

Databases

Graph Databases Motivation

André Santanchè e Patrícia Cavoto
Institute of Computing - UNICAMP
August 2015

Graph Database

Graph as the basic data model of a database

Why Graphs?

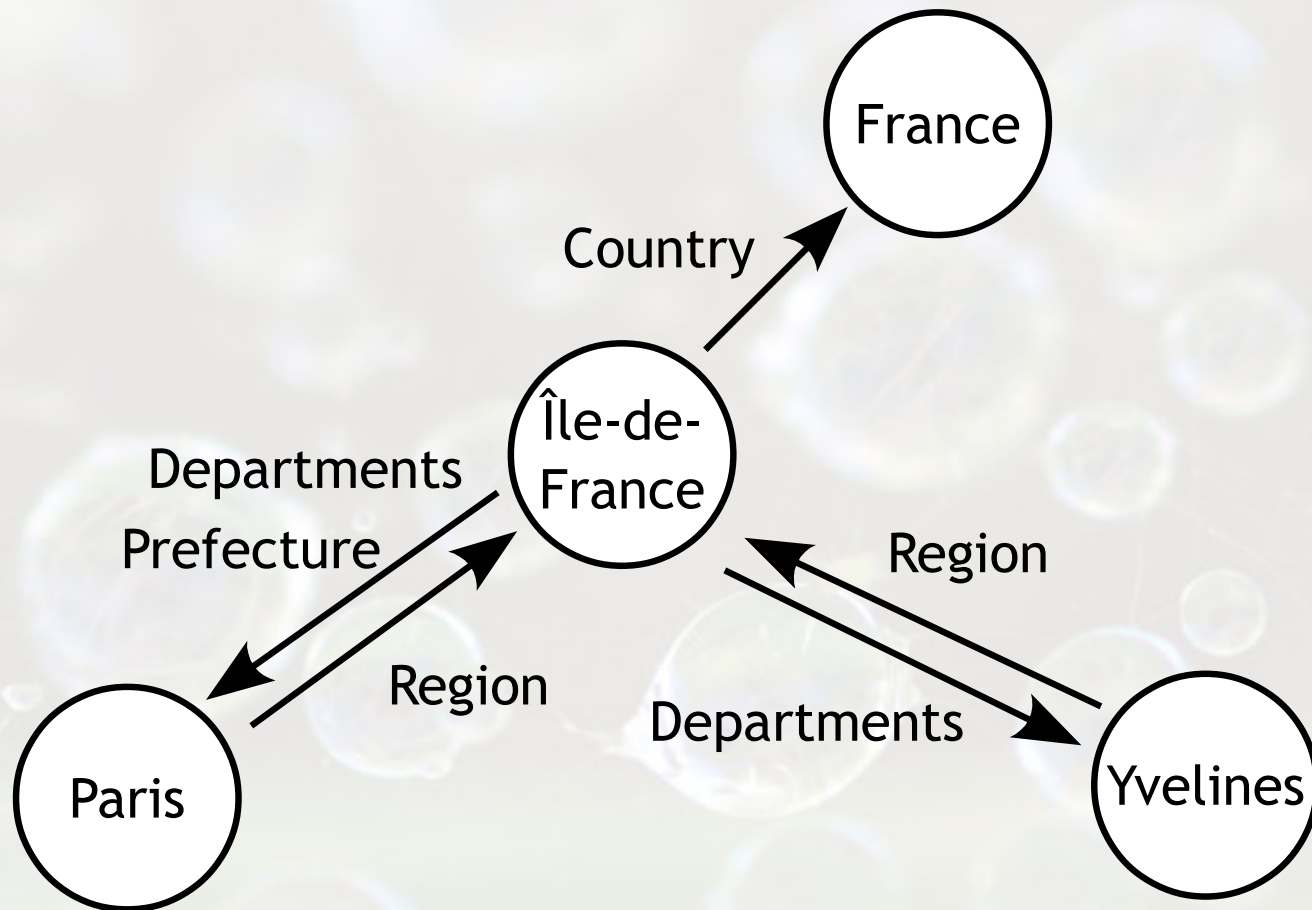
Why Graphs?

- The Web Effect
 - Linked Data
 - Social Networks and Social Content
- Sharing and interconnecting
- Complex Networks

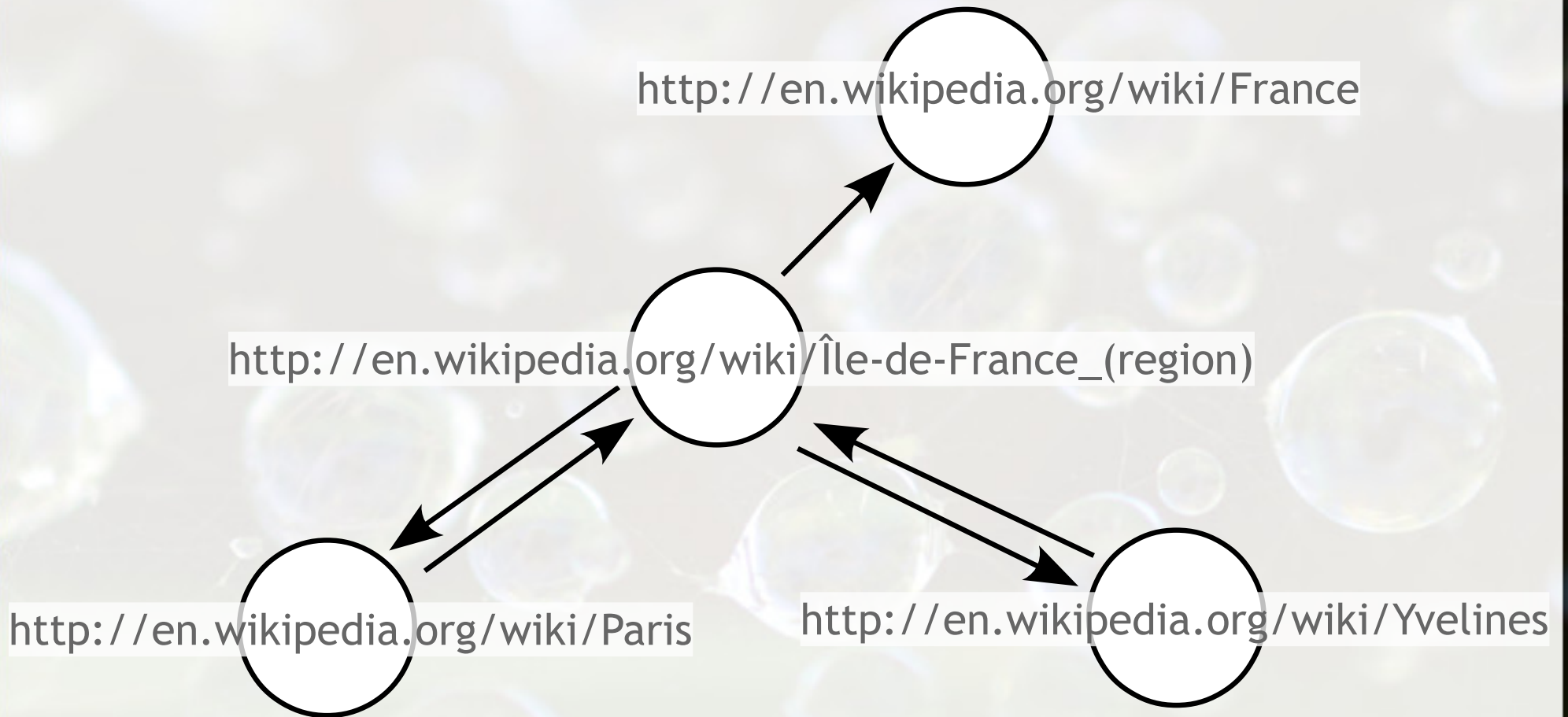


Why Graphs?
Linked Data

DBPedia



DBPedia (URIs)



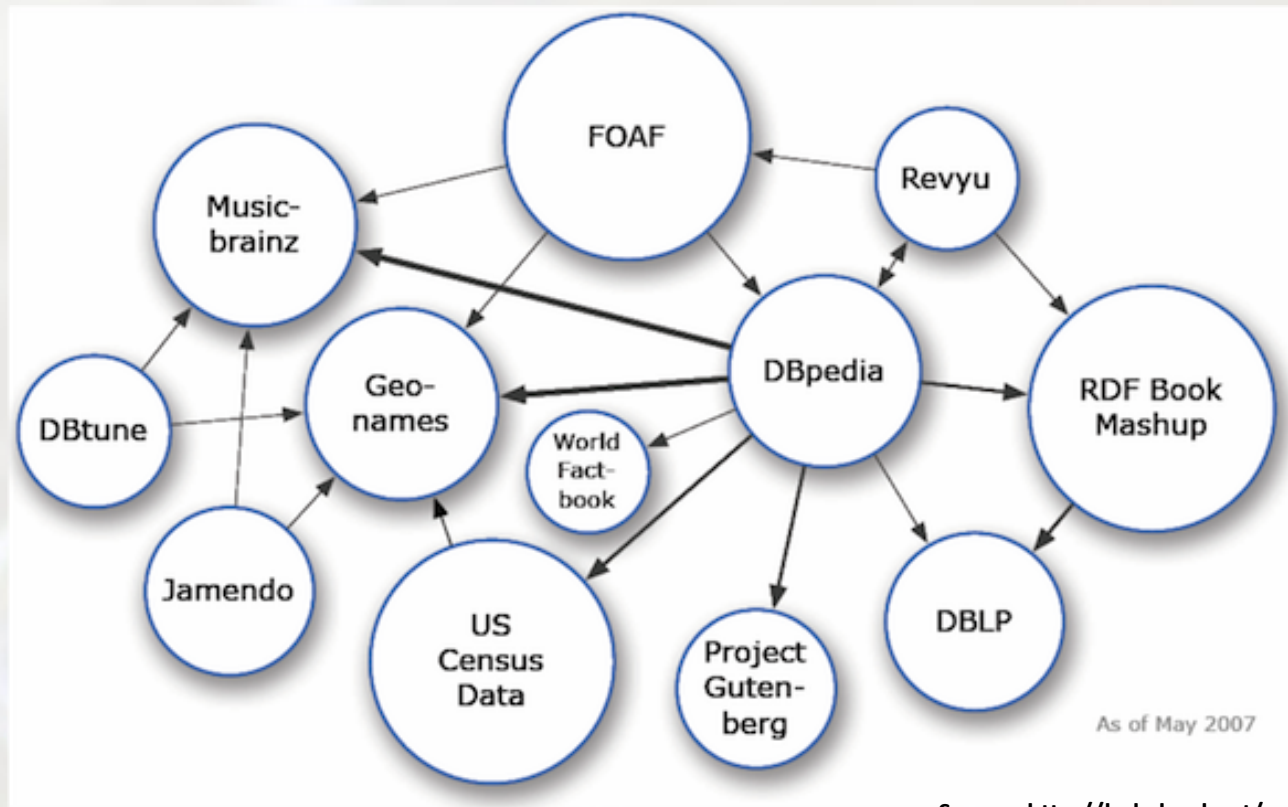
DBPedia - English

- **4 million things**
- **3.22 million classified in a consistent ontology**
 - 832,000 persons
 - 639,000 places (427,000 populated)
 - 372,000 creative works
 - 116,000 music albums; 78,000 films; 18,500 video games
 - 209,000 organizations
 - 226,000 species
 - 5,600 diseases.

