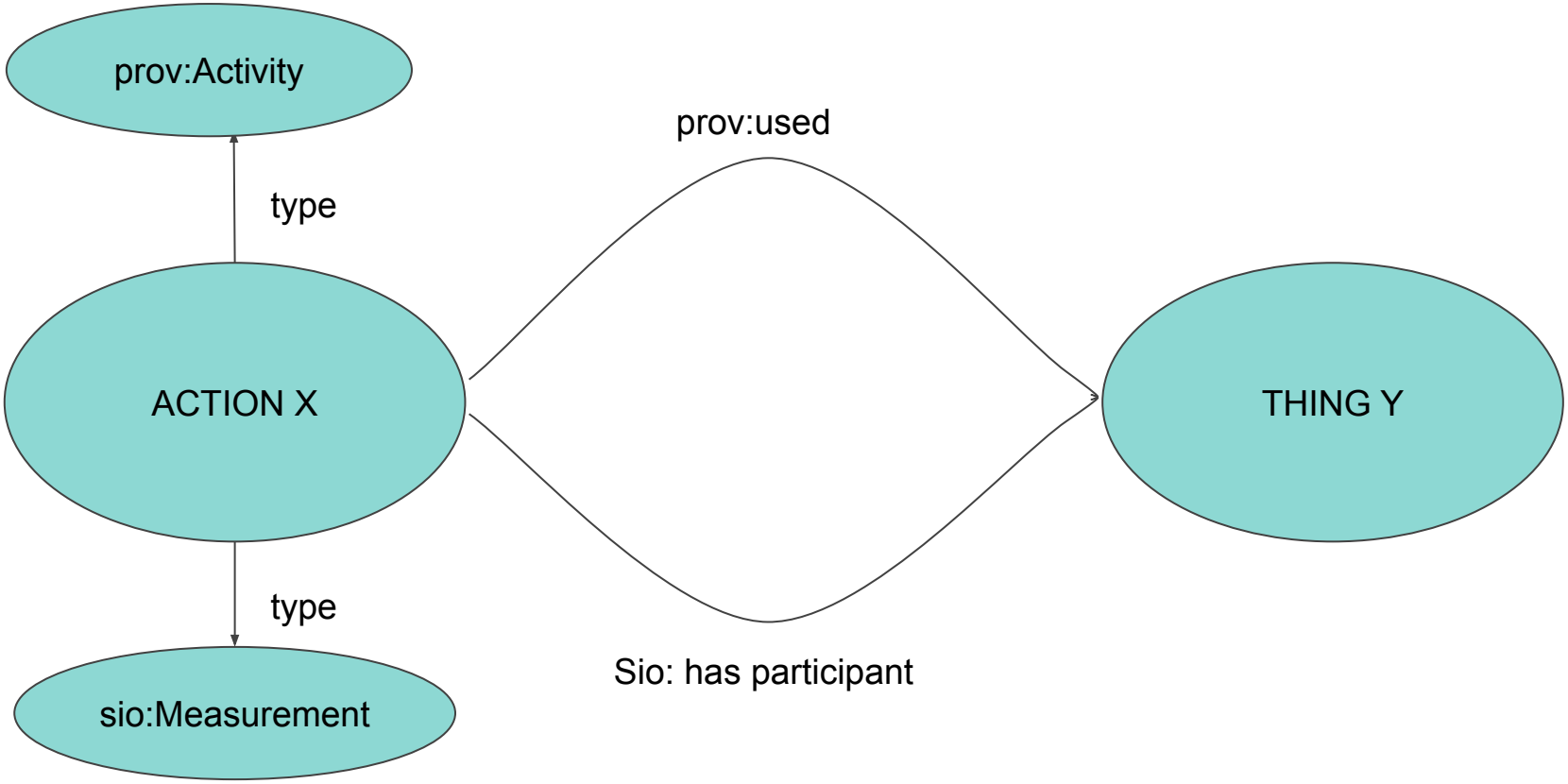


Conceptual overlap between core ontologies





It is often a good idea to use both!



Compatibility with more ontologies makes you more interoperable!

Modelling Sample Data

In the Google Drive folder there are two files with data extracted from:

Riedel, Judith, Romeis, Jörg, & Meissle, Michael. (2016). *Update and expansion of the database of bio-ecological information on non-target arthropod species* [Data set]. Zenodo. http://doi.org/10.5281/zenodo.285525

The dataset contains observations of insects "grazing on" crops over all of Europe over several years.

Looking at the data, we will begin to build a semantic model, **using SIO as our primary core**

SpeciesAbundancePub2015.tsv

Open with Goog

A	B	C	D	E	F
observationID	cropCode	cropName	speciesID	countryCode	countryText
2147365908	A059G	Maize	290307251	ES	Spain
2147365909	A059G	Maize	290307131	ES	Spain
2147365910	A059G	Maize	290307252	ES	Spain
2147365911	A059G	Maize	290307205	ES	Spain
2147365216	A059G	Maize	290307128	ES	Spain

gle Sheets

G	H	I	J	K
startYear	duration	longitude	latitude	collectionRemarks
2000	3	0.36	41.36	Farm scale study. Sp
2000	3	0.36	41.36	Farm scale study. Sp
2000	3	0.36	41.36	Farm scale study. Sp
2000	3	0.36	41.36	Farm scale study. Sp
2002	1	0.5701	41.7955	2 fields: isogenic and

SpeciesInfoPub2015.tsv

	A	B	C
	speciesID	GBIF	species
	290307346	urn:lsid:faunaeur.org	Allygidius commutat
	290308575	urn:lsid:faunaeur.org	Phytocoris varipes
	290310057	urn:lsid:faunaeur.org	Callipterinella minut
	290308656	urn:lsid:faunaeur.org	Hyperdiplosis bryanti
	290308539	urn:lsid:faunaeur.org	Scolopostethus decc

SpeciesAbundancePub2015.tsv

Open with Goog

A	B	C	D	E	F
observationID	cropCode	cropName	speciesID	countryCode	countryText
2147365908	A059G	Maize	290307251	ES	Spain
2147365909	A059G	Maize	290307131	ES	Spain
2147365910	A059G	Maize	290307252	ES	Spain
2147365911	A059G	Maize	290307205	ES	Spain
2147365216	A059G	Maize	290307128	ES	Spain

speciesID is a key between tables

SpeciesInfoPub2015.tsv

A	B	C
speciesID	GBIF	species
290307346	urn:lsid:faunaeur.org	Allygidius commutat
290308575	urn:lsid:faunaeur.org	Phytocoris varipes
290310057	urn:lsid:faunaeur.org	Callipterinella minut
290308656	urn:lsid:faunaeur.org	Hyperdiplosis bryanti
290308539	urn:lsid:faunaeur.org	Scolopostethus decc

Summary of data features

This is a database of pests eating plants

It contains:

- Information about Infections
- Information about pests/pathogens/parasites
- Information about hosts
- Infection observations
- At a location
- In a country
- Over a time-period

Hosts are identified by their EU foodtype code

Infection observations and pests have local ID numbers

Pests are also annotated by a special ID (LSID)

Geographic locations are properties, with no identifiers

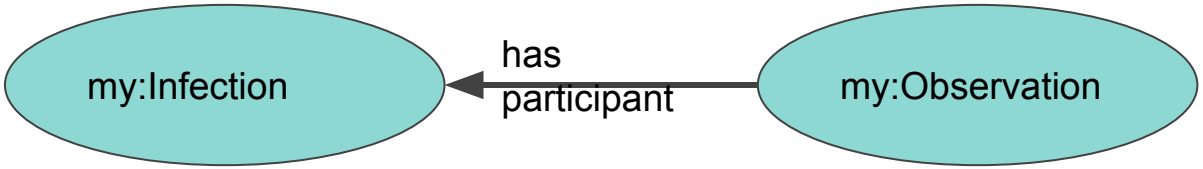
Countries are identified by their ISO 3166-2 Country Codes (e.g. ES)

Start with the infections

In words: These are observations of parasitic infections

Start with the infections

In words: These are observations of parasitic infections

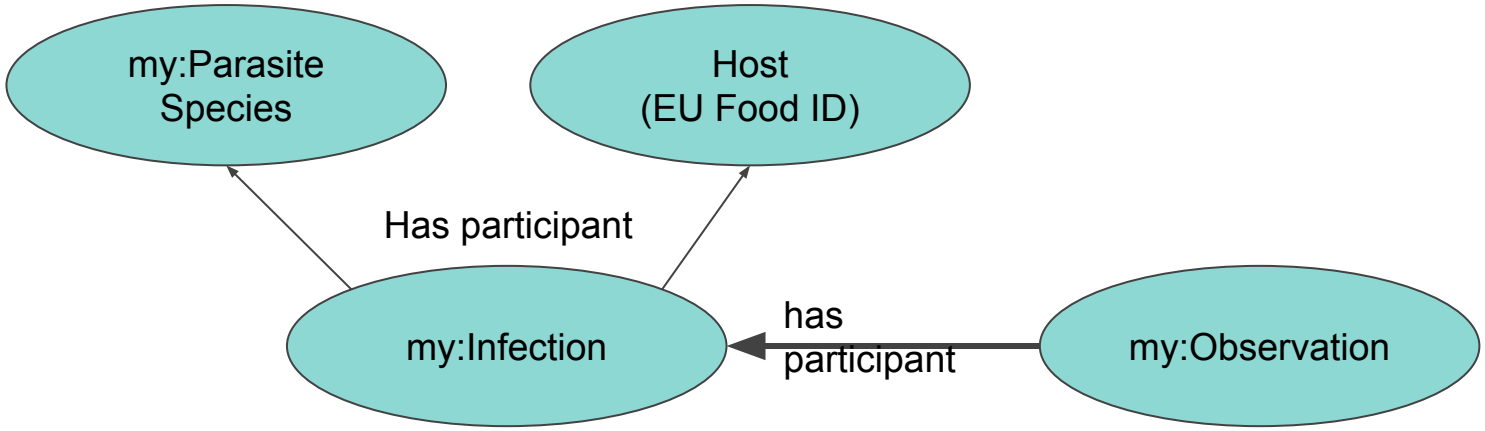


An observation is an Activity, an infection is also an Activity

In SIO, an Activity is allowed to be a “participant in” another Activity

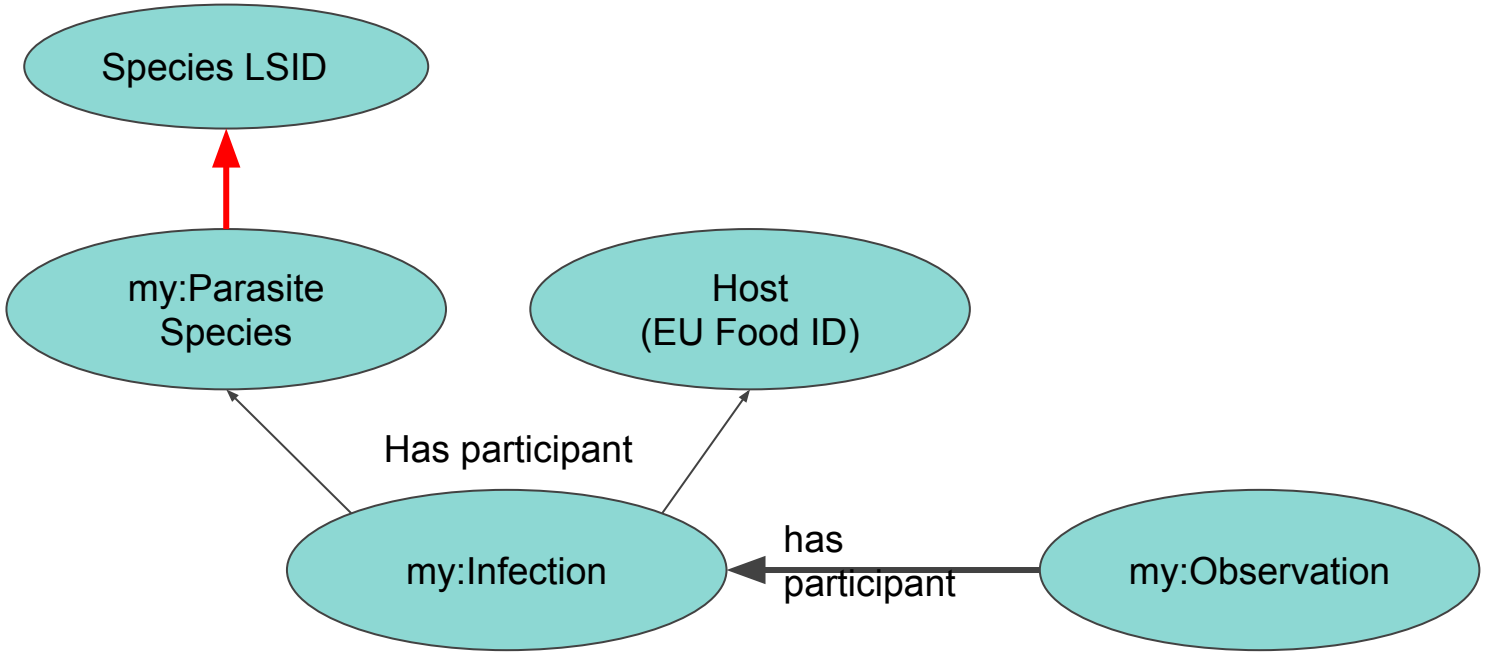
Start with the infections

Parasitic infections also have participants (host and parasite)