DBpedia - International

- 119 languages
- 24.9 million things
- 16.8 million interlinked with English
- 12.6 million unique things

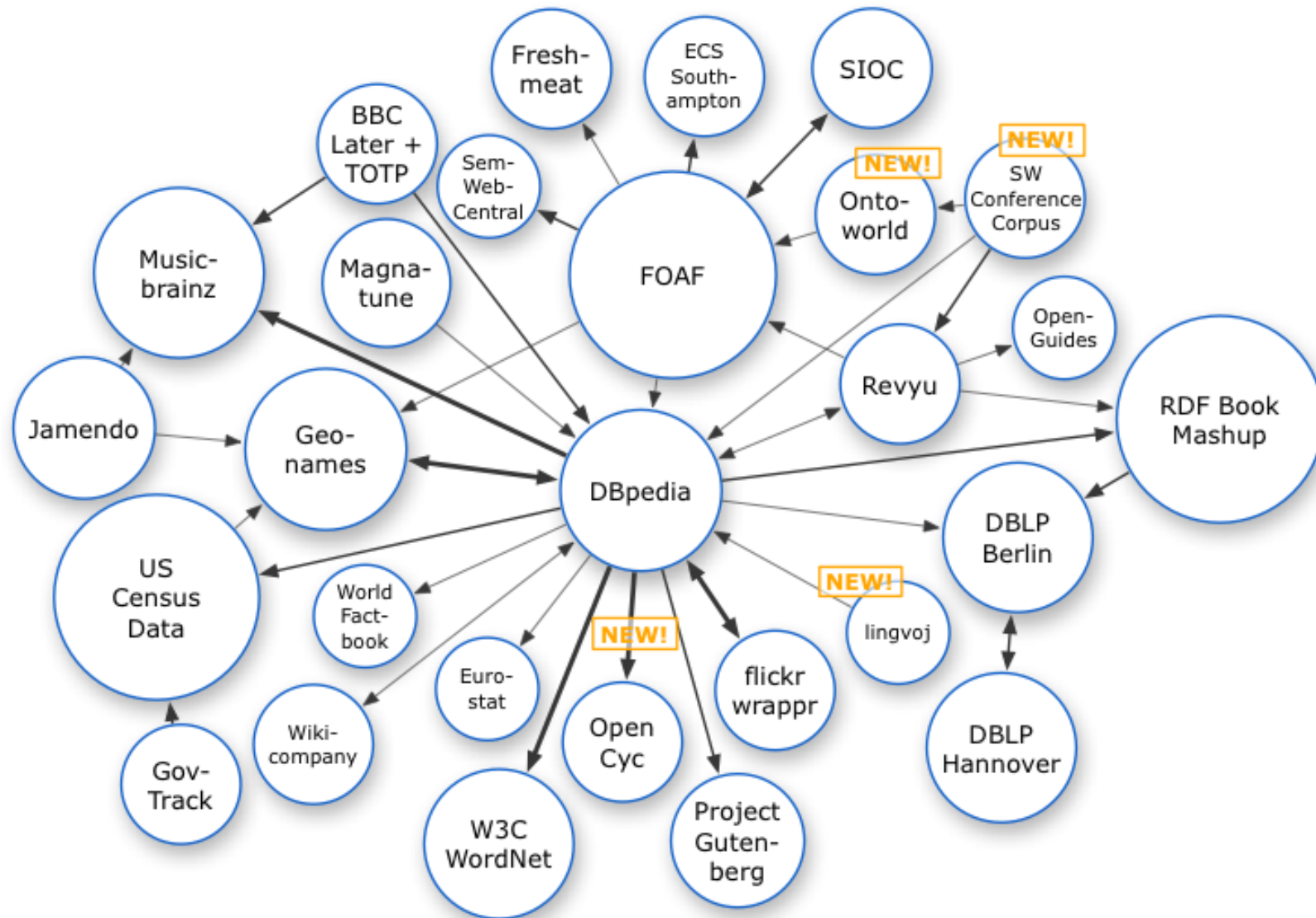
Linked Data



Source: <http://lod-cloud.net/>

Datasets published following Linked Data 'format': **05/2007**

Linked Data

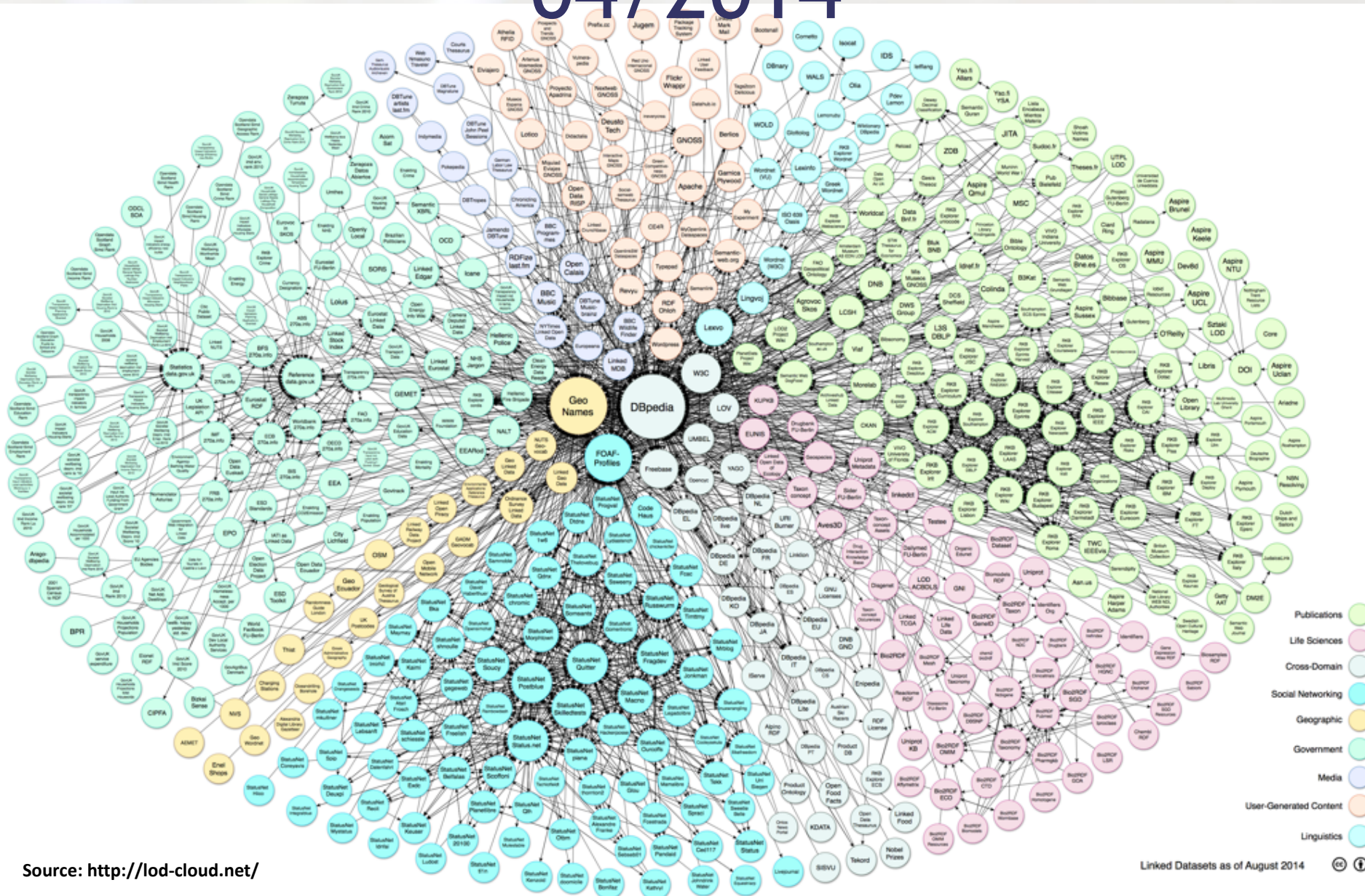


Source: <http://lod-cloud.net/>

Datasets published following Linked Data 'format': **11/2007**

Linked Data

04/2014



Source: <http://lod-cloud.net/>

Linked Datasets as of August 2014 

Why Graphs?

Social Networks and Social Content

Searching Pet

Pet Shop Boys

Anirudh Koul



pet shop boys, pet, concert, tour



cat, kittens, eyes,
ears, pet, animal



dog, pet, animal,
funny, glasses



dog, pet, alaskan malamute



cat, kitty, eyes,
pretty



cat, kitten, garden,
pet



recycle, pet, plastic bottle,
polyethylene terephthalate



recycle, pet, plastic bottle,
polyethylene terephthalate



wine, pet, bottle

sfroehlich1121



cat, kittens, eyes,
ears, pet, animal

shorty_nz_2000



dog, pet, animal,
funny, glasses

Edward Corpuz



dog, pet, alaskan malamute

sfroehlich1121



cat, kitty, eyes,
pretty

Jay Woodworth



cat, kitten, garden,
pet

Nemo's great uncle



recycle, pet, plastic bottle,
polyethylene terephthalate

Karl Baron



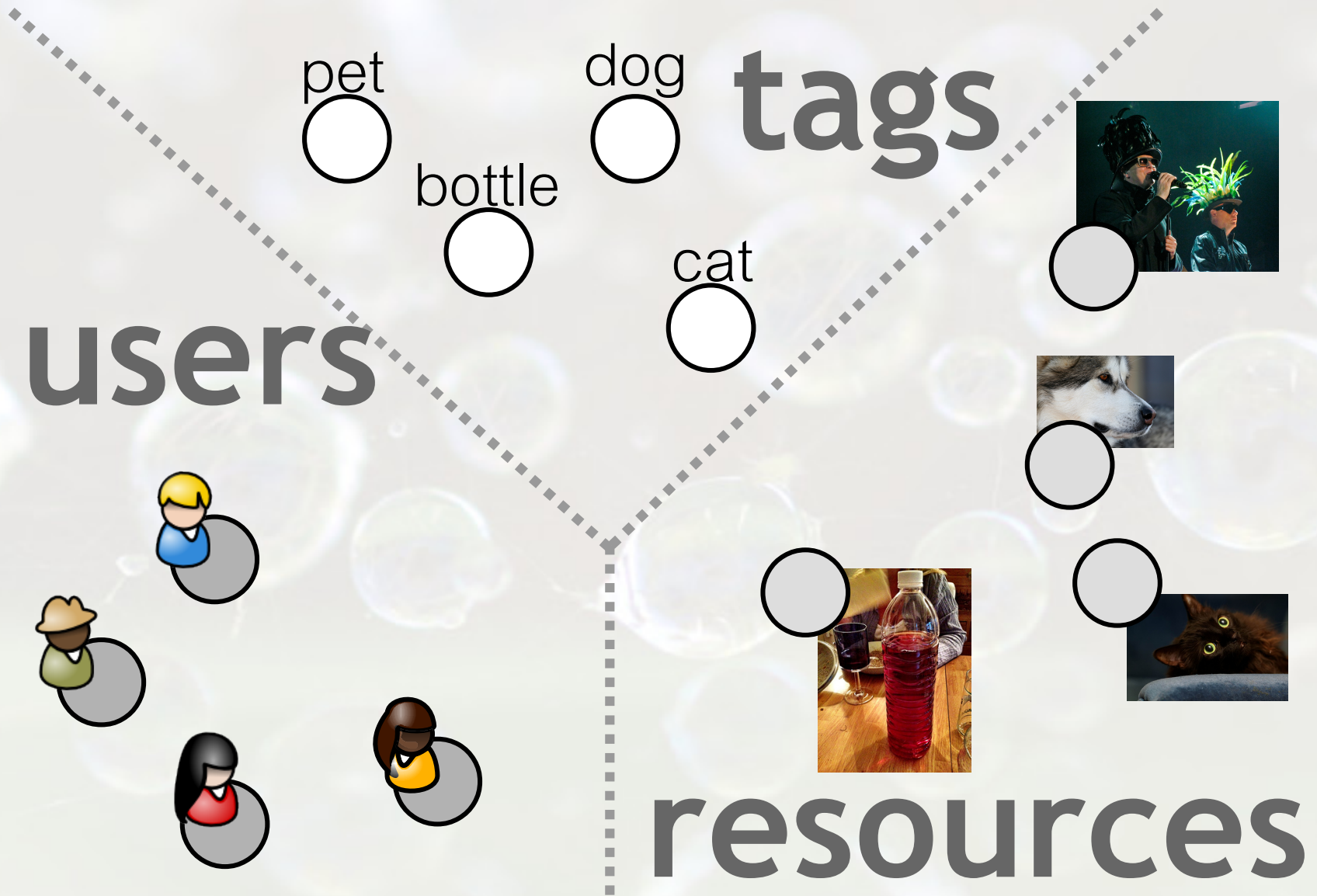
wine, pet, bottle

FaceMePLS

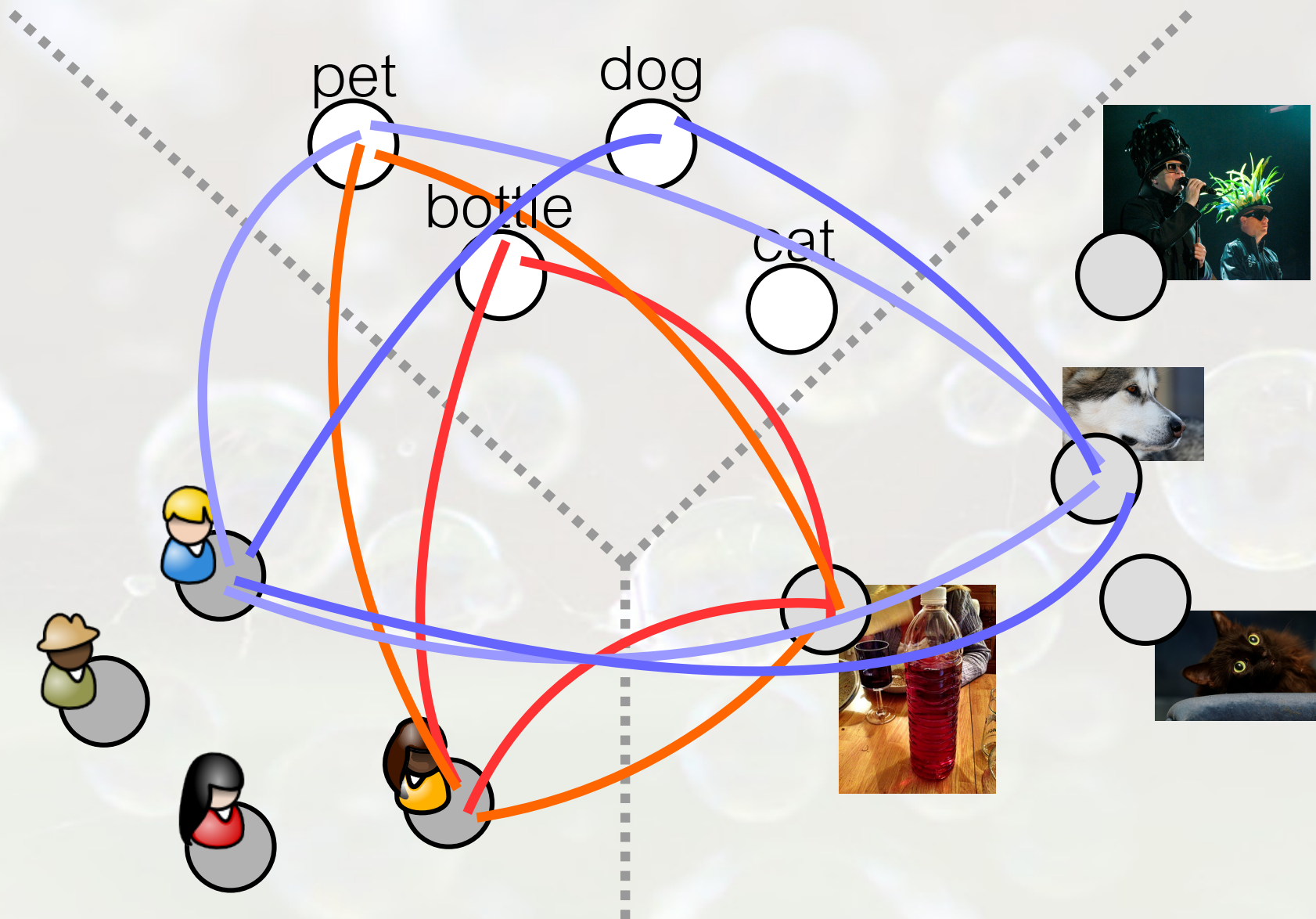


recycle, pet, plastic bottle,
polyethylene terephthalate

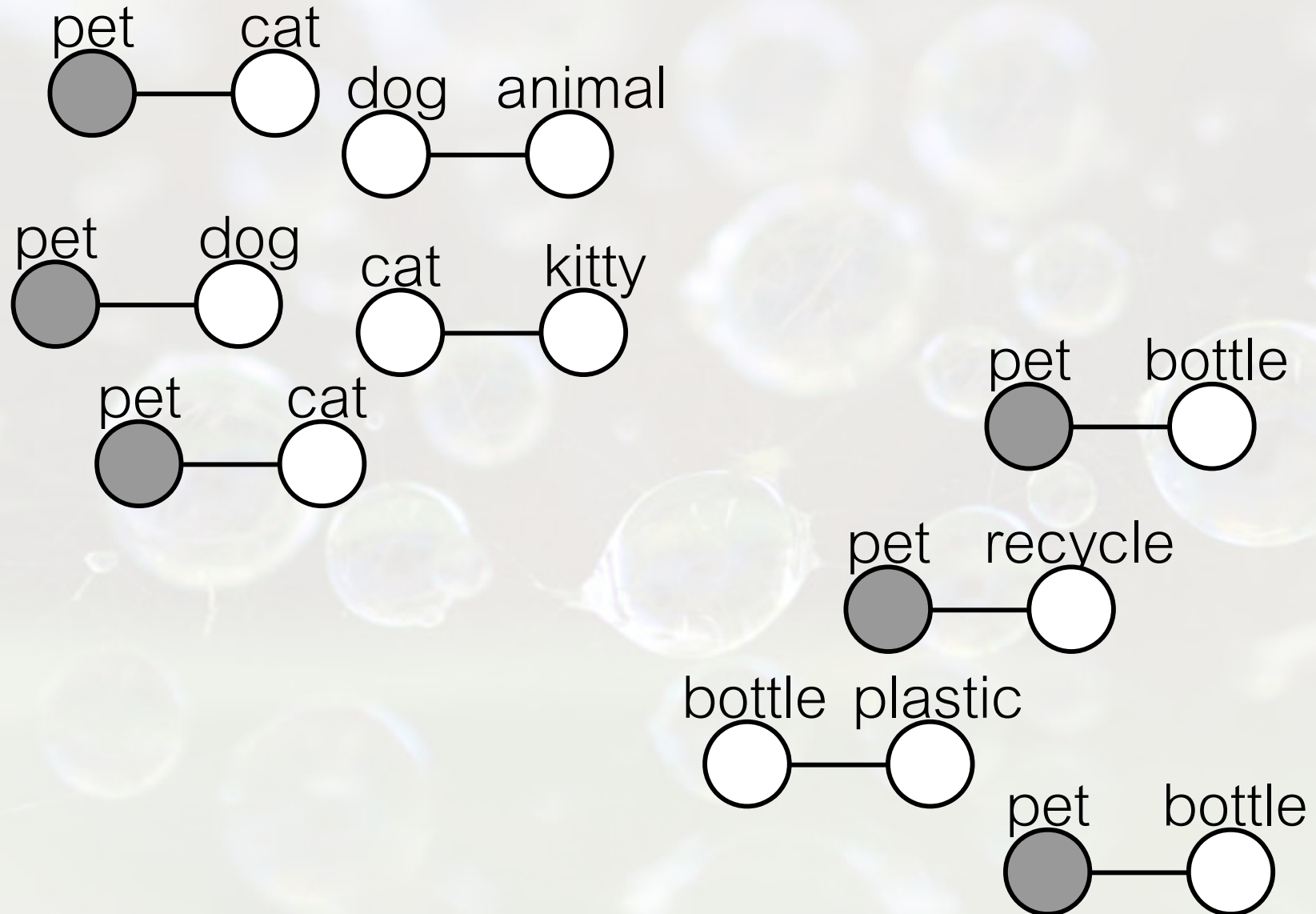
Graph users/tags/resources



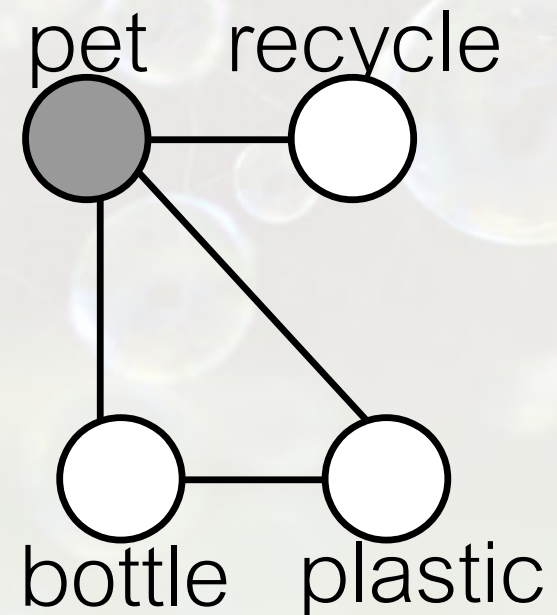
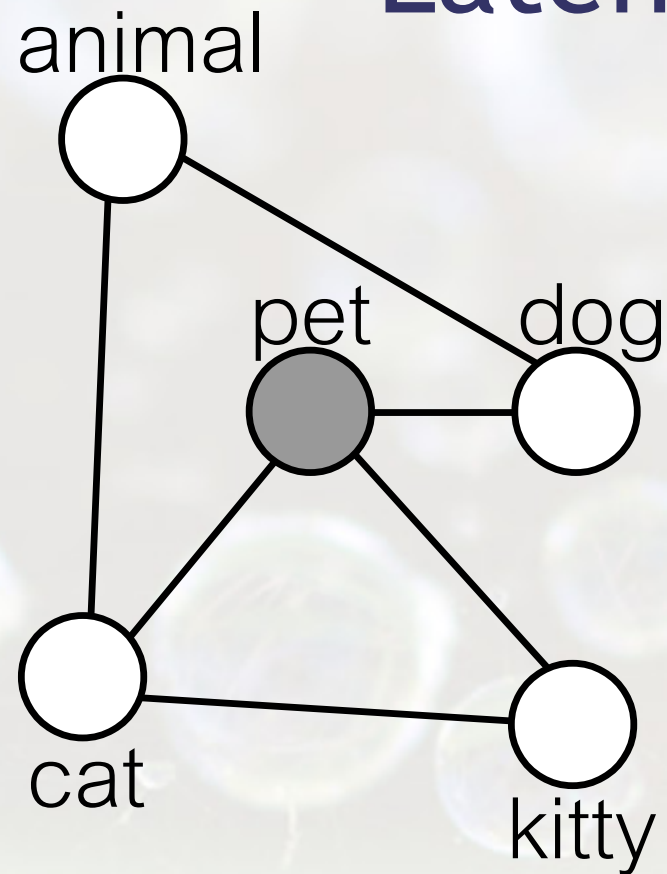
Graph users/tags/resources