Start with the infections

Those parasitic species also have an external ID number (LSID)... SIO has a property that can represent this



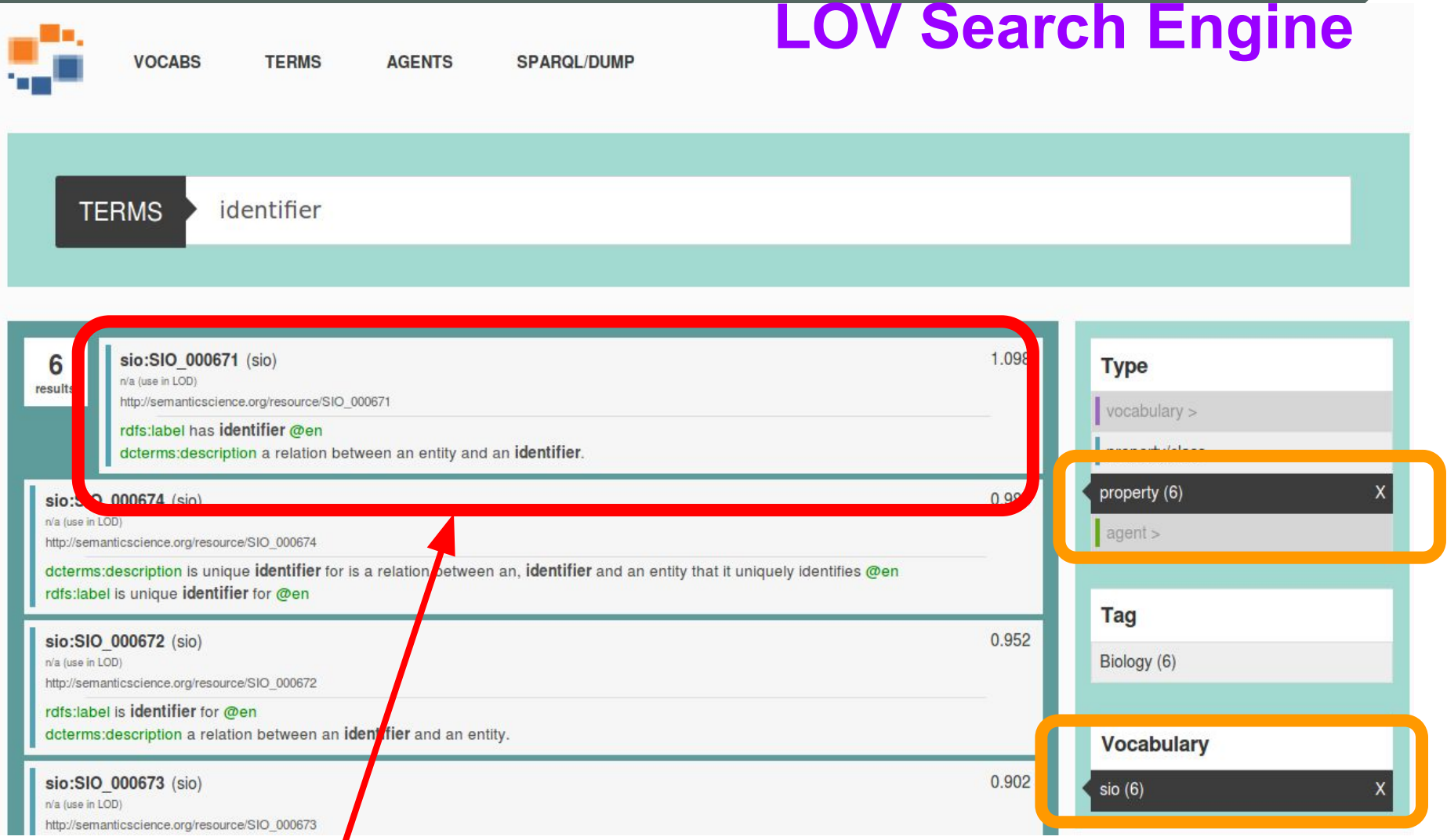
Searching Ontologies

https://lov.linkeddata.es/dataset/lov/

https://www.ebi.ac.uk/ols/index

(these two lookup services are partially redundant. EBI's service is primarily biological)

LOV Search Engine



SIO:has-identifier

NOTE: I filtered for Properties (predicates) that come from SIO

http://semanticscience.org/resource/has-identifier

has identifier [SIO:000671]

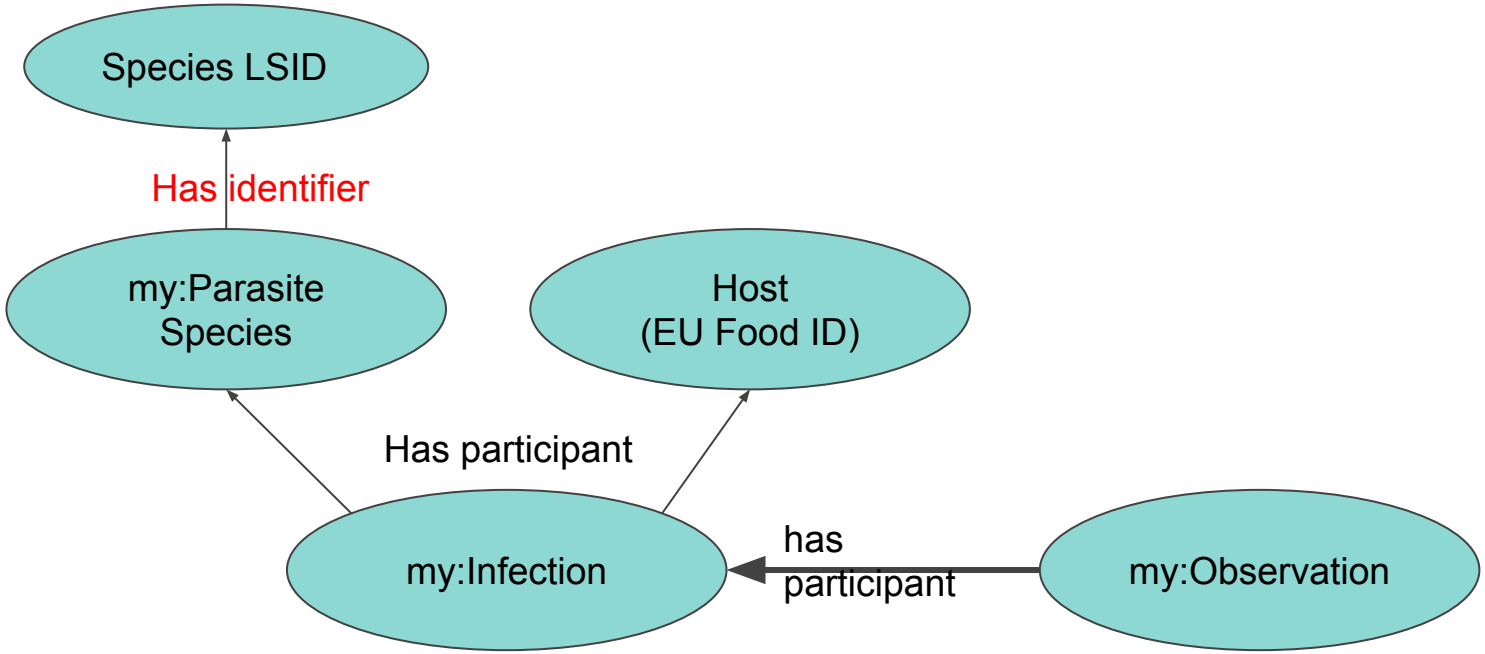
label has identifier
identifier SIO:000671
description a relation between an entity and an identifier.

term type ObjectProperty
subproperty of has attribute
superproperty of has unique identifier
inverse of is identifier for

is defined by http://semanticscience.org/ontology/sio.owl

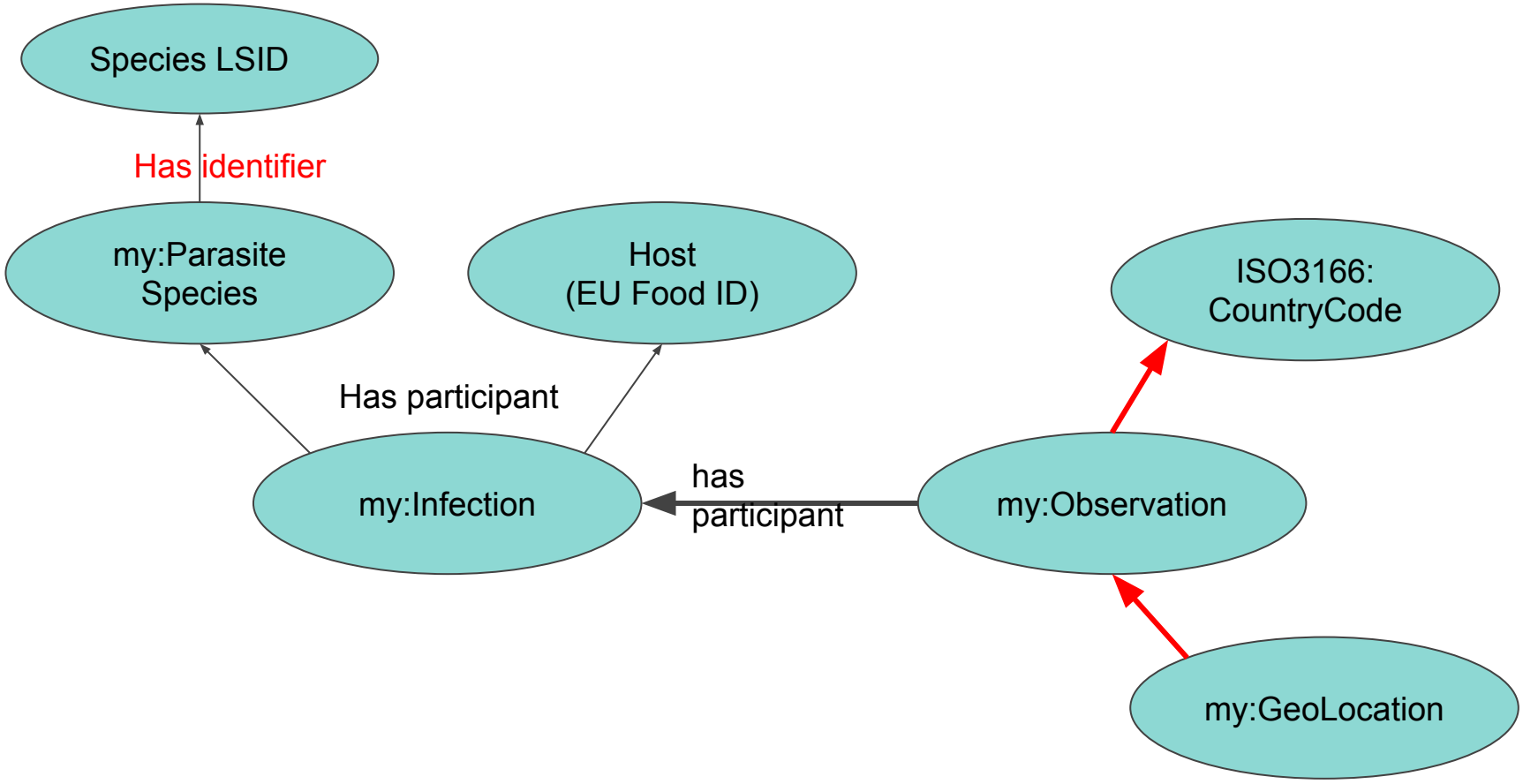
Start with the infections

Those parasitic species also have an external ID number (LSID)... SIO has a property that can represent this