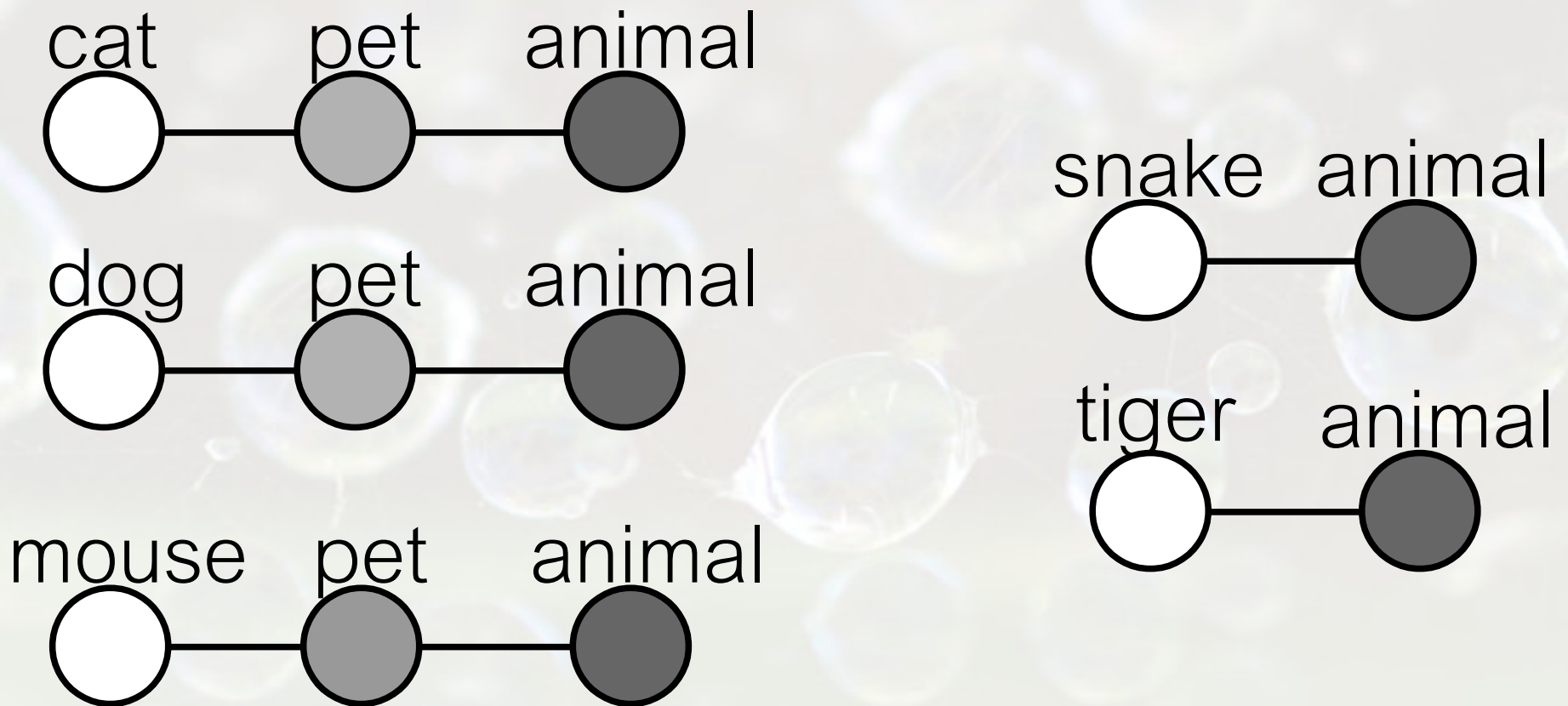
Co-occurrences and Latent Semantics



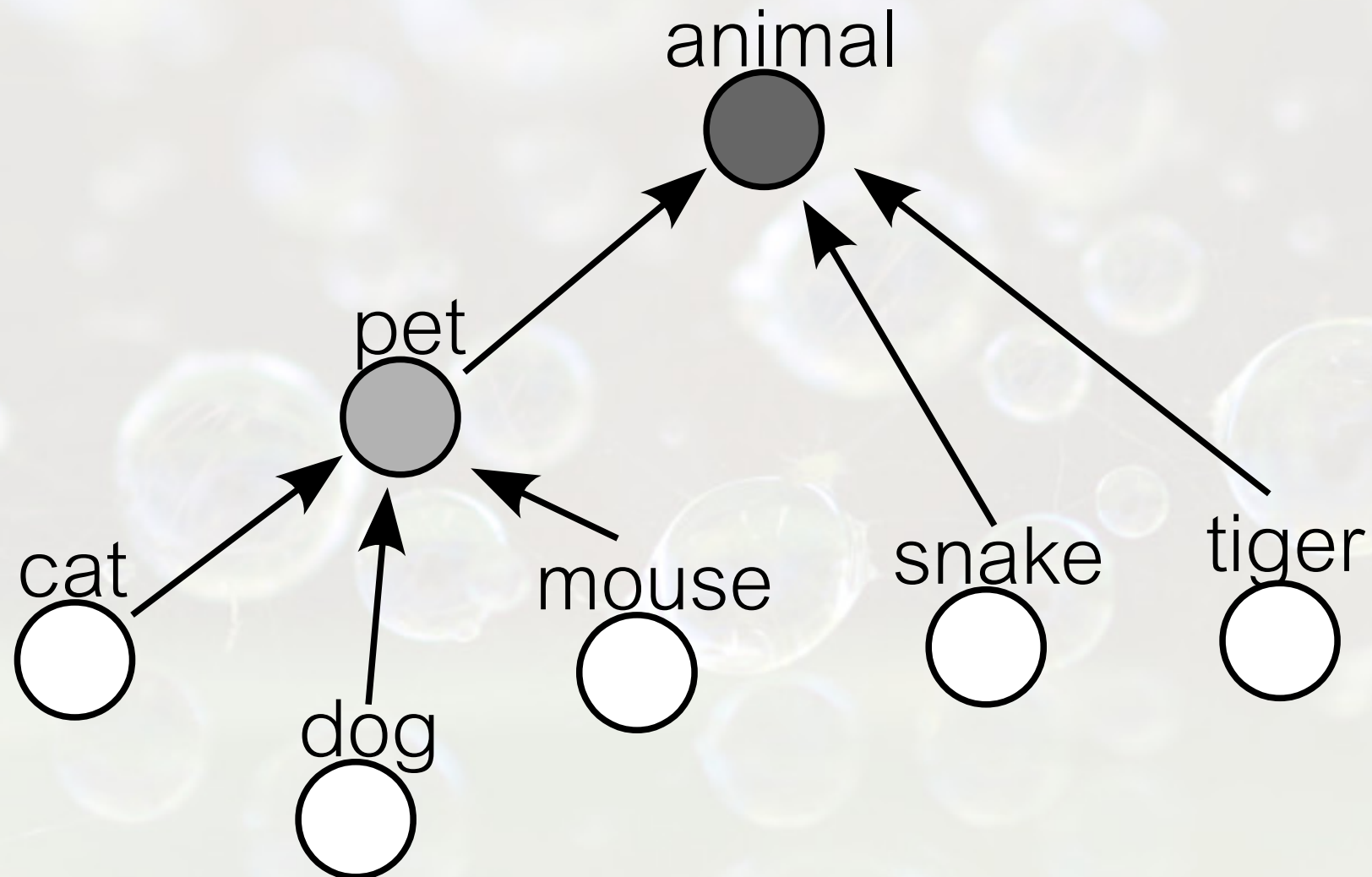
Co-occurrences and Latent Semantics



Co-occurrences and Latent Semantics



Co-occurrences and Latent Semantics



Social Effect

Suggested/Reinforced Tags

sfroehlich1121



black, cat, kitty, katze, long, hair, blue, eyes, pretty,
canon, t1i, 500d, ef 100mm f/2.8 usm macro