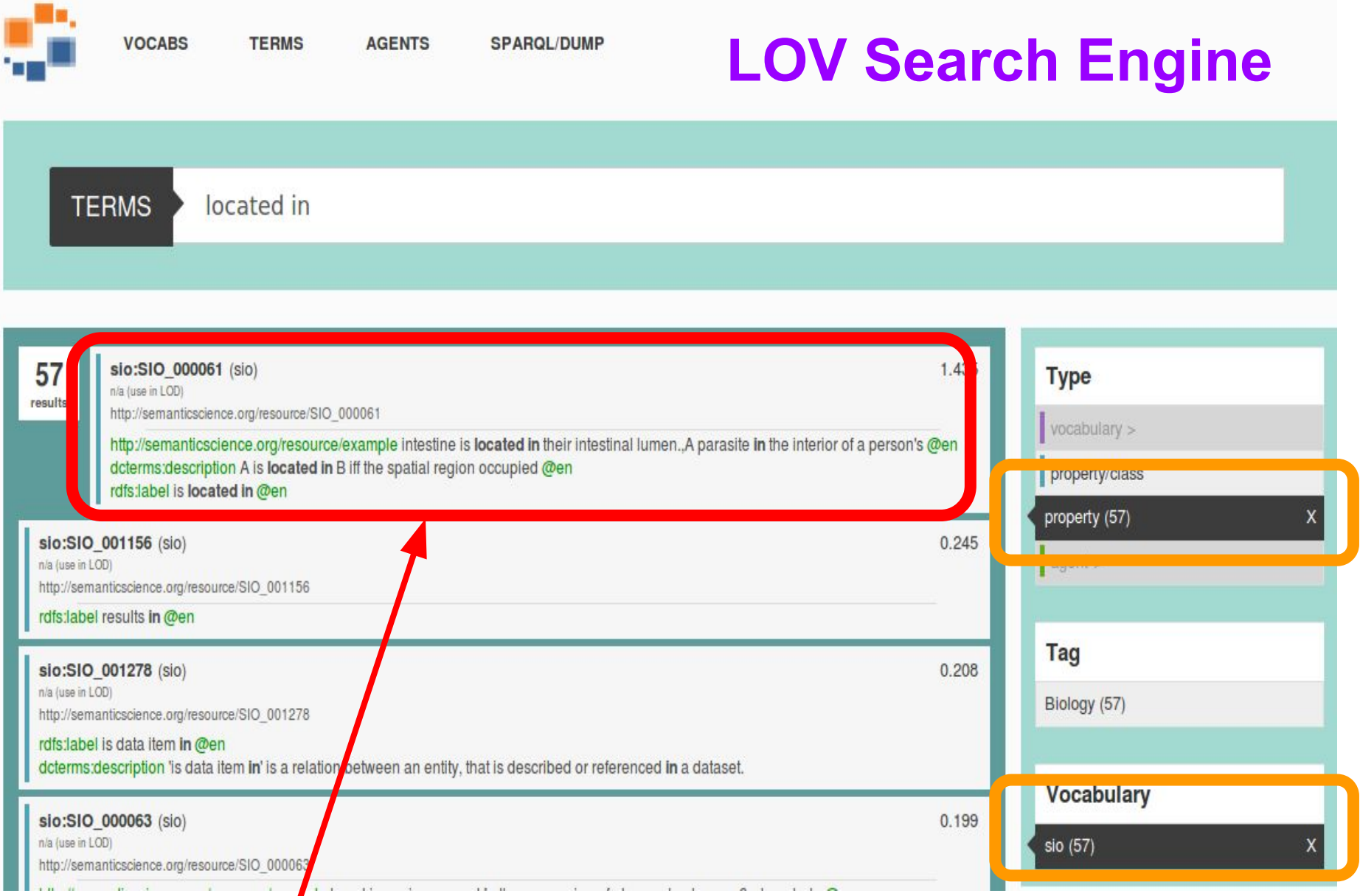


Start with the infections

Observations happen in a country, and at a geographical location (I'm going to mix things up a bit and reverse the "direction" of the geo predicate)



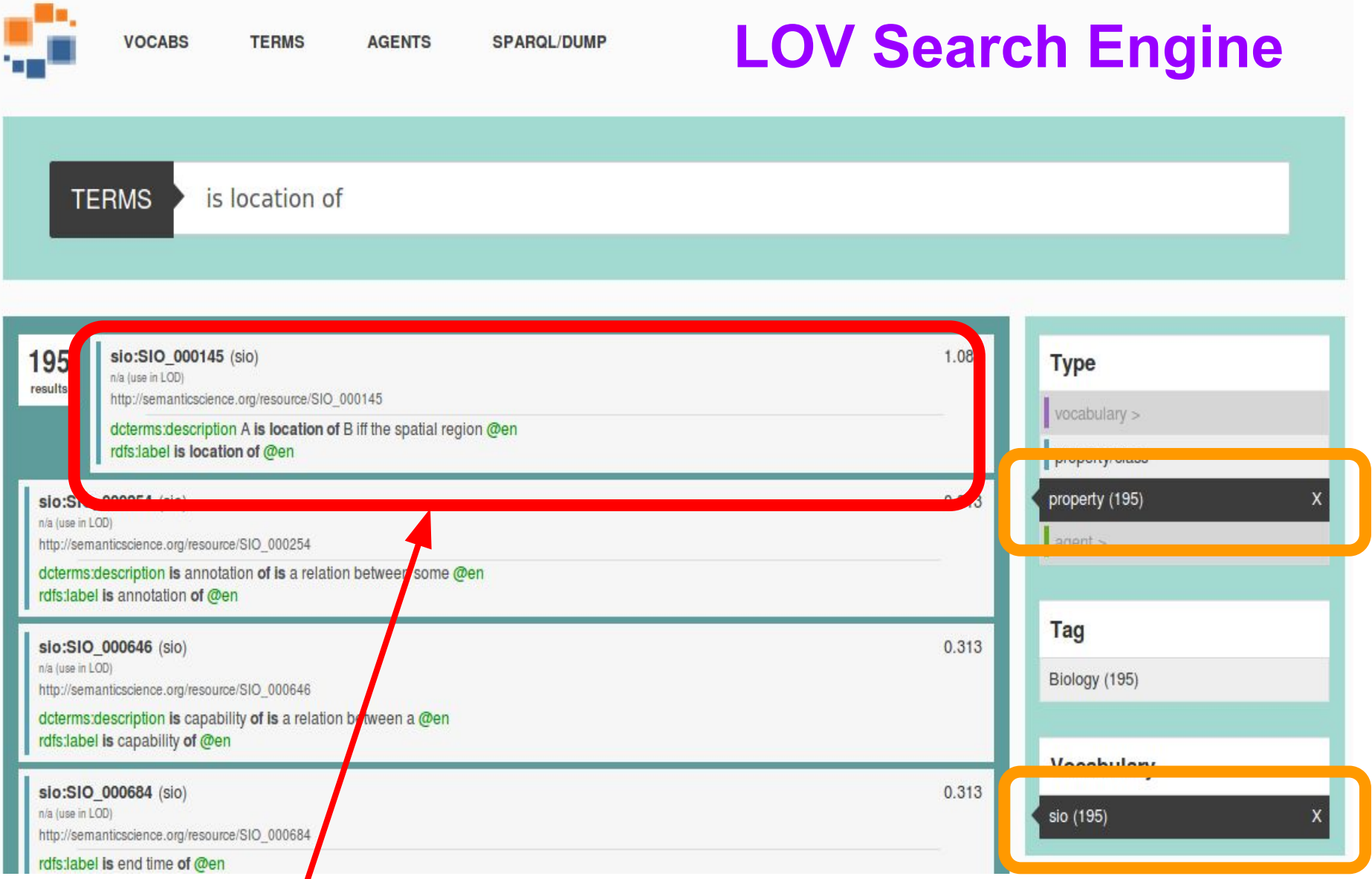
LOV Search Engine



SIO:is-located-in

NOTE: I filtered for Properties (predicates) that come from SIO

LOV Search Engine

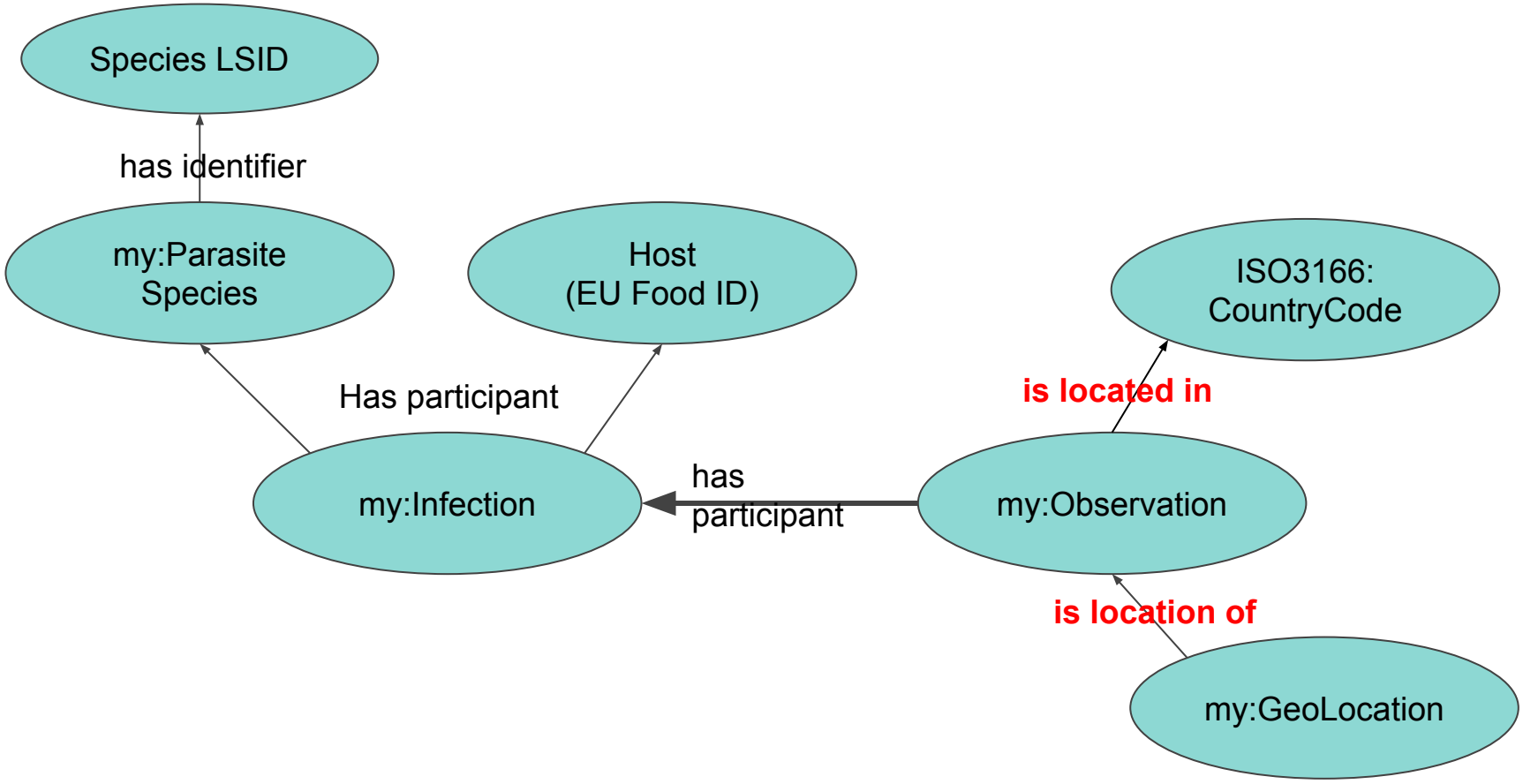


SIO:is-location-of

NOTE: I filtered for Properties (predicates) that come from SIO

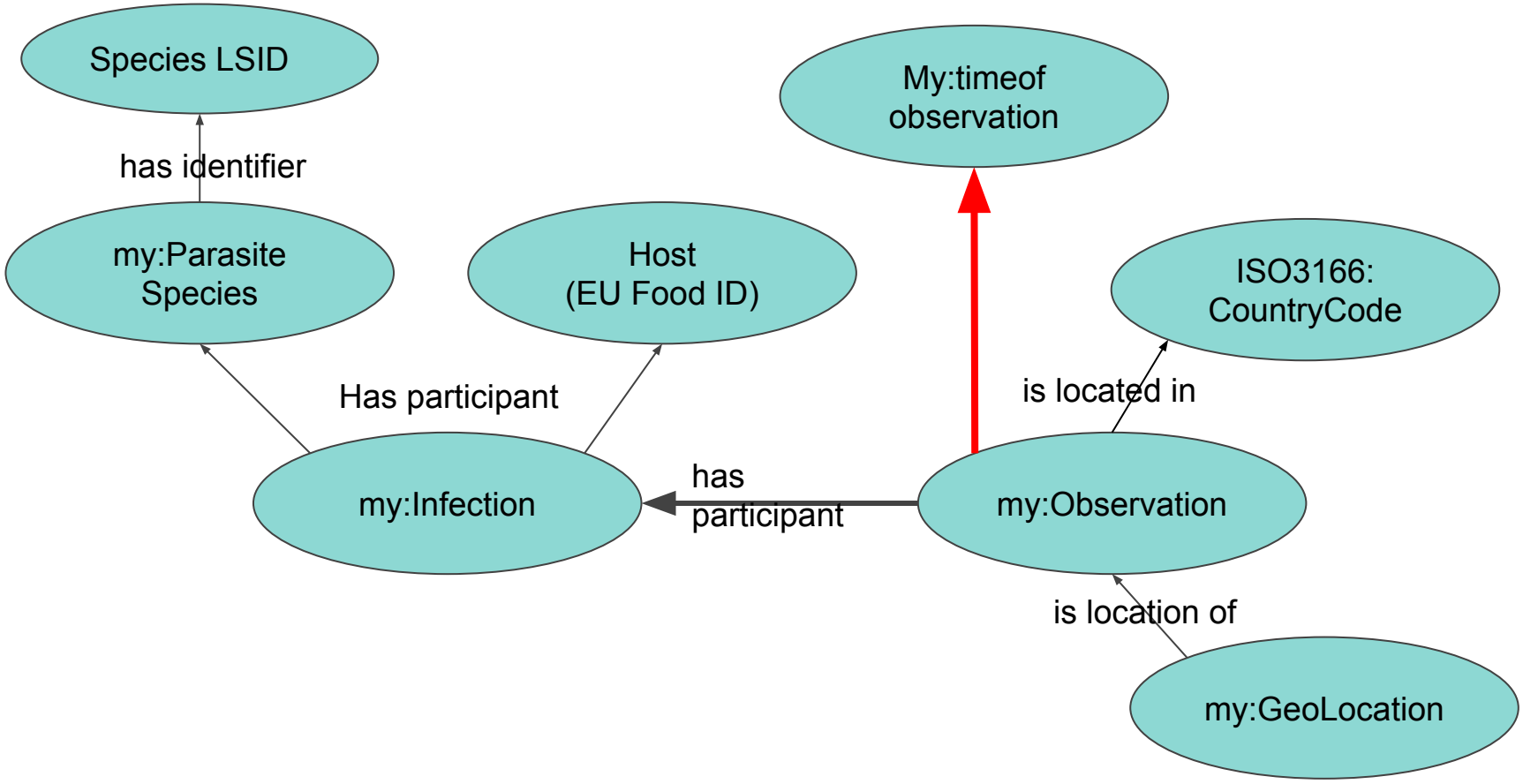
Start with the infections

Observations happen in a country, and at a geographical location (I'm going to mix things up a bit and reverse the "direction" of the geo predicate)



Start with the infections

Finally, we have the time of observation



The correct predicate was quite hard to find!