Social Effect

Suggested/Reinforced Tags

sfroehlich1121



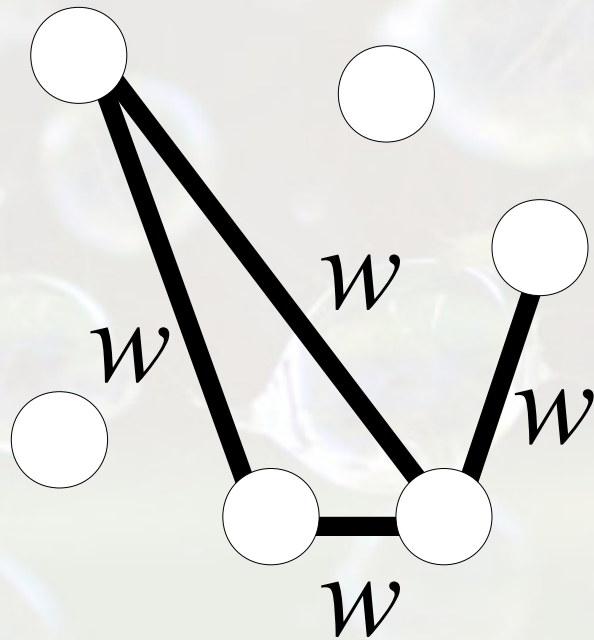
black, **cat**, **kitty**, katze, long,

hair, blue, eyes, **pretty**,
canon, t1i, 500d, ef 100mm f/2.8 usm macro

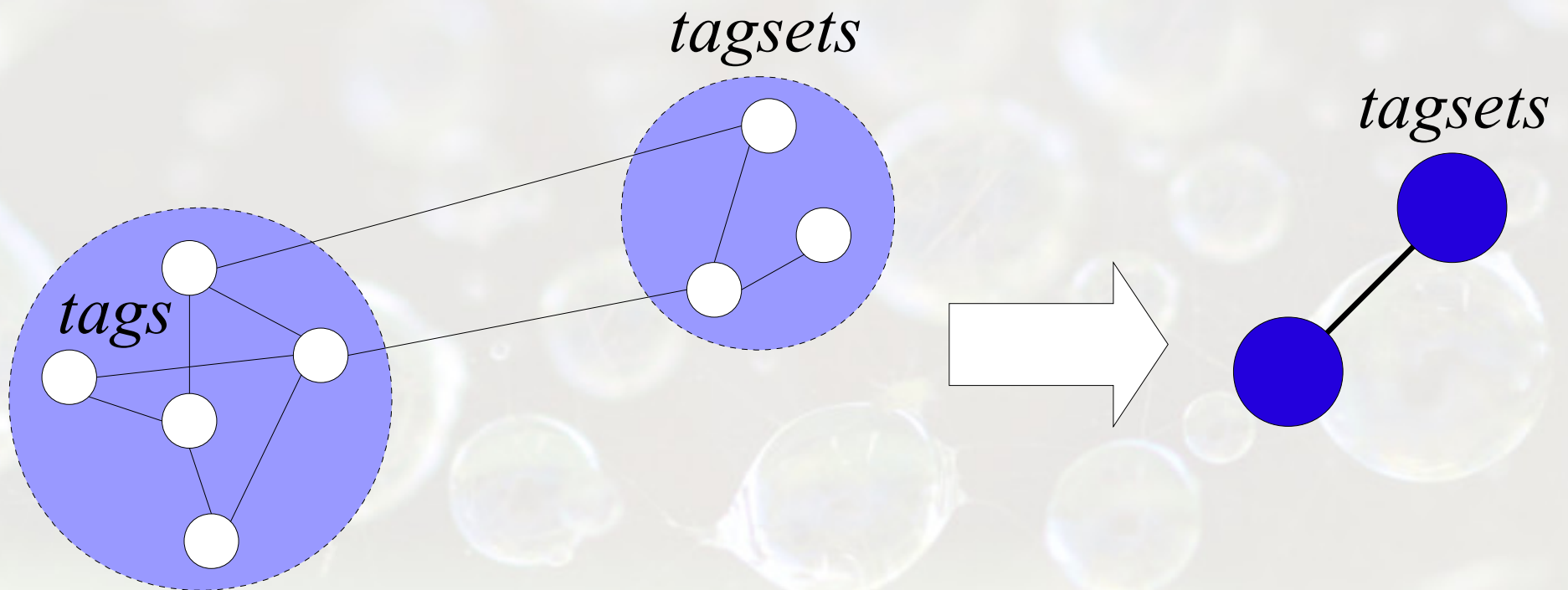
M1

co-occurrences in a folksonomy

tags / co-occurrences



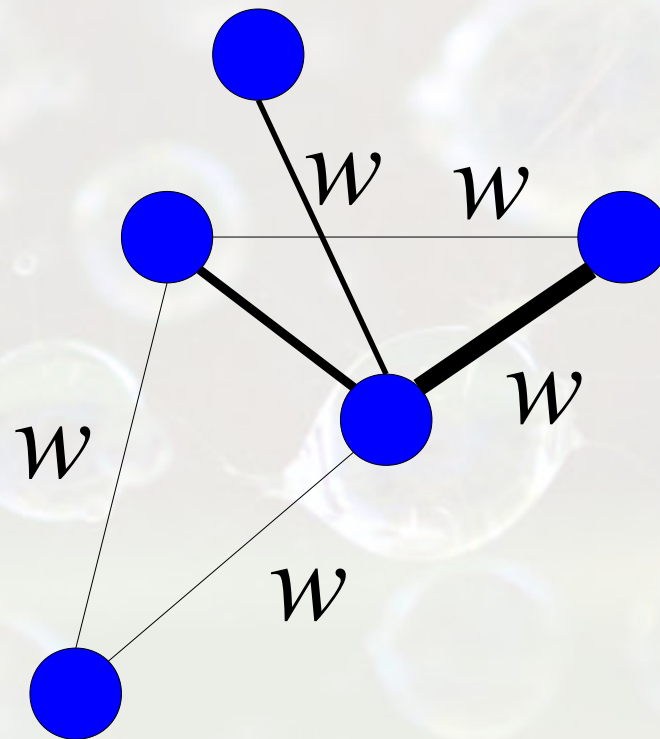
Tagsets



M2

tagsets and their co-occurrences

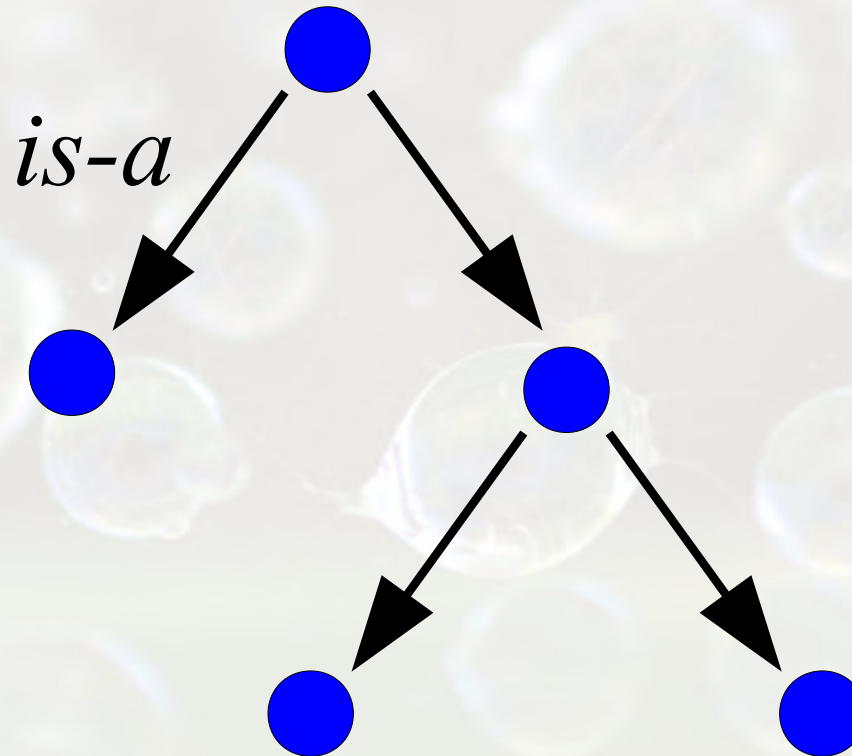
tagsets/co-occurrences



M4

Social Ontology

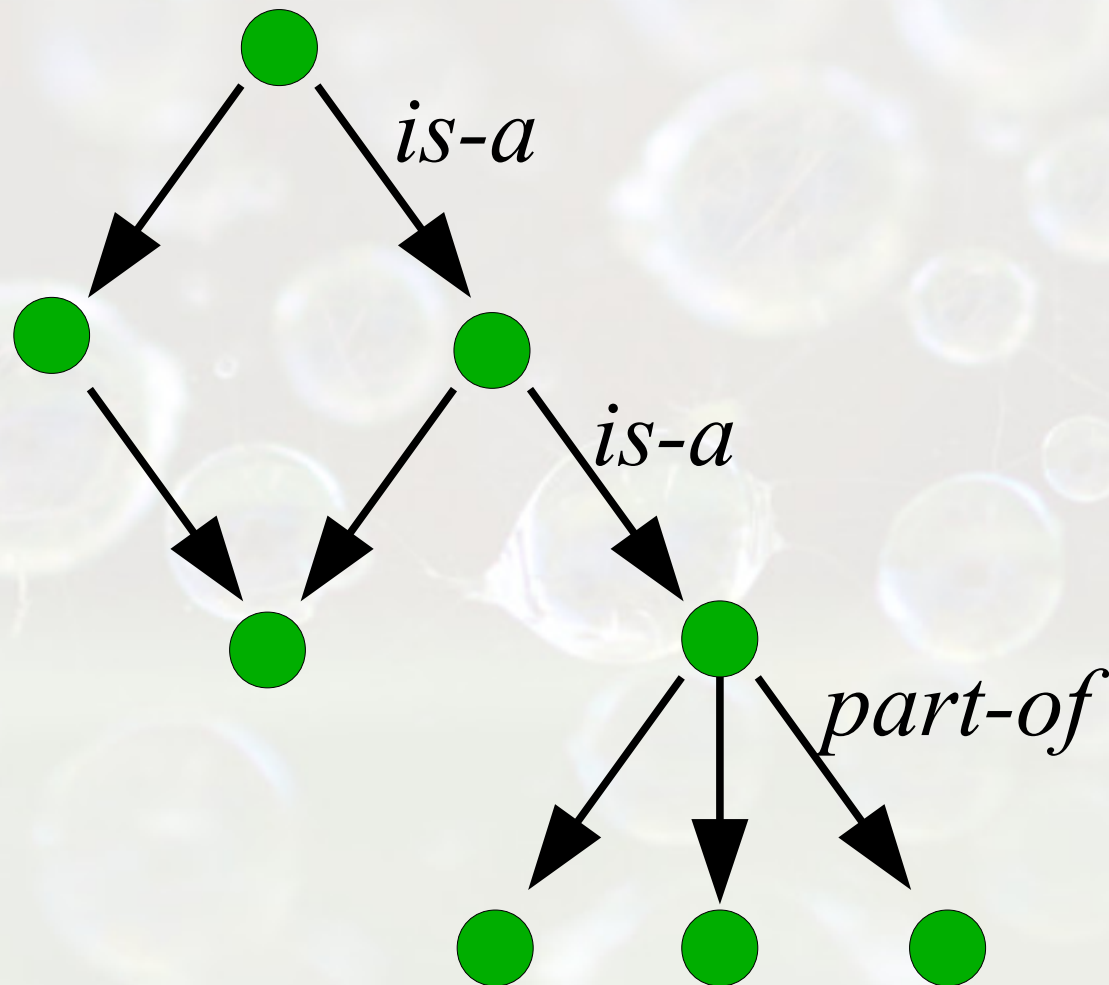
tagsets / typed relations



M3

simplified model of an ontology

concepts/relations



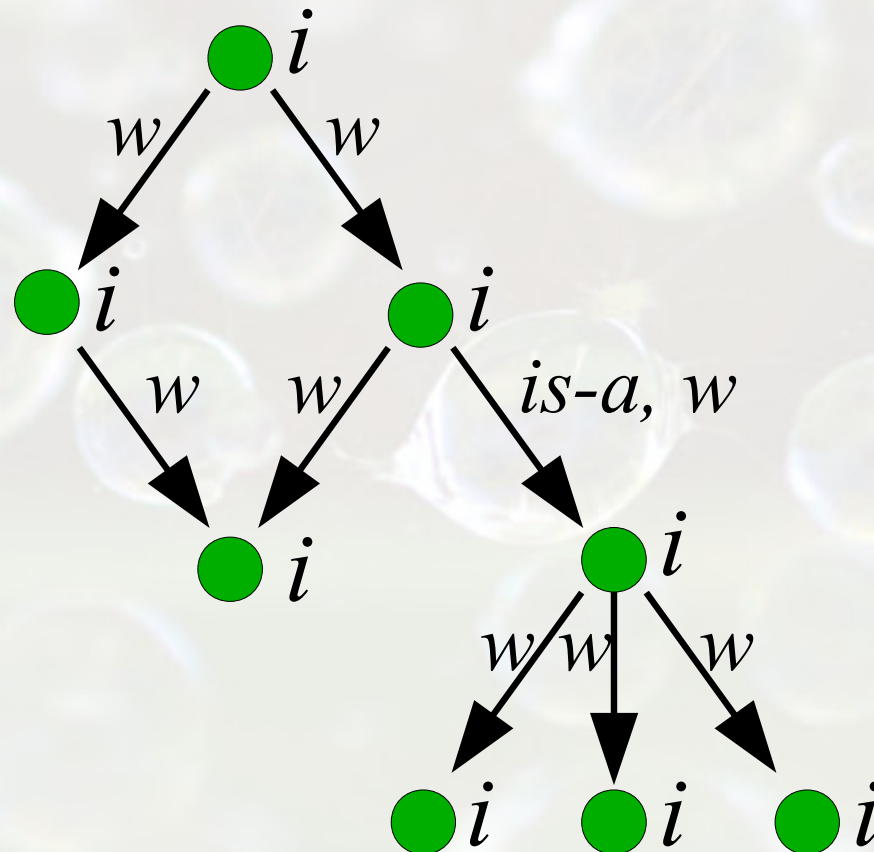
M5

Folksonomized Ontology (FO)

concepts / typed relations (ontology)

+


information content / co-occurrences (folksonomy)



**Meaning of Pet
needs
meaning of other tags?**

Matrix

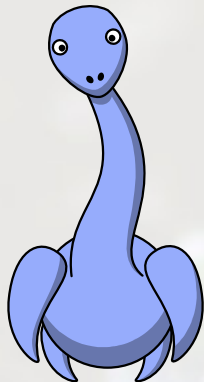
**What is the meaning of the word
Love?**

The background of the slide features a close-up photograph of water droplets on a green leaf. The droplets are of various sizes and are in sharp focus, with some showing internal reflections. The leaf's surface is visible, showing fine veins and a slightly textured appearance. The overall color palette is dominated by greens and blues, with the white highlights of the water droplets.

Why Graphs?
Sharing and Interconnecting

Models to Describe

Describing Prehistoric Animals



MNHN A. C. 8592

Is a	Plesiosaurus dolichodeirus	
Origin	Lyme Regis	England
Recognized	1824	
Size	5	

Describing Prehistoric Animals



SIPB R 90		
Is a	Plesiosaurus dolichodeirus	
Origin	Lyme Regis	England
Recognized	1830	
Size	5	



STC223		
Is a	Plesiosaurus gurgitis	
Origin	St. Croix	Switzerland
Recognized	1964	
Size	3.5	



MNHN 1912.20		
Is a	Triceratops horridus	
Origin	Lance Creek	EUA
Recognized	1889	
Size	9	



Sue

FMNH PR2081		
Is a	Tyrannosaurus rex	
Origin	Hell Creek	EUA
Recognized	1990	
Size	12.3	

Table

Id	Is a	Origin Place	Origin Country	Recognized	Size
MNHN A. C. 8592	Plesiosaurus dolichodeirus	Lyme Regis	England	1824	5
SIPB R 90	Plesiosaurus dolichodeirus	Lyme Regis	England	1830	5
STC223	Plesiosaurus gurgitis	St. Croix	Switzerland	1964	3.5
MNHN 1912.20	Triceratops horridus	Lance Creek	EUA	1889	9
FMNH PR2081	Tyrannosaurus rex	Hell Creek	EUA	1990	12.3