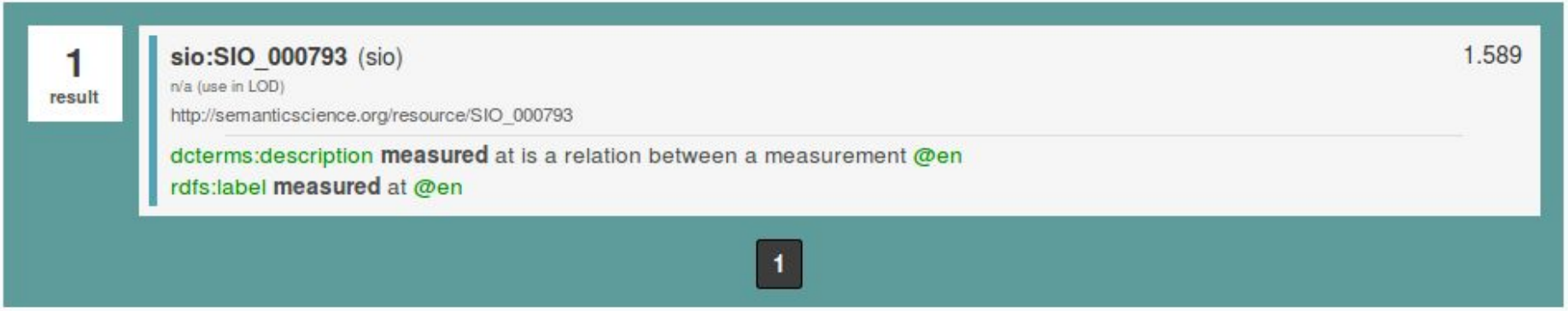
When you know an ontology well, it becomes easy to find a predicate

I am a co-author of SIO, and this is one predicate that I didn't know!

It took me about 10 minutes to find it :-)

I searched: **"time", "happened", "occurred", and "measured"**

"Measured" gave me the predicate "**measured at**"... but "**measured at**" could be spatial or temporal! So I had to look it up and be sure!

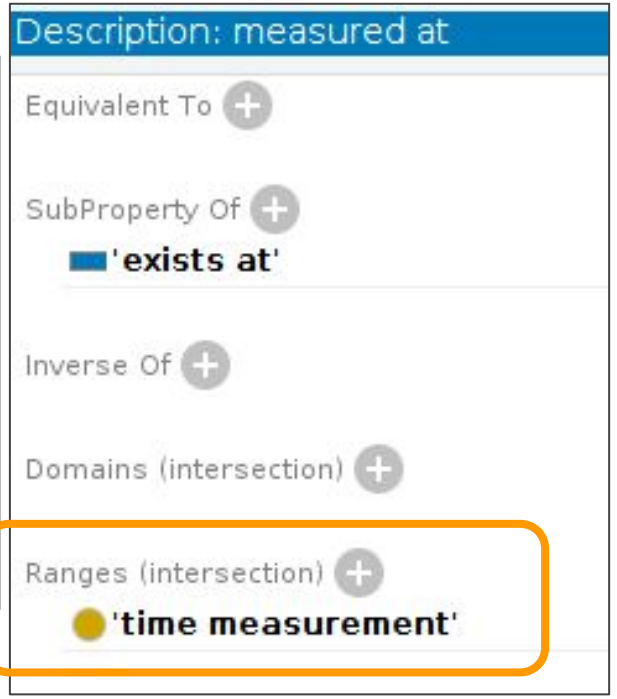


measured at [SIO:000793]

label measured at
identifier SIO:000793
description measured at is a relation between a measurement value and the time measurement.

term type ObjectProperty
subproperty of exists at

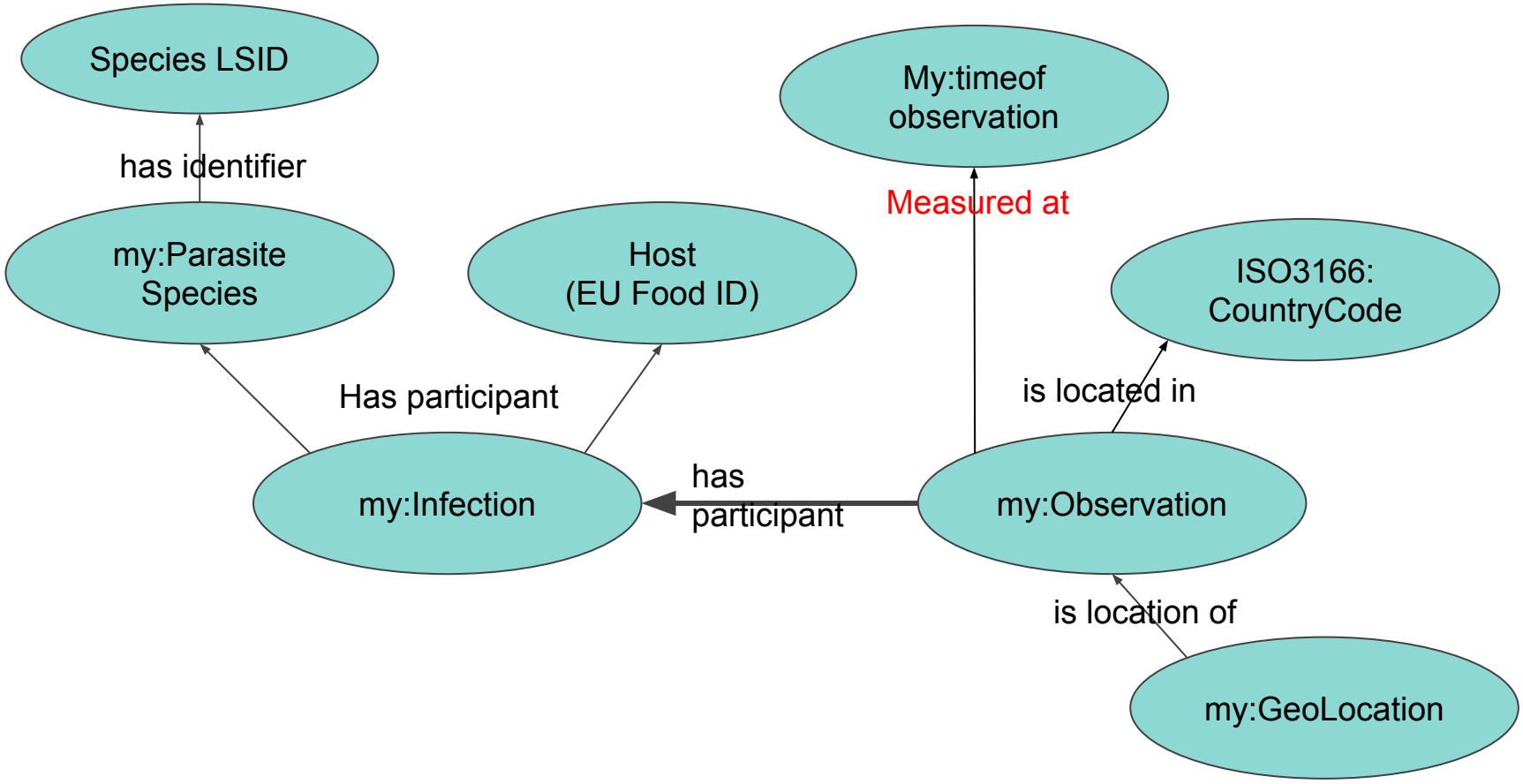
is defined by http://semanticscience.org/ontology/sio.owl



Bingo!

Start with the infections

Finally, we have the time of observation



“Good” RDF has a type on every node

We have our skeleton data model

Next step is to decide the Class of each of those nodes… what does it represent?

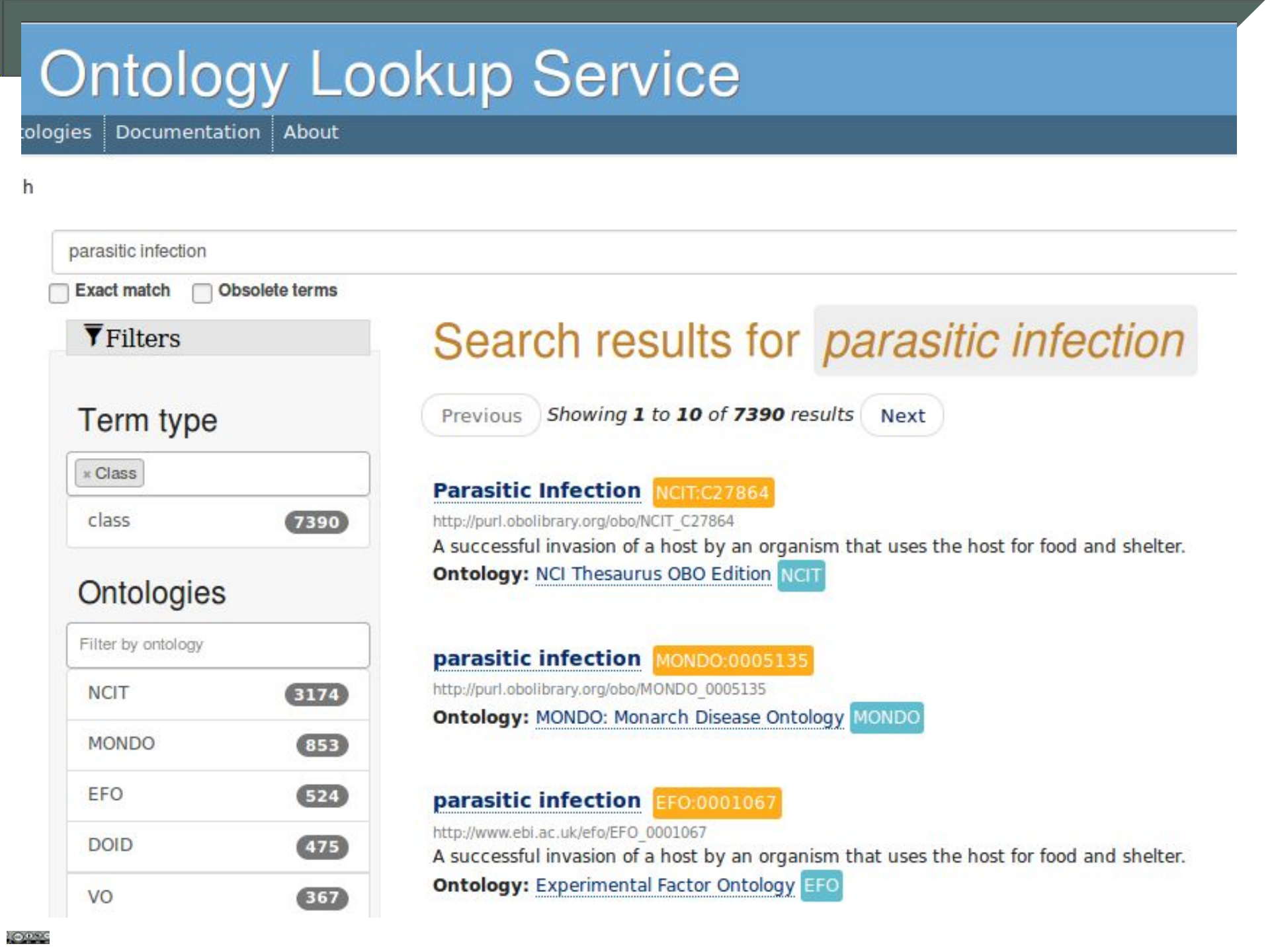
Again, we’re going to use the Ontology Lookup Service(s)

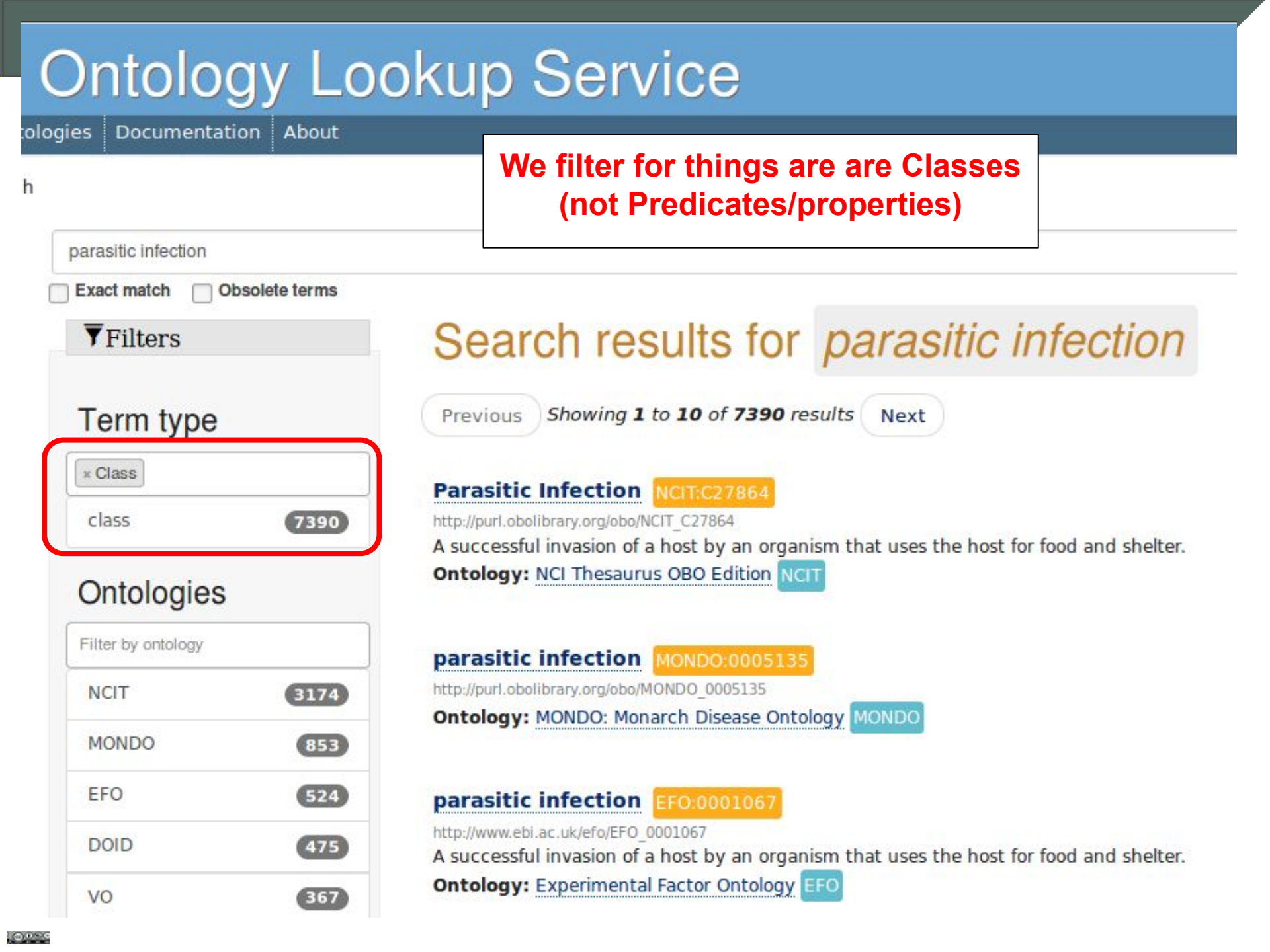
The only interesting cases are:

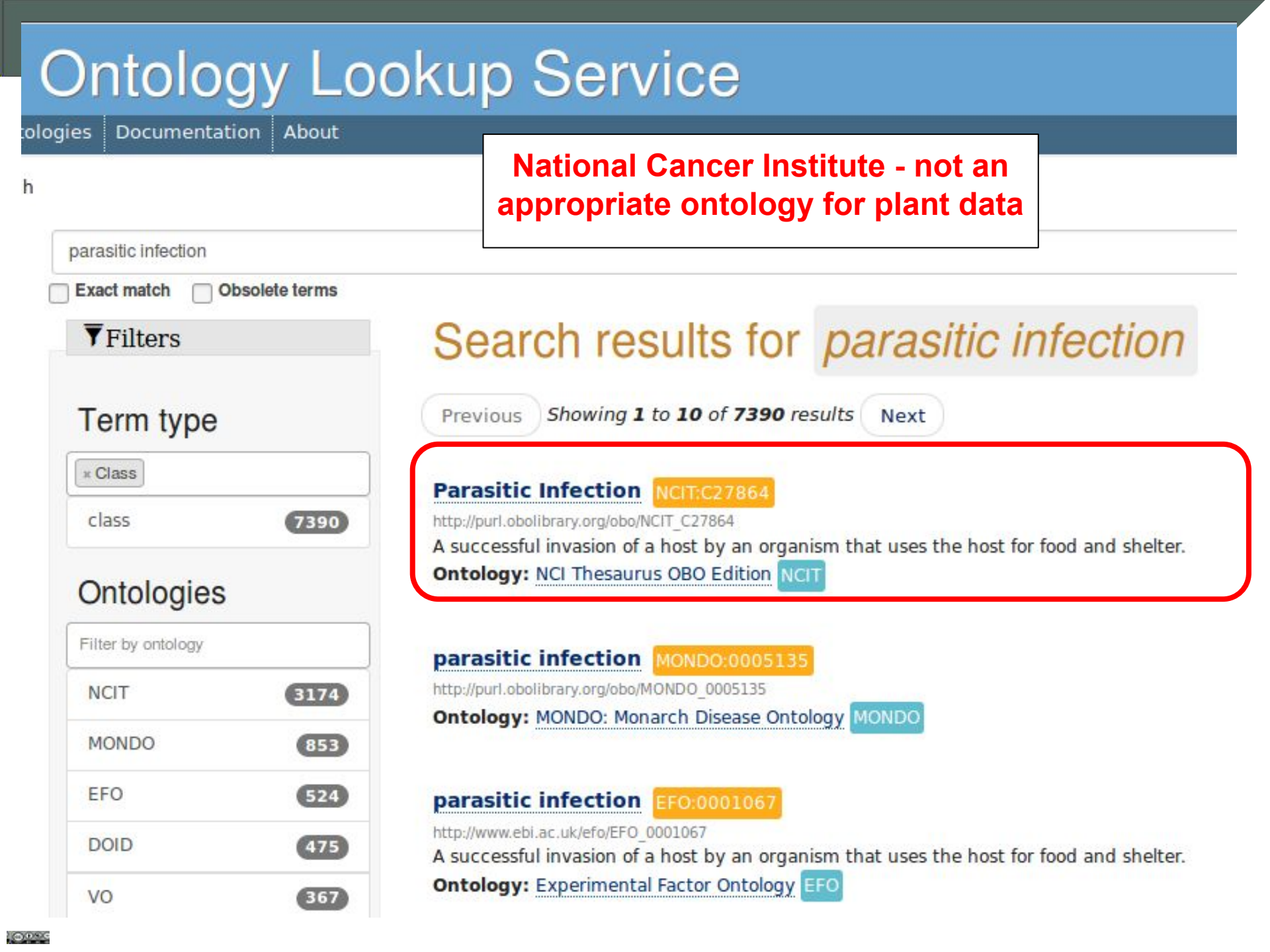
- the “infection” node, because that concept does not exist in SIO
- The Geographic Location node, because SIO does not have internal models for geolocations