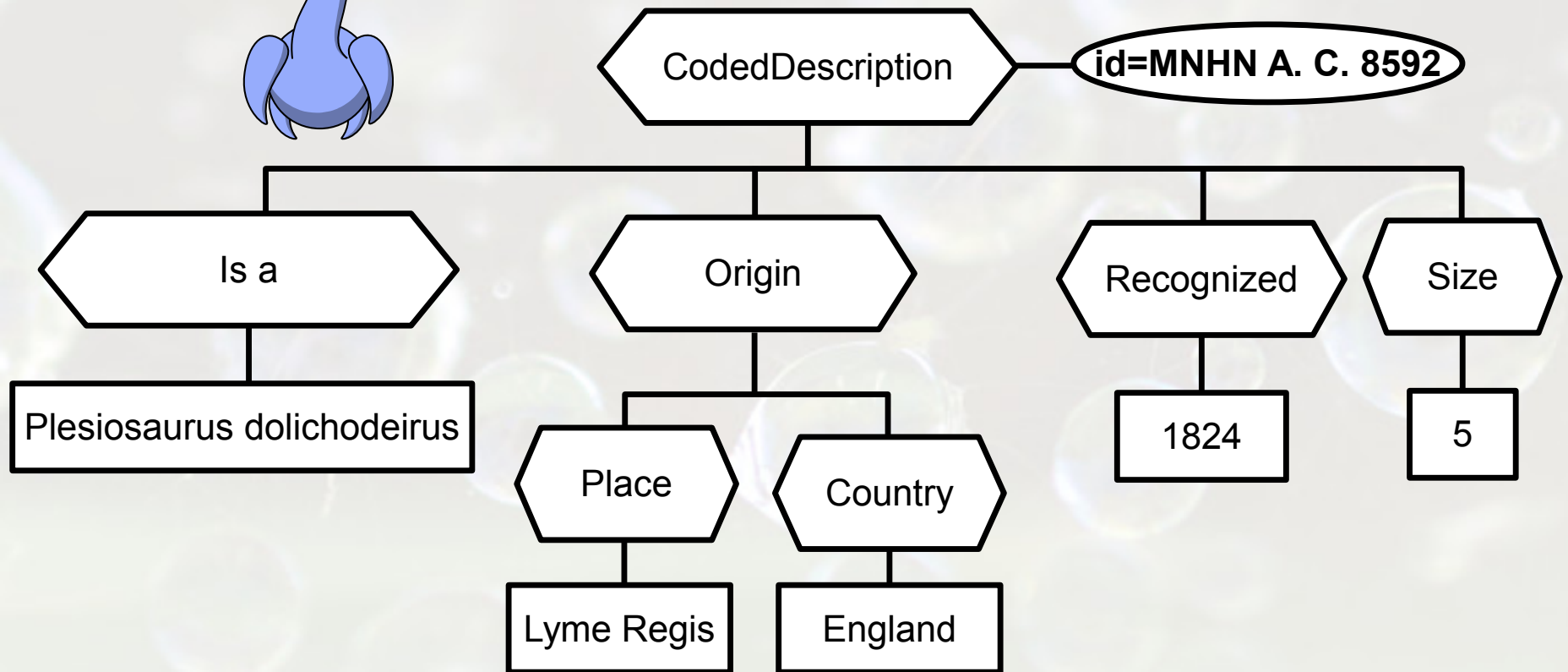
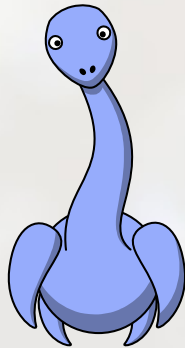


Table

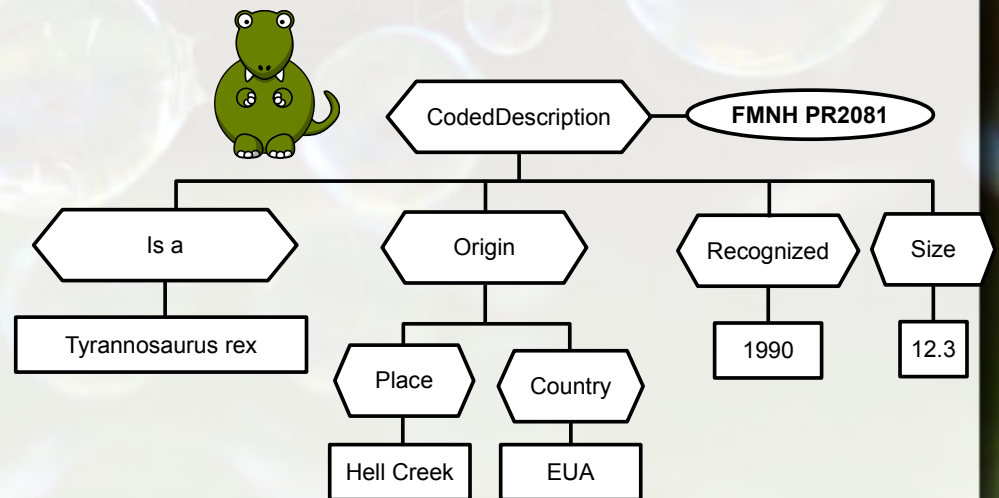
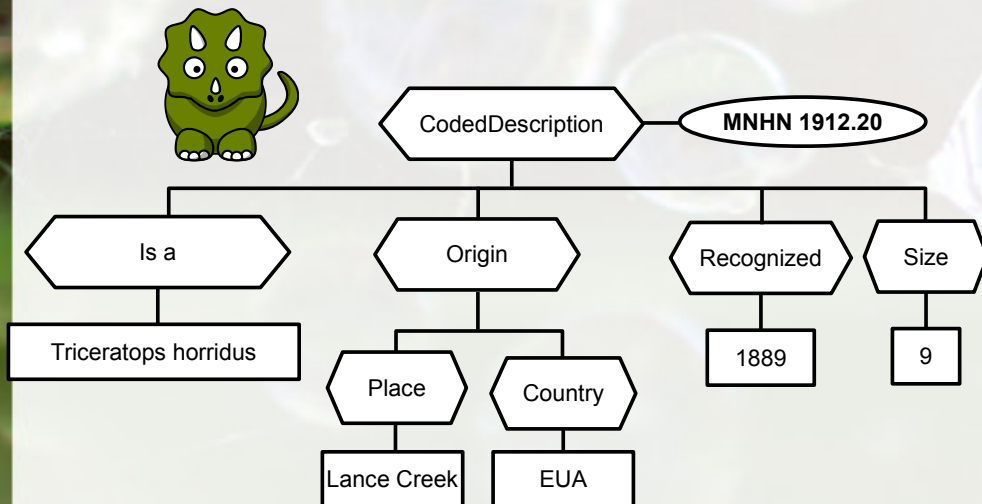
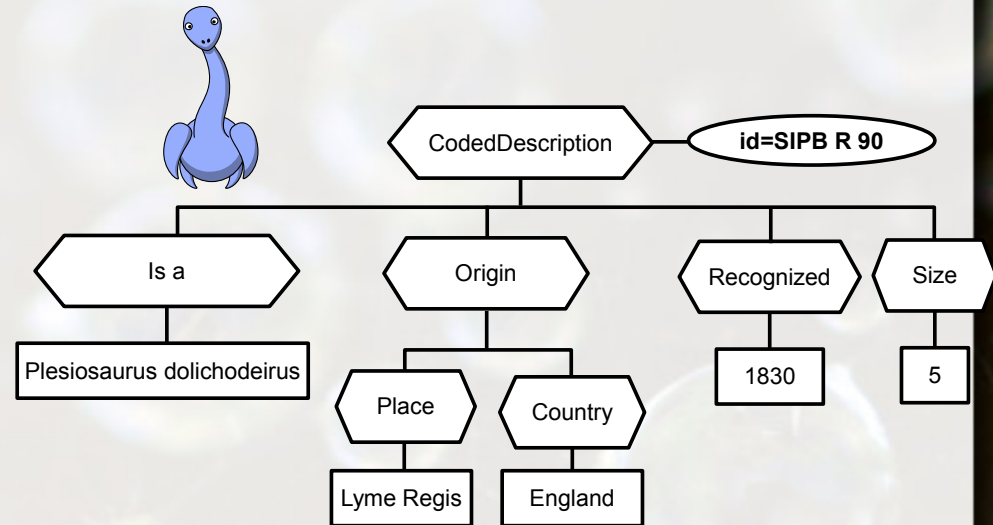
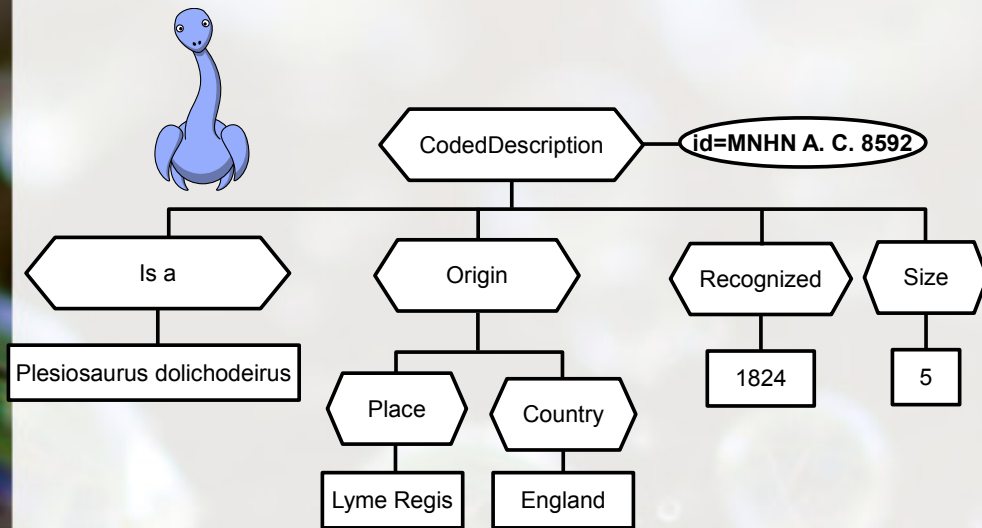
Excellent to Manage Data with
Predictable Static Schema

Sharing?

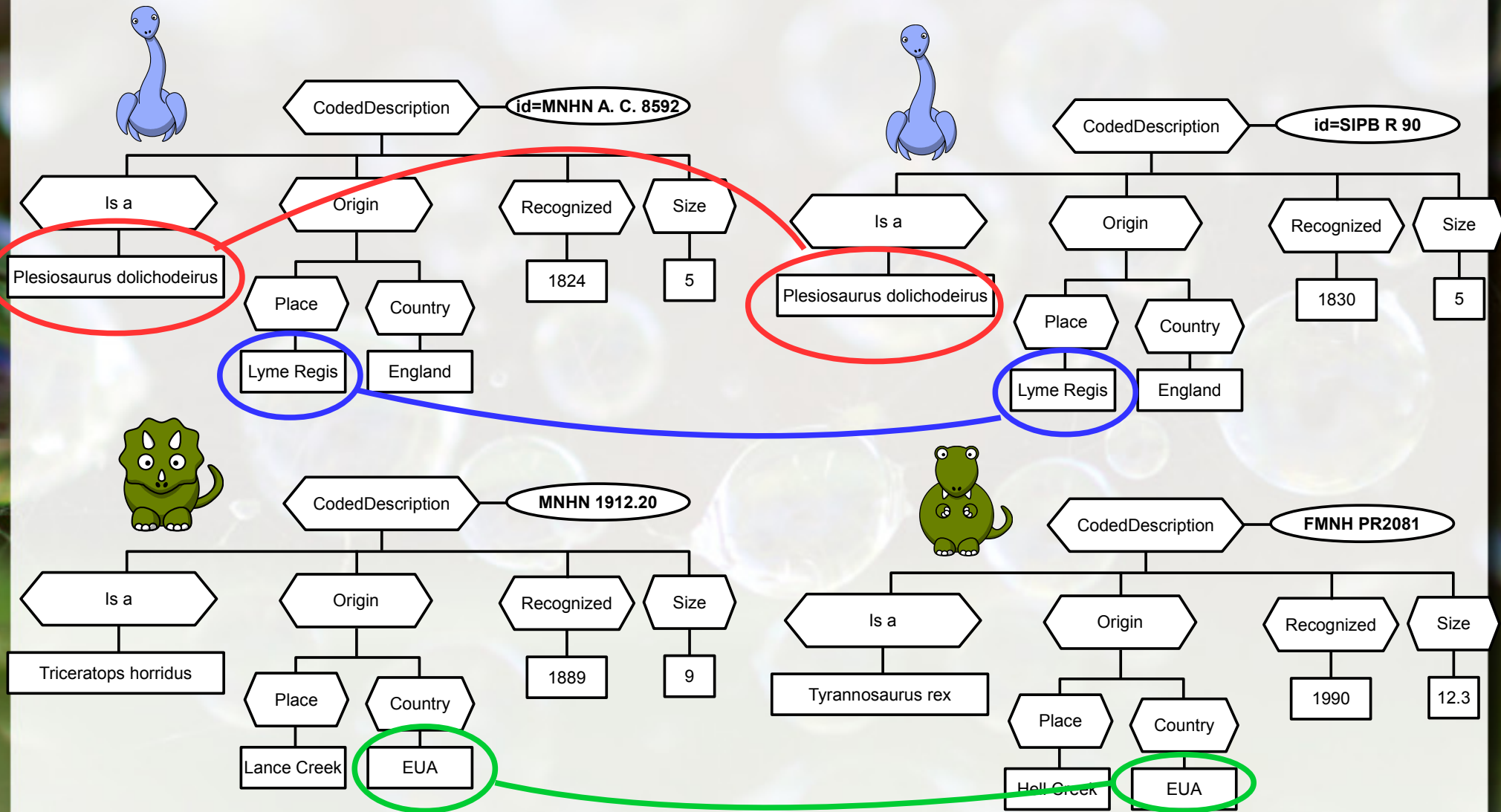
Documents and XML



Documents and XML



Documents and XML



How much hierarchical is your data?