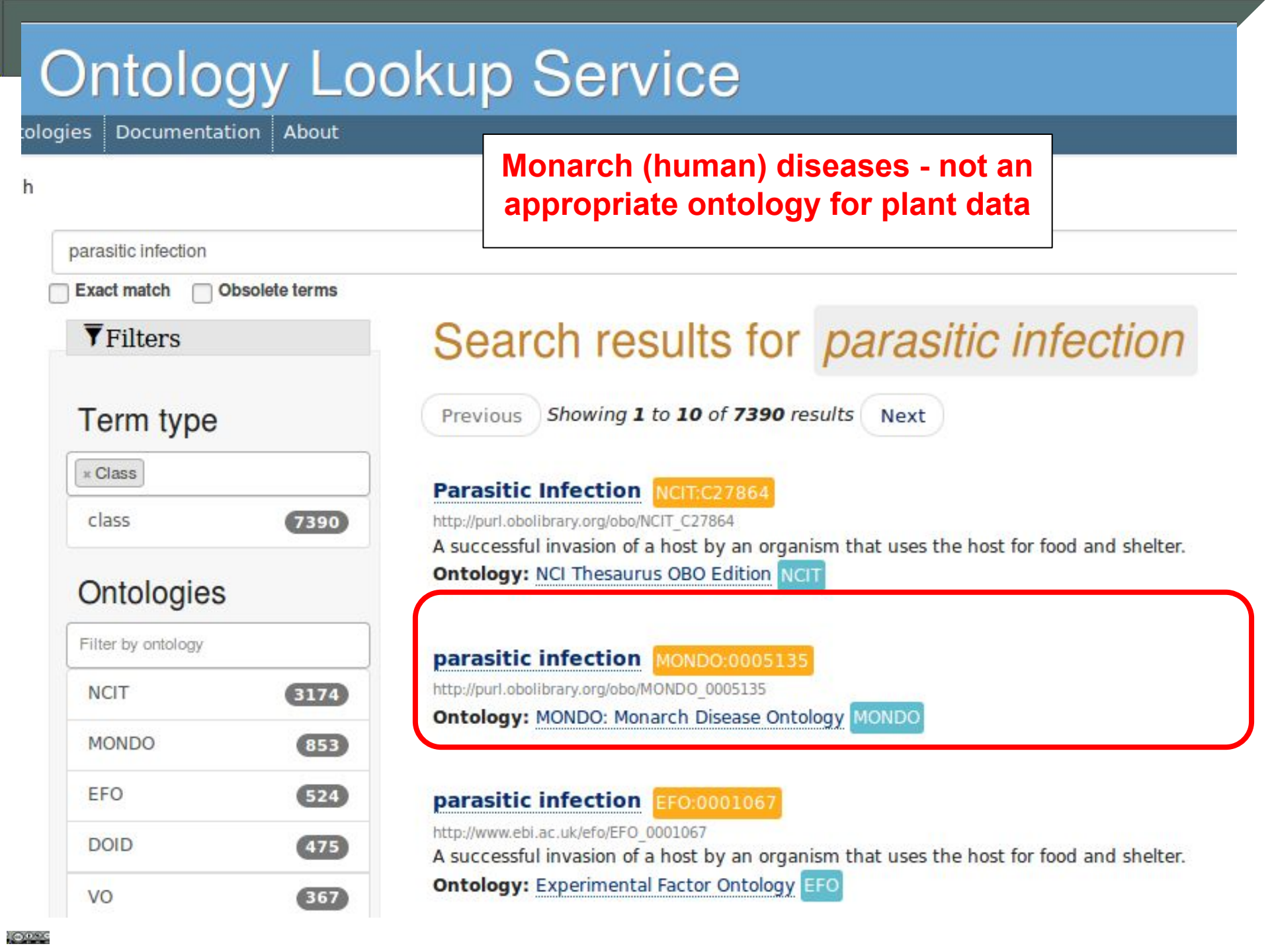
Find what we want

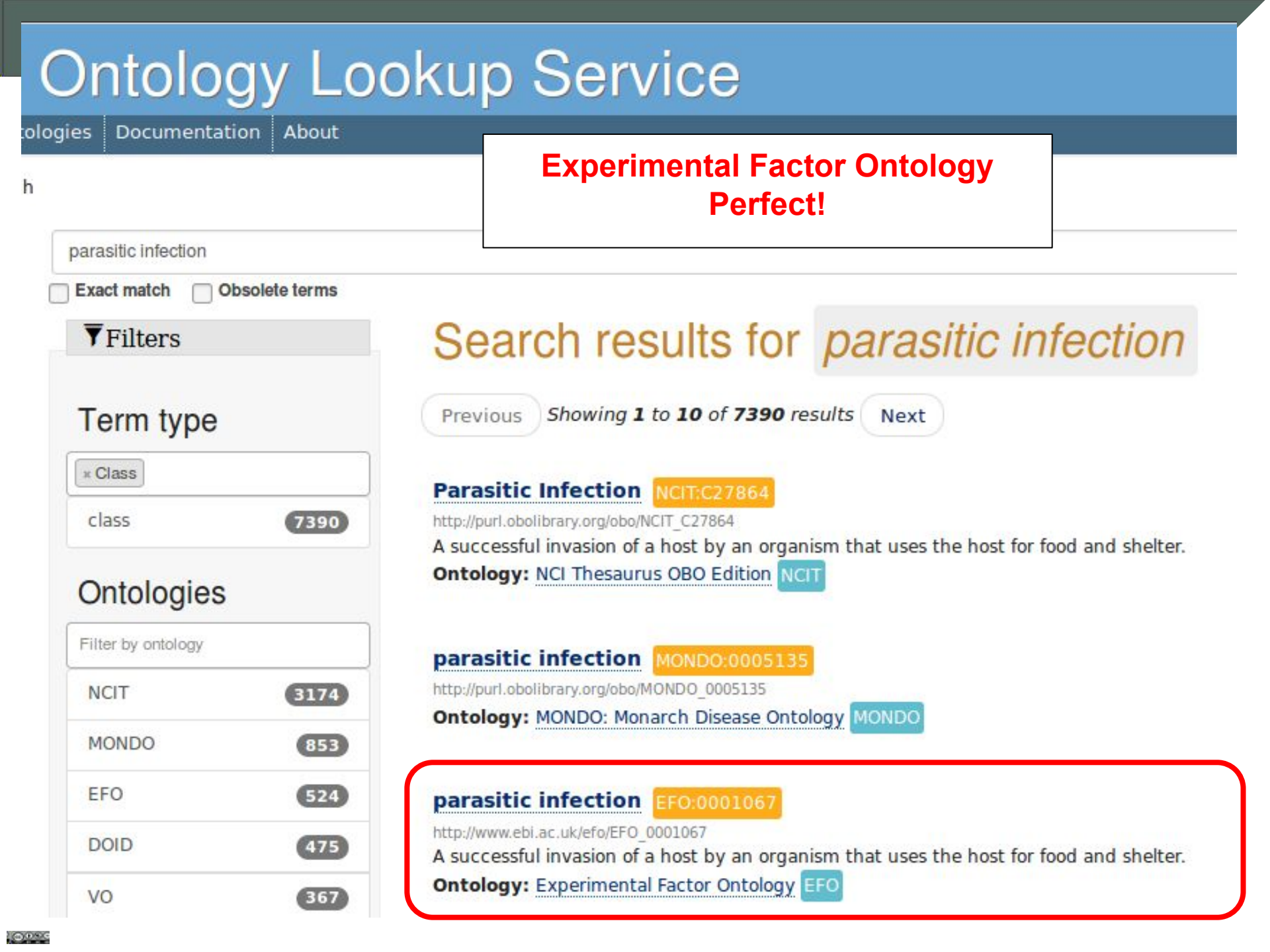
Using the EBI Ontology Lookup Service, I searched for “infection” filtering only for Classes


Ontology Lookup Service
tologies
Documentation
About
h
parasitic infection
Exact match
Obsolete terms
Filters
Term type
× Class
class 7390
Ontologies
Filter by ontology
NCIT 3174
MONDO 853
EFO 524
DOID 475
VO 367
Search results for parasitic infection
Previous
Showing 1 to 10 of 7390 results
Next
Parasitic Infection
NCIT:C27864
http://purl.obolibrary.org/obo/NCIT_C27864
A successful invasion of a host by an organism that uses the host for food and shelter.
Ontology: NCI Thesaurus OBO Edition NCIT
parasitic infection
MONDO:0005135
http://purl.obolibrary.org/obo/MONDO_0005135
Ontology: MONDO: Monarch Disease Ontology MONDO
parasitic infection
EFO:0001067
http://www.ebi.ac.uk/efo/EFO_0001067
A successful invasion of a host by an organism that uses the host for food and shelter.
Ontology: Experimental Factor Ontology EFO


Ontology Lookup Service
tologies
Documentation
About
We filter for things are are Classes
(not Predicates/properties)
h
parasitic infection
Exact match
Obsolete terms
Filters
Term type
× Class
class 7390
Ontologies
Filter by ontology
NCIT 3174
MONDO 853
EFO 524
DOID 475
VO 367
Search results for parasitic infection
Previous
Showing 1 to 10 of 7390 results
Next
Parasitic Infection
NCIT:C27864
http://purl.obolibrary.org/obo/NCIT_C27864
A successful invasion of a host by an organism that uses the host for food and shelter.
Ontology: NCI Thesaurus OBO Edition NCIT
parasitic infection
MONDO:0005135
http://purl.obolibrary.org/obo/MONDO_0005135
Ontology: MONDO: Monarch Disease Ontology MONDO
parasitic infection
EFO:0001067
http://www.ebi.ac.uk/efo/EFO_0001067
A successful invasion of a host by an organism that uses the host for food and shelter.
Ontology: Experimental Factor Ontology EFO


Ontology Lookup Service
tologies
Documentation
About
National Cancer Institute - not an appropriate ontology for plant data
h
parasitic infection
Exact match
Obsolete terms
Filters
Term type
× Class
class 7390
Ontologies
Filter by ontology
NCIT 3174
MONDO 853
EFO 524
DOID 475
VO 367
Search results for parasitic infection
Previous
Showing 1 to 10 of 7390 results
Next
Parasitic Infection NCIT:C27864
http://purl.obolibrary.org/obo/NCIT_C27864
A successful invasion of a host by an organism that uses the host for food and shelter.
Ontology: NCI Thesaurus OBO Edition NCIT
parasitic infection MONDO:0005135
http://purl.obolibrary.org/obo/MONDO_0005135
Ontology: MONDO: Monarch Disease Ontology MONDO
parasitic infection EFO:0001067
http://www.ebi.ac.uk/efo/EFO_0001067
A successful invasion of a host by an organism that uses the host for food and shelter.
Ontology: Experimental Factor Ontology EFO


Ontology Lookup Service
tologies
Documentation
About
Monarch (human) diseases - not an appropriate ontology for plant data
h
parasitic infection
Exact match
Obsolete terms
Filters
Term type
× Class
class 7390
Ontologies
Filter by ontology
NCIT 3174
MONDO 853
EFO 524
DOID 475
VO 367
Search results for parasitic infection
Previous
Showing 1 to 10 of 7390 results
Next
Parasitic Infection NCIT:C27864
http://purl.obolibrary.org/obo/NCIT_C27864
A successful invasion of a host by an organism that uses the host for food and shelter.
Ontology: NCI Thesaurus OBO Edition NCIT
parasitic infection MONDO:0005135
http://purl.obolibrary.org/obo/MONDO_0005135
Ontology: MONDO: Monarch Disease Ontology MONDO
parasitic infection EFO:0001067
http://www.ebi.ac.uk/efo/EFO_0001067
A successful invasion of a host by an organism that uses the host for food and shelter.
Ontology: Experimental Factor Ontology EFO

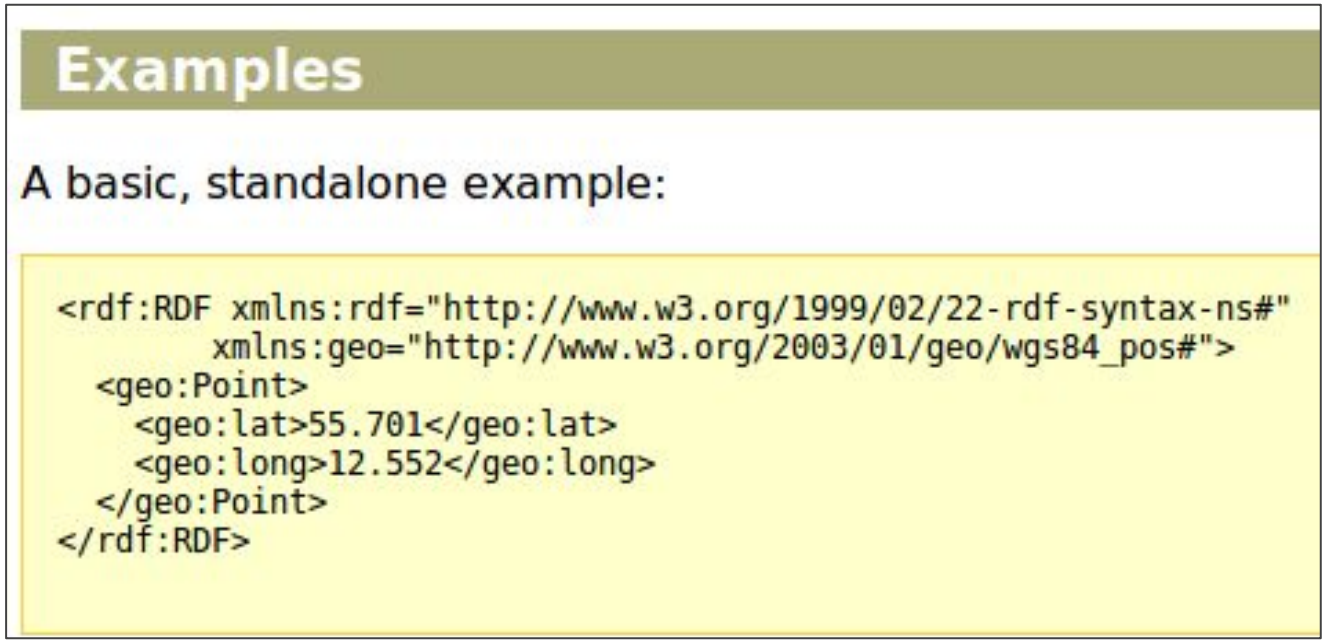

Ontology Lookup Service
tologies
Documentation
About
Experimental Factor Ontology
Perfect!
h
parasitic infection
Exact match
Obsolete terms
Filters
Term type
× Class
class 7390
Ontologies
Filter by ontology
NCIT 3174
MONDO 853
EFO 524
DOID 475
VO 367
Search results for parasitic infection
Previous
Showing 1 to 10 of 7390 results
Next
Parasitic Infection NCIT:C27864
http://purl.obolibrary.org/obo/NCIT_C27864
A successful invasion of a host by an organism that uses the host for food and shelter.
Ontology: NCI Thesaurus OBO Edition NCIT
parasitic infection MONDO:0005135
http://purl.obolibrary.org/obo/MONDO_0005135
Ontology: MONDO: Monarch Disease Ontology MONDO
parasitic infection EFO:0001067
http://www.ebi.ac.uk/efo/EFO_0001067
A successful invasion of a host by an organism that uses the host for food and shelter.
Ontology: Experimental Factor Ontology EFO

For the geographic location

I could have used Google to find the Geo ontology

(I knew where it was already)

https://www.w3.org/2003/01/geo/



Examples

A basic, standalone example:

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#">
  <geo:Point>
    <geo:lat>55.701</geo:lat>
    <geo:long>12.552</geo:long>
  </geo:Point>
</rdf:RDF>
```

For the geographic location

How to “read” RDF+XML

(we could also look-up this information in the ontology browsers! ...I just want to show you how to interpret examples you may find on the Web…)

This is the Class (“Point” in the geo: namespace)

Examples

A basic, standalone example:

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#">
  <geo:Point>
    <geo:lat>55.701</geo:lat>
    <geo:long>12.552</geo:long>
  </geo:Point>
</rdf:RDF>
```

For the geographic location

How to “read” RDF+XML

(we could also look-up this information in the ontology browsers! ...I just want to show you how to interpret examples you may find on the Web…)

These are the properties (“lat”, and “long” in the geo: namespace)

Examples

A basic, standalone example:

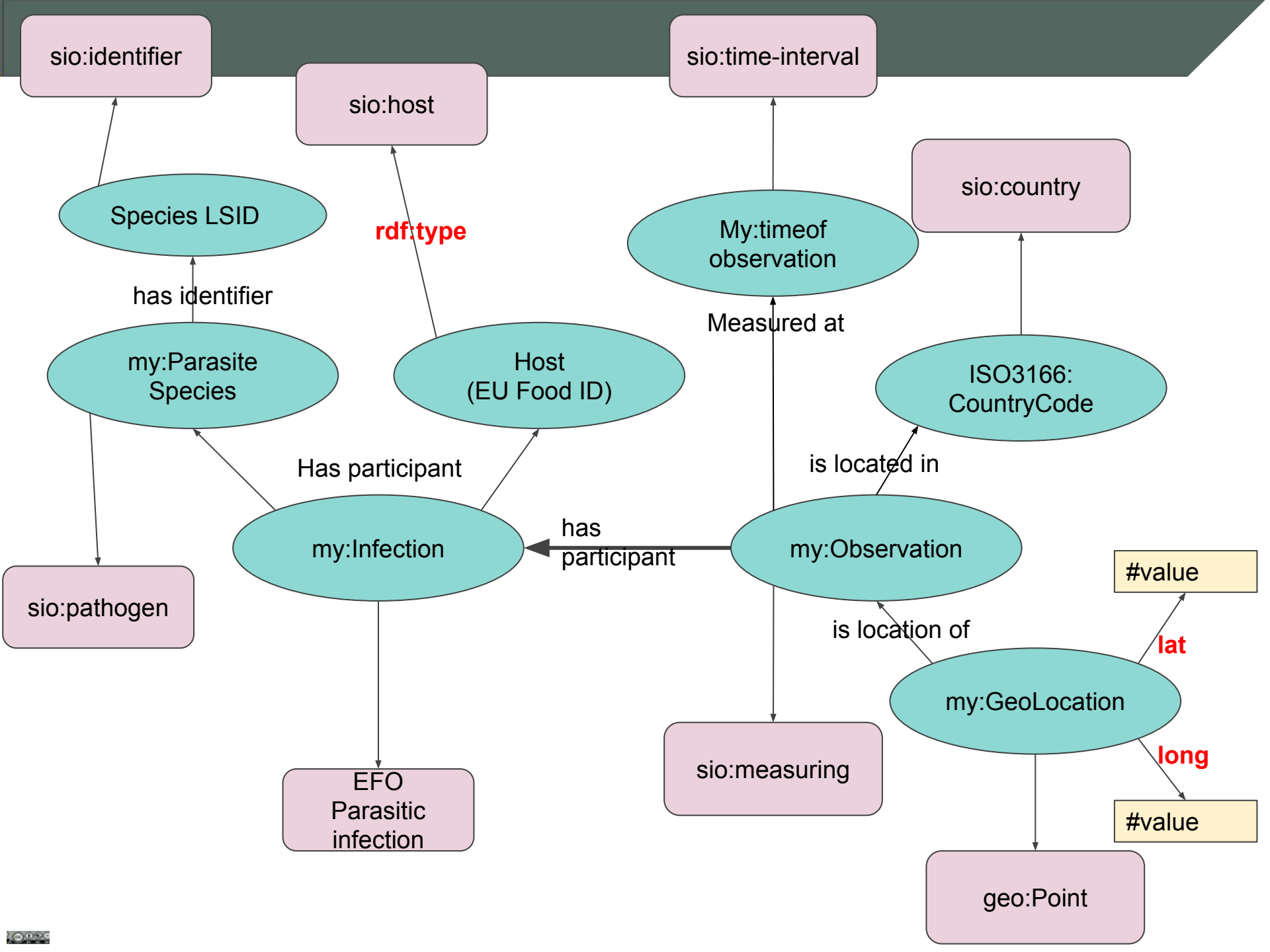
```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#">
  <geo:Point>
    <geo:lat>55.701</geo:lat>
    <geo:long>12.552</geo:long>
  </geo:Point>
</rdf:RDF>
```