Back to the Table

Le Muséum national d'Histoire naturelle

includes

Id	Is a	Origin Place	Origin Country	Recognized	Size
MNHN A. C. 8592	Plesiosaurus dolichodeirus	Lyme Regis	England	1824	5
SIPB R 90	Plesiosaurus dolichodeirus	Lyme Regis	England	1830	5
STC223	Plesiosaurus gurgitis	St. Croix	Switzerland	1964	3.5
MNHN 1912.20	Triceratops calicornis	Lance Creek	EUA	1888	9
MNHN 1912.20b	Triceratops horridus	Lance Creek	EUA	1889	9
FMNH PR2081	Tyrannosaurus rex	Hell Creek	EUA	1990	12.3

renamed

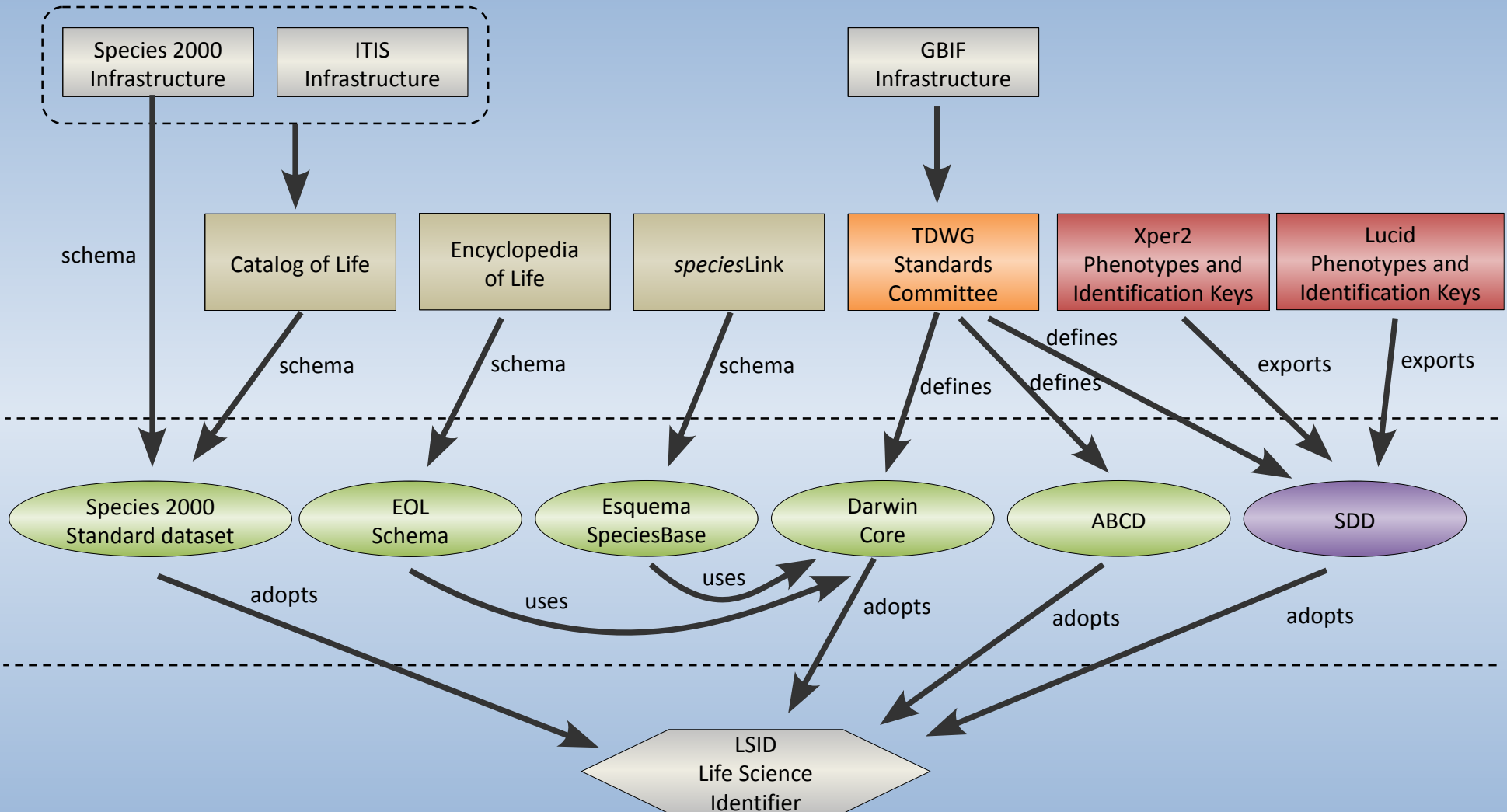
near



The tradeoff of Static Schemas

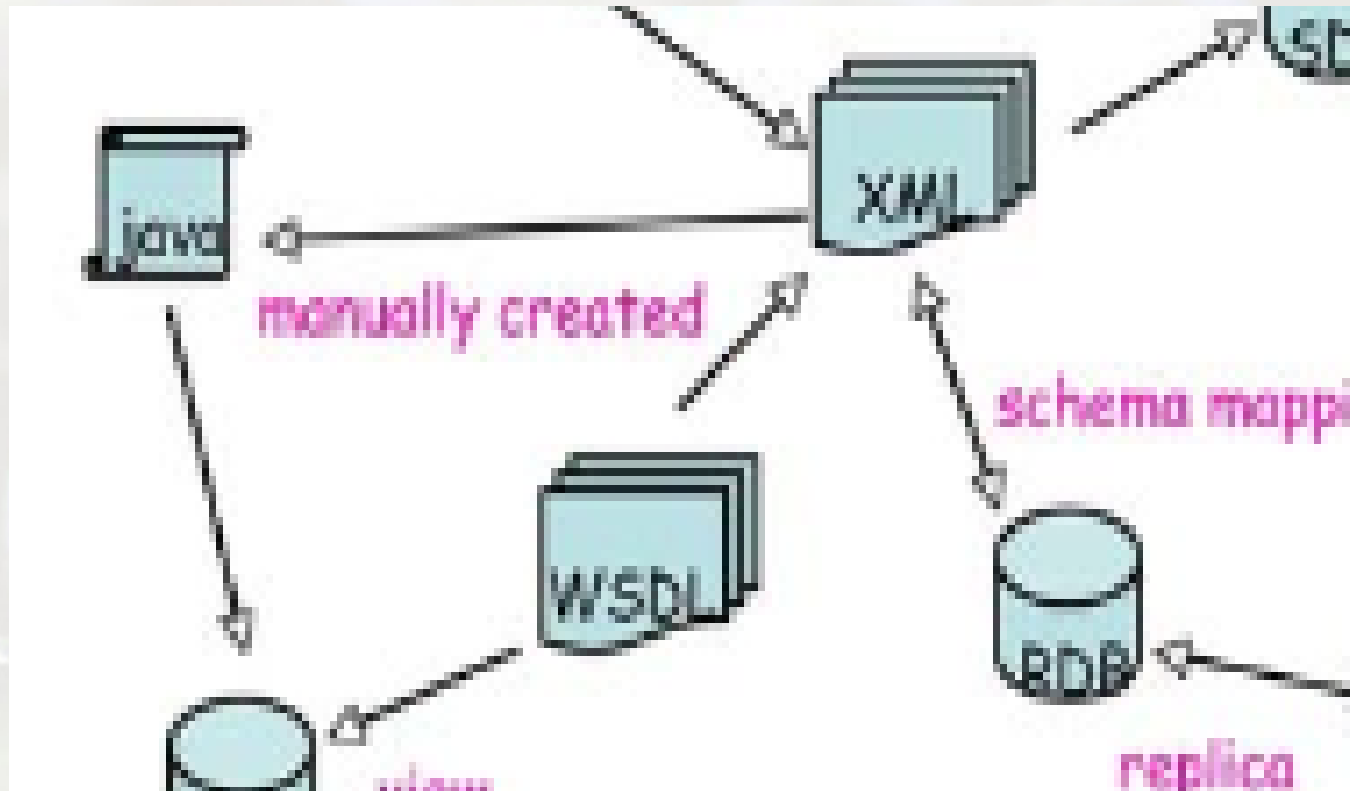
XML Schema

Not designed to mix



DataSpaces

Pay-as-you-go Integration



Franklin, M., Halevy, A., & Maier, D. (2005). From databases to dataspace: a new abstraction for information management. SIGMOD Rec., 34(4), 27–33.

DataSpaces and Linked Data

The background of the slide is a close-up photograph of a spider web. The web is composed of fine, silken threads that are slightly out of focus, creating a soft, ethereal texture. Numerous dew drops of various sizes are scattered across the web, some appearing as bright, clear spheres and others as more blurred, shimmering droplets. The lighting is soft and diffused, highlighting the intricate patterns of the web and the delicate nature of the dew.

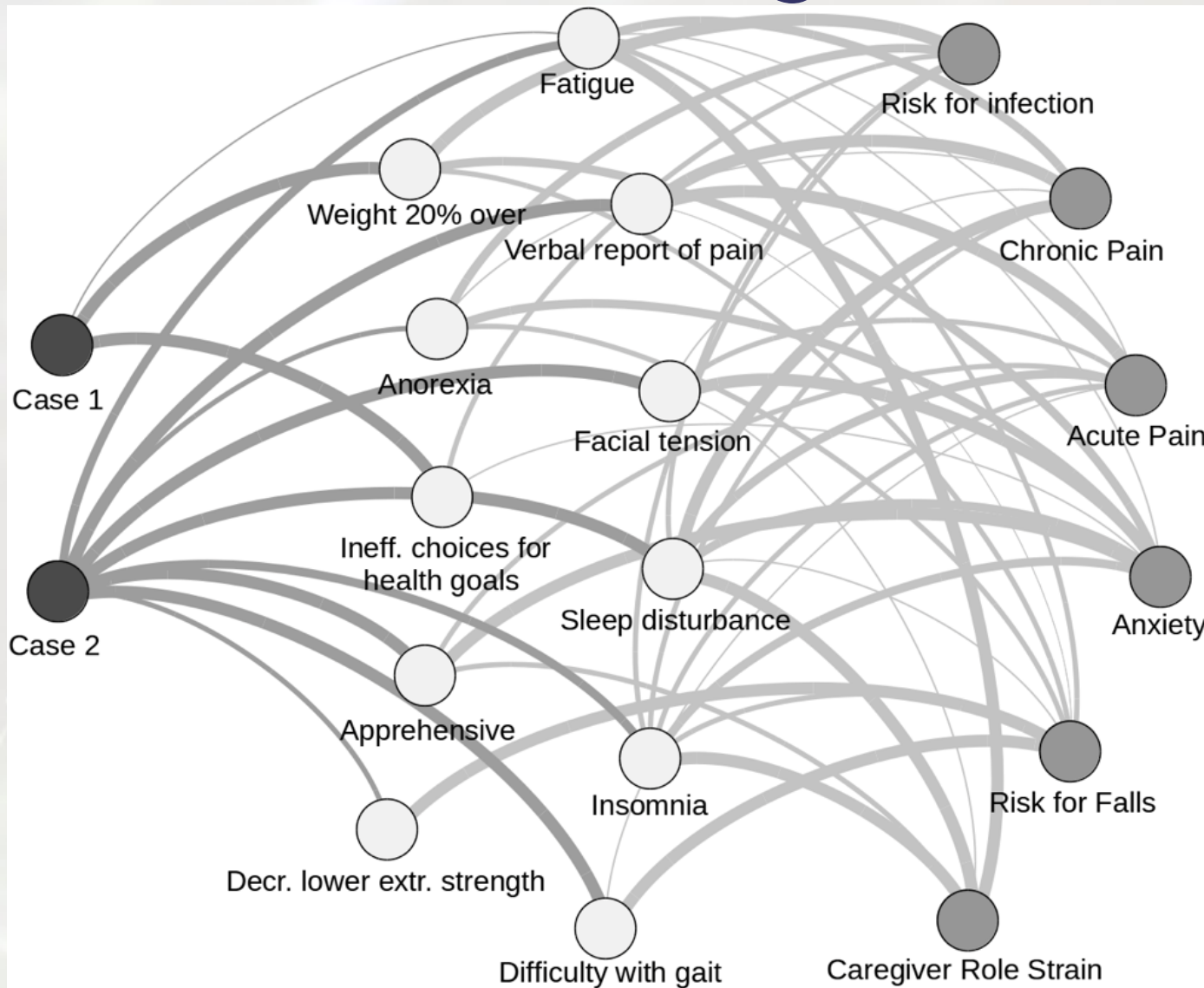
Why Graphs?
Complex Networks

Complex Networks

- Developed steadily since 1999
- Discrete systems are represented in terms of entities and relationships