All of the other nodes we type by the terms we find

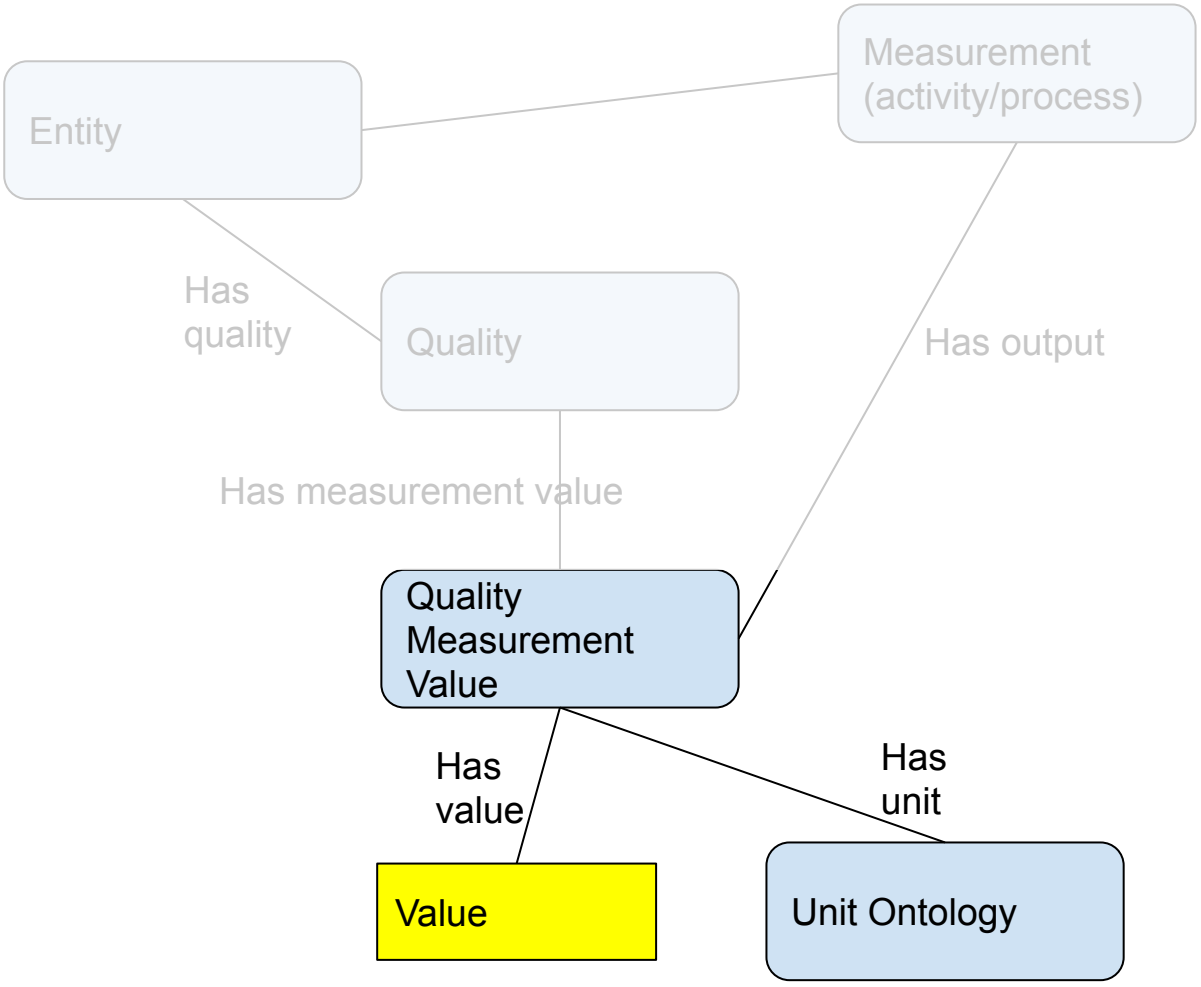
The predicate in every case is **rdf:type**

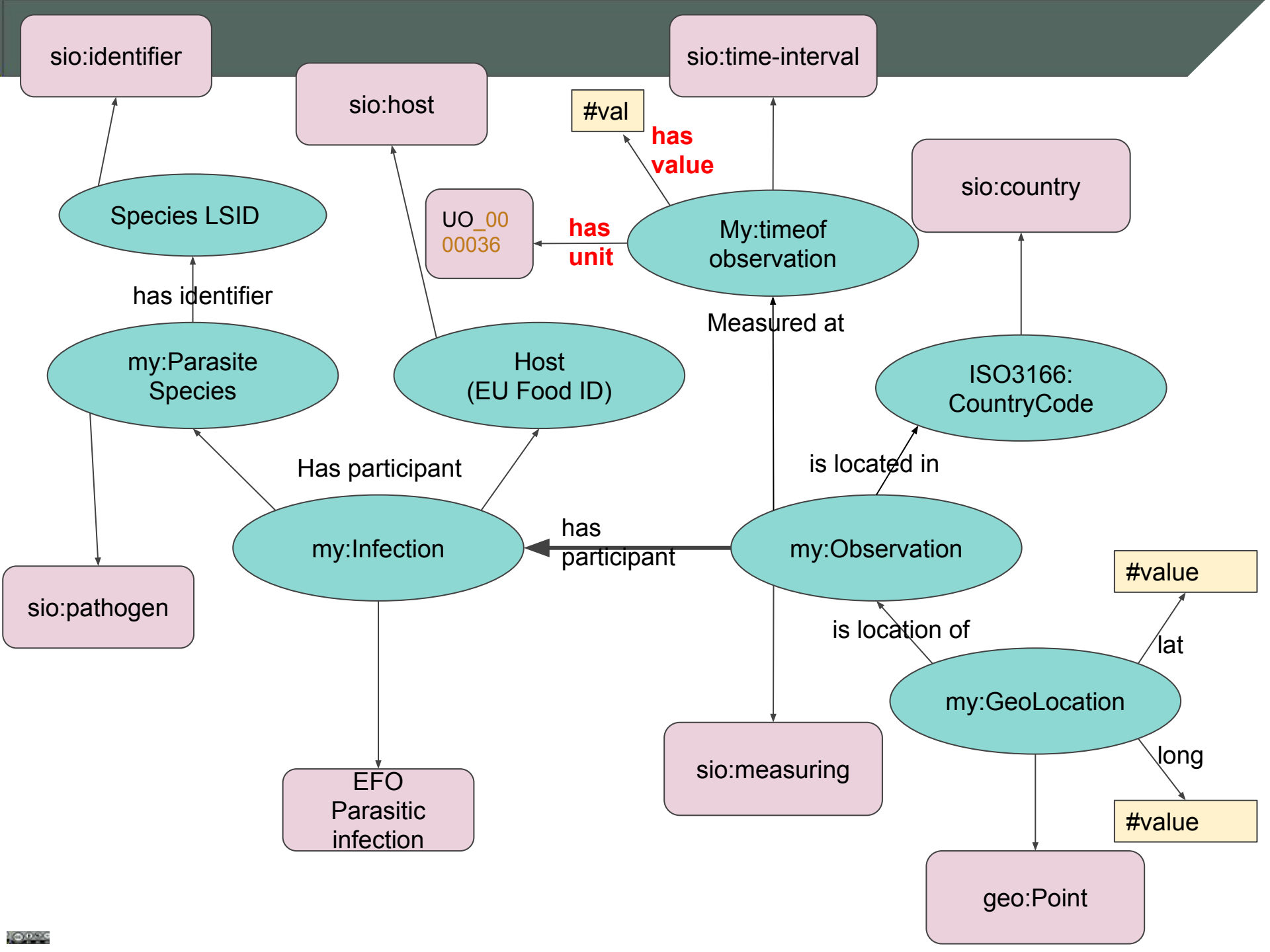
If you ever see a prefix, like "rdf", and you don't know what the expanded address is, just surf to:

http://prefix.cc


sio:identifier
sio:host
sio:time-interval
Species LSID
rdf:type
My:timeof observation
sio:country
has identifier
Measured at
my:Parasite Species
Host (EU Food ID)
ISO3166: CountryCode
Has participant
is located in
my:Infection
has participant
my:Observation
sio:pathogen
#value
is location of
lat
my:GeoLocation
sio:measuring
long
#value
EFO Parasitic infection
geo:Point

Capturing values in SIO: there is ONLY ONE WAY!

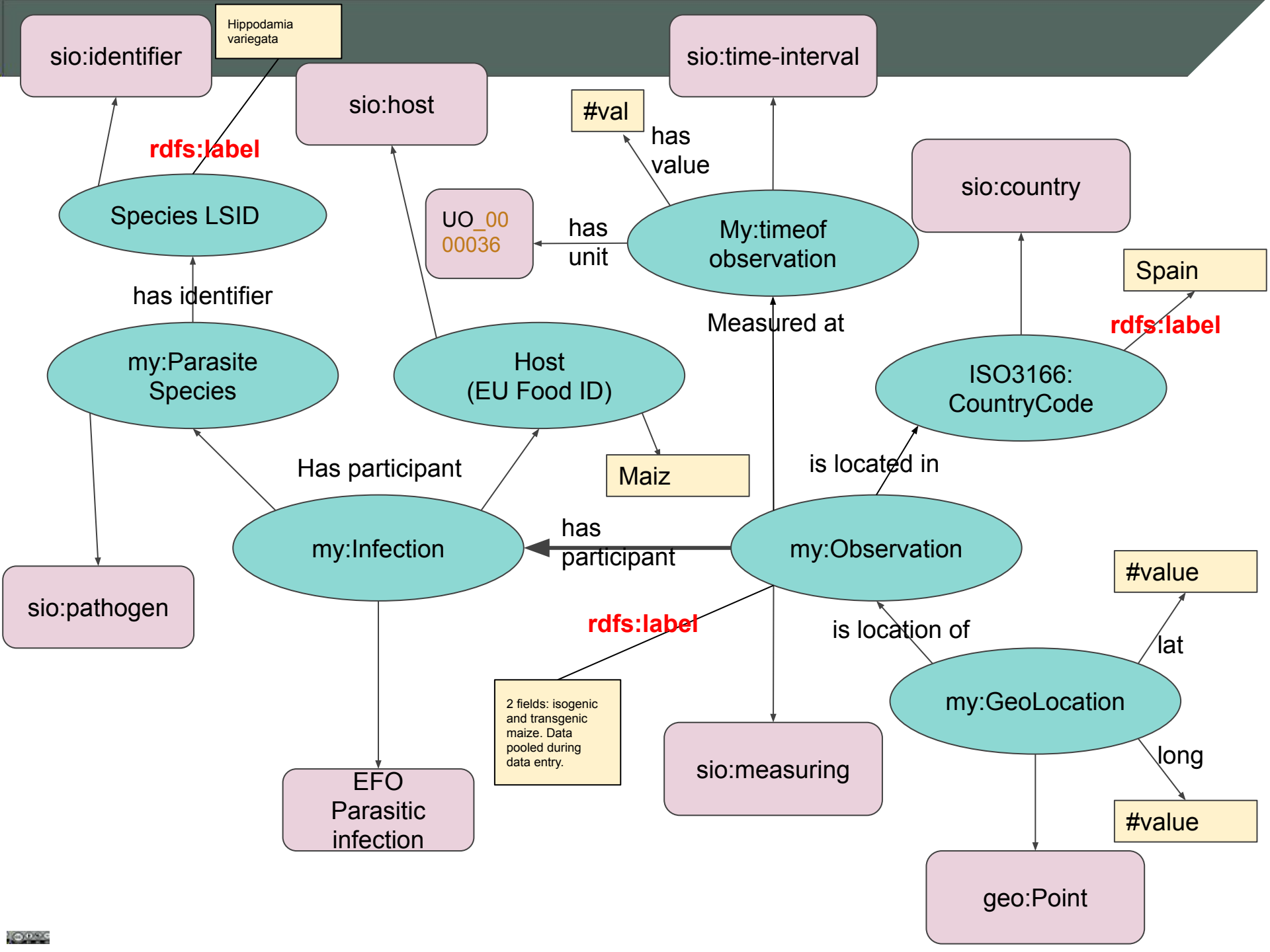



sio:identifier
sio:host
sio:time-interval
#val
has value
Species LSID
UO_0000036
has unit
My:timeof observation
sio:country
has identifier
Measured at
my:Parasite Species
Host (EU Food ID)
ISO3166: CountryCode
Has participant
is located in
my:Infection
has participant
my:Observation
sio:pathogen
#value
is location of
lat
my:GeoLocation
sio:measuring
long
#value
EFO Parasitic infection
geo:Point

And all other information just becomes a “rdfs:label”

All of the other information is for “human readers”, so it becomes a label

The final graph looks like this:


sio:identifier
Hippodamia variegata
sio:host
sio:time-interval
#val
has value
rdfs:label
Species LSID
UO_0000036
has unit
My:timeof observation
sio:country
Spain
has identifier
Measured at
rdfs:label
my:Parasite Species
Host (EU Food ID)
ISO3166: CountryCode
Has participant
Maiz
is located in
my:Infection
has participant
my:Observation
sio:pathogen
#value
rdfs:label
is location of
lat
2 fields: isogenic and transgenic maize. Data pooled during data entry.
sio:measuring
my:GeoLocation
long
EFO Parasitic infection
#value
geo:Point

Next Step - pick URIs for each node

Considerations:

- URLs should resolve whenever possible (FAIR A1. (Meta)data are retrievable by their identifier using a standardised communications protocol)
- Where possible, we should create outward links (FAIR I3. (Meta)data include qualified references to other (meta)data)

Addressing the second issue first:

There are three data elements that we can outwardly link

SpeciesAbundancePub2015.tsv

Open with Goog

A	B	C	D	E	F
observationID	cropCode	cropName	speciesID	countryCode	countryText
2147365908	A059G	Maize	290307251	ES	Spain
2147365909	A059G	Maize	290307131	ES	Spain

This is a code used to identify crops in UK government databases

http://data.food.gov.uk/codes/foodtype/hierarchy/facet/source/_**A059G**

Food Standards Agency

ALPHA This is a new service – your feedback will help us to improve it.

Browse About Food Types Advanced Search Submit

http://data.food.gov.uk/codes / foodtype / id / _A059G

Entry: Maize (live plants)

URI: http://data.food.gov.uk/codes/foodtype/id/A059G
Notation: A059G

SpeciesAbundancePub2015.tsv

Open with Goog

A	B	C	D	E	F
observationID	cropCode	cropName	speciesID	countryCode	countryText
2147365908	A059G	Maize	290307251	ES	Spain
2147365909	A059G	Maize	290307131	ES	Spain

This is an ISO standard country code, used by (many) including WikiPedia

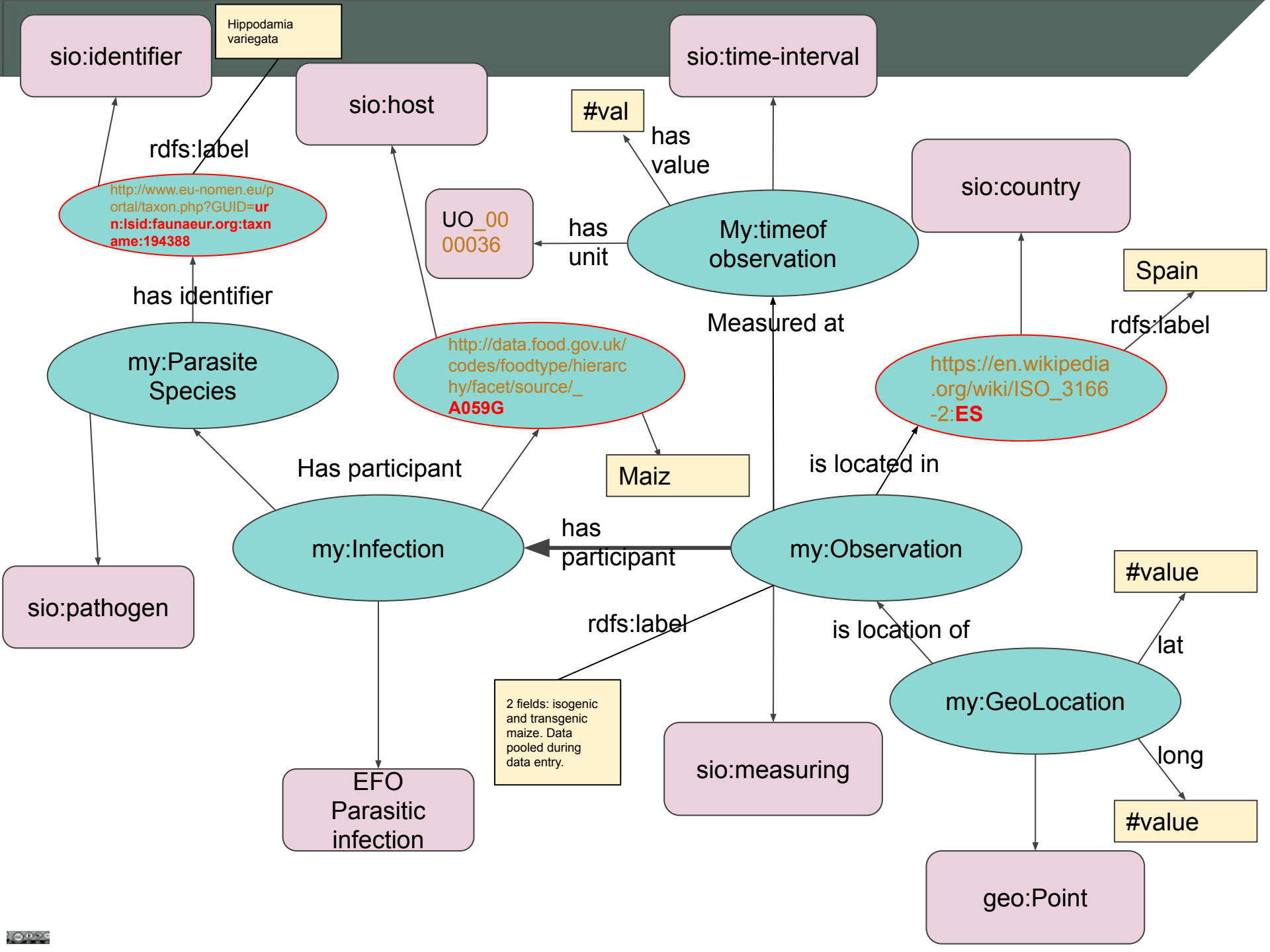
https://en.wikipedia.org/wiki/ISO_3166-2:**ES**

SpeciesInfoPub2015.tsv

	A	B	C
	speciesID	GBIF	species
	290307346	urn:lsid:faunaeur.org	Allygidius commutat
	290308575	urn:lsid:faunaeur.org	Phytocoris varipes

These are Life Sciences Identifiers (LSIDs), sed by the biodiversity community. They can be resolved to information about species:

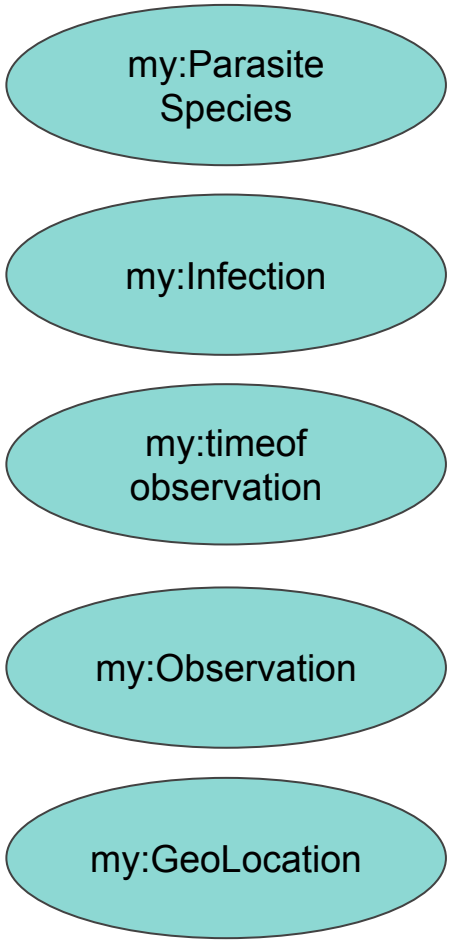
http://www.eu-nomen.eu/portal/taxon.php?GUID=**urn:lsid:faunaeur.org:taxname:194388**


sio:identifier
Hippodamia variegata
sio:host
sio:time-interval
#val
has value
rdfs:label
http://www.eu-nomen.eu/portal/taxon.php?GUID=urn:lsid:faunaeur.org:taxname:194388
UO_0000036
has unit
My:timeof observation
sio:country
Spain
rdfs:label
has identifier
Measured at
my:Parasite Species
http://data.food.gov.uk/codes/foodtype/hierarchy/facet/source/_A059G
https://en.wikipedia.org/wiki/ISO_3166-2:ES
Has participant
Maiz
is located in
my:Infection
has participant
my:Observation
sio:pathogen
#value
rdfs:label
is location of
lat
2 fields: isogenic and transgenic maize. Data pooled during data entry.
my:GeoLocation
sio:measuring
long
EFO Parasitic infection
#value
geo:Point

We know the ontology terms

sio:identifier	http://semanticscience.org/resource/
UO_0000036	http://purl.obolibrary.org/obo/uo.owl#
EFO Parasitic infection	http://www.ebi.ac.uk/efo/efo.owl#
geo:Point	http://www.w3.org/2003/01/geo/wgs84_pos#

What is “my”?



We are now at the point where we have to make decisions about how to “mint” our own identifiers.

Remember - the identifiers should resolve, so we will base them on the URL of a server/folder that we have access to.

I’ll show you a useful pattern that can be used in many scenarios (even Zenodo uploads!)

Our server:

http://training.fairdata.solutions/DAV/home/LDP/gofair/

SpeciesAbundancePub2015.tsv

Open with Goog

A	B	C	D	E	F
observationID	cropCode	cropName	speciesID	countryCode	countryText
2147365908	A059G	Maize	290307251	ES	Spain
2147365909	A059G	Maize	290307131	ES	Spain

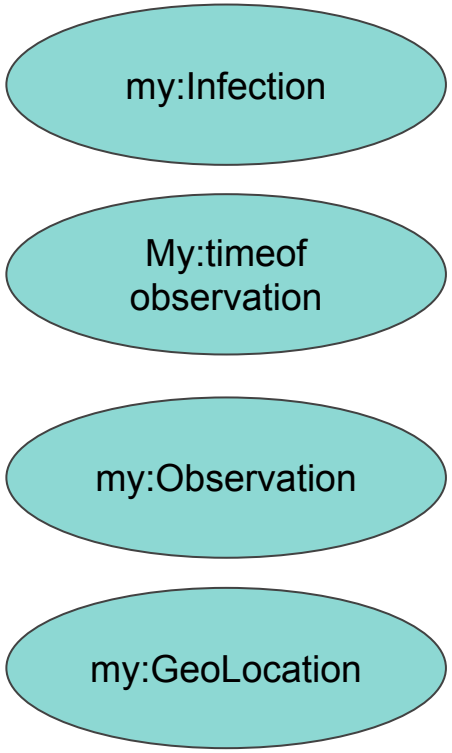
Because “observations” are the central part of our model, we will simply use these as the Record IDs

http://training.fairdata.solutions/DAV/home/LDP/gofair/obs_2147365908

```
@prefix train <http://training.fairdata.solutions/DAV/home/LDP/gofair/> .

train:obs_2147365908
```

Now use “document fragments” to identify nodes



train:obs_2147365908#infection

train:obs_2147365908#time

train:obs_2147365908#observation

train:obs_2147365908#location

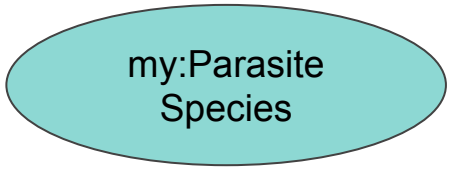
http://training.fairdata.solutions/DAV/home/LDP/gofair/obs_2147365908#location

Species is a “key” in the data.

I *could* model it in exactly the same way as the previous four

However, that results in duplicate information in the database

So we will model it as its own record



train:species_{speciesID}

http://training.fairdata.solutions/DAV/home/LDP/gofair/species_123456

Because servers download whatever is before the

It is possible to use this pattern for a wide range of scenarios.

E.g. in Zenodo, you can reserve a DOI before you publish your data

That DOI becomes part of a predictable URL

E.g. 10.5281/zenodo.**1435421**

Files are:

https://zenodo.org/record/**1435421**/files/MY_FILENAME_HERE

So you can create RDF with document fragments:

https://zenodo.org/record/**1435421**/files/MY_FILENAME_HERE#SomeNode