(Luciano da F. Costa, 2013)

Patient x Diagnosis



(Gomes-Jr, 2013)

Complex Networks

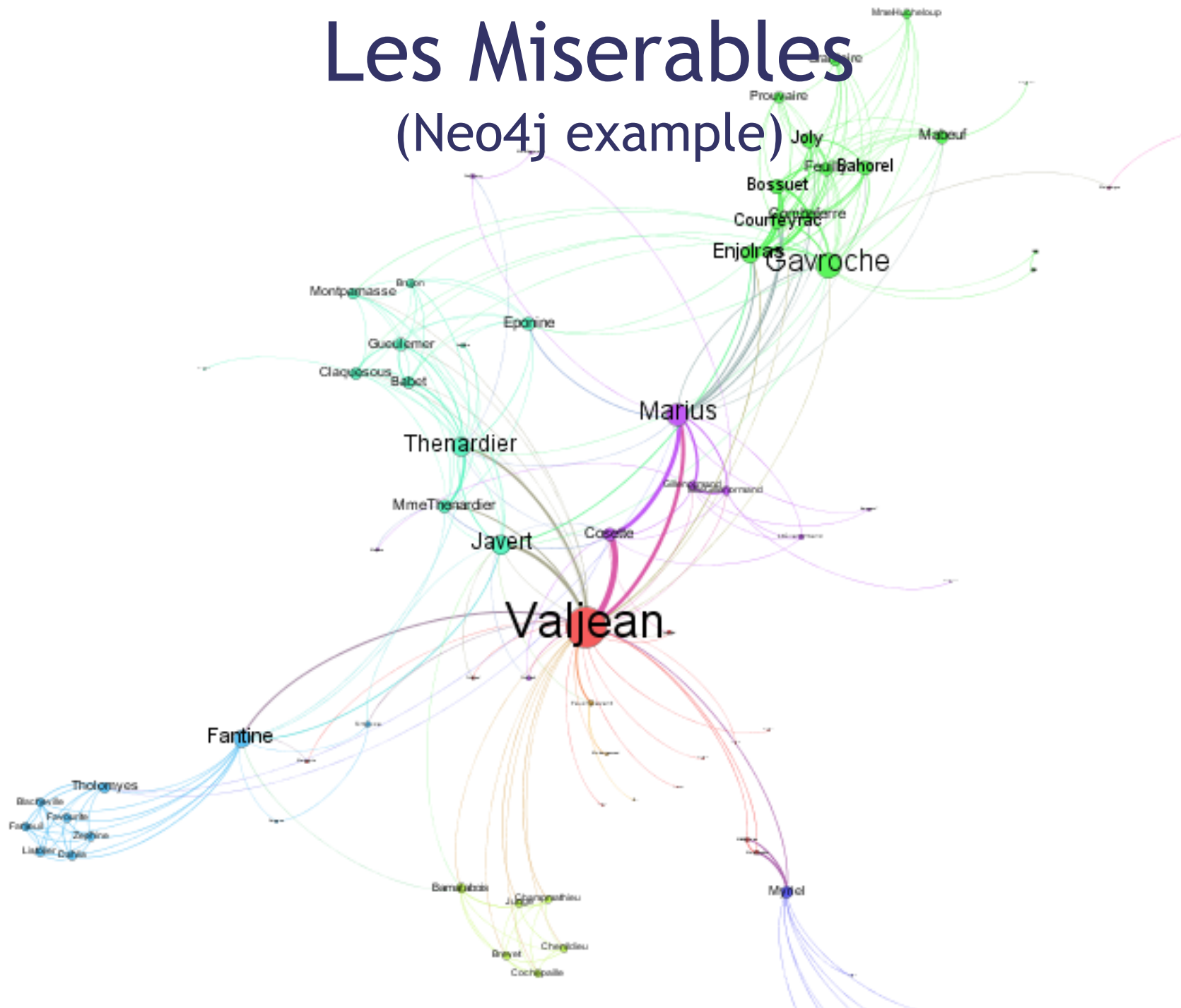
- Physical relationships
 - neurons - nodes; connections - edges
- Force relationships
 - grains - nodes; force vectors - edges
- Social relationships
- Conceptual relationships

Complex Networks

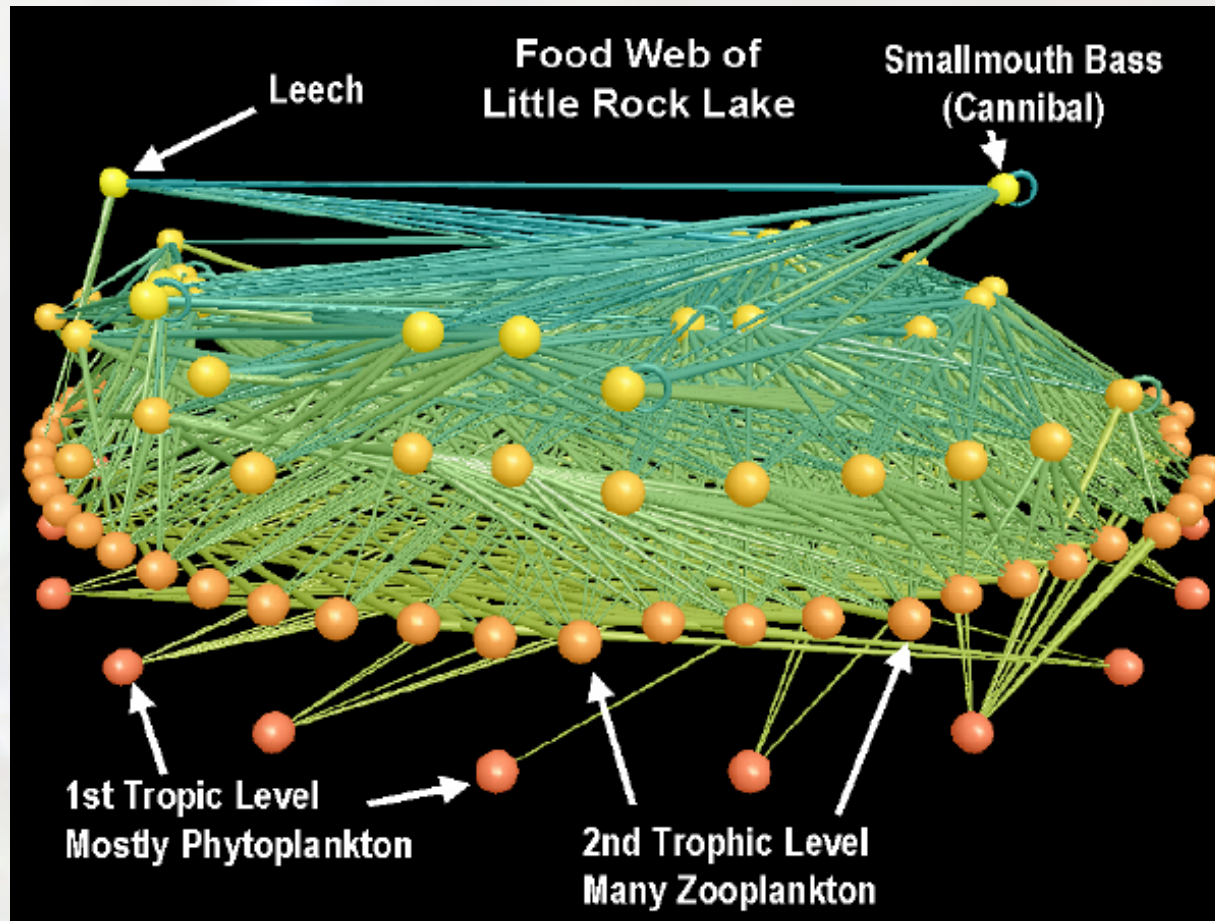
Examples

<http://www-personal.umich.edu/~mejn/networks/>

Les Miserables (Neo4j example)

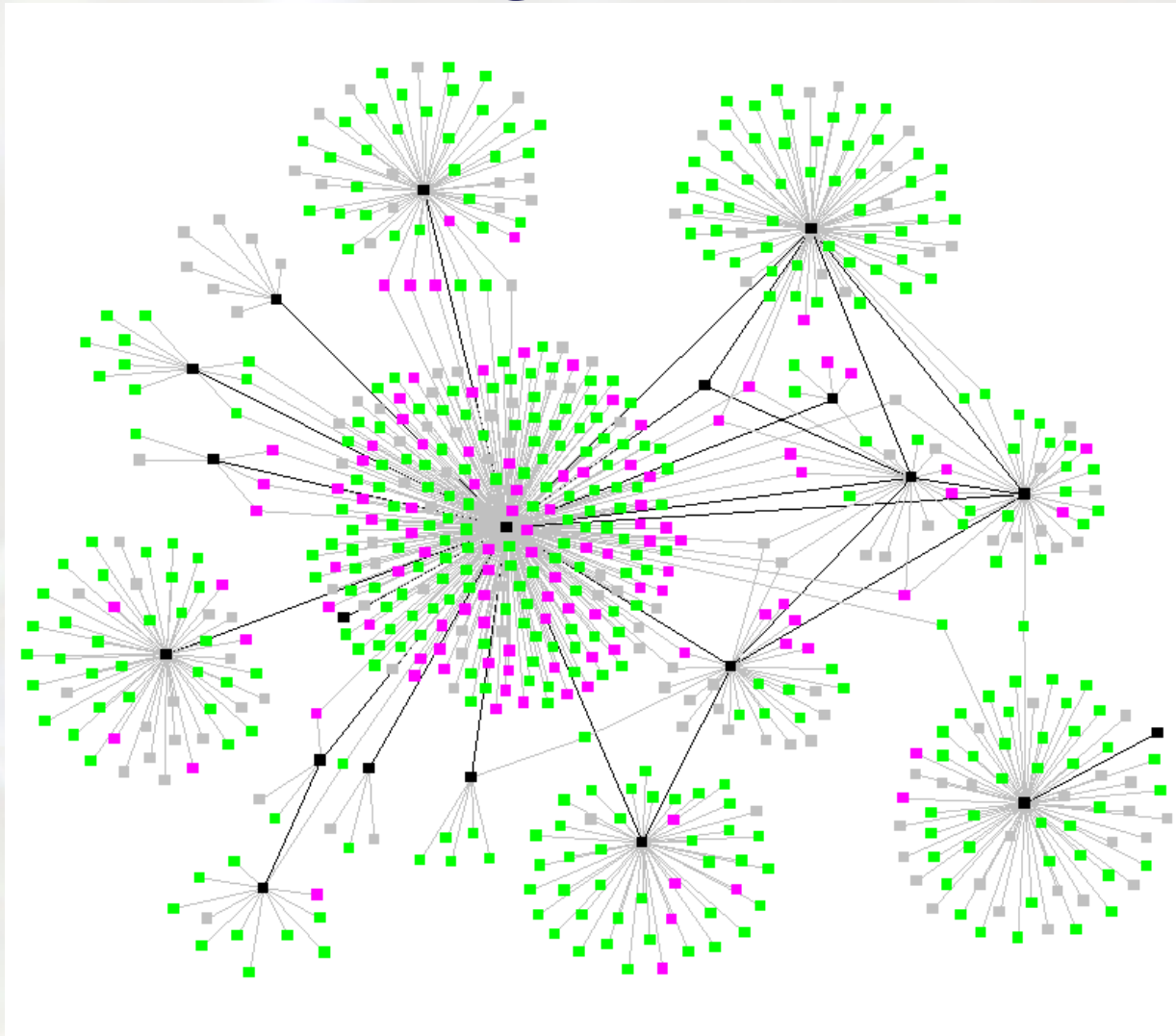


Freshwater food web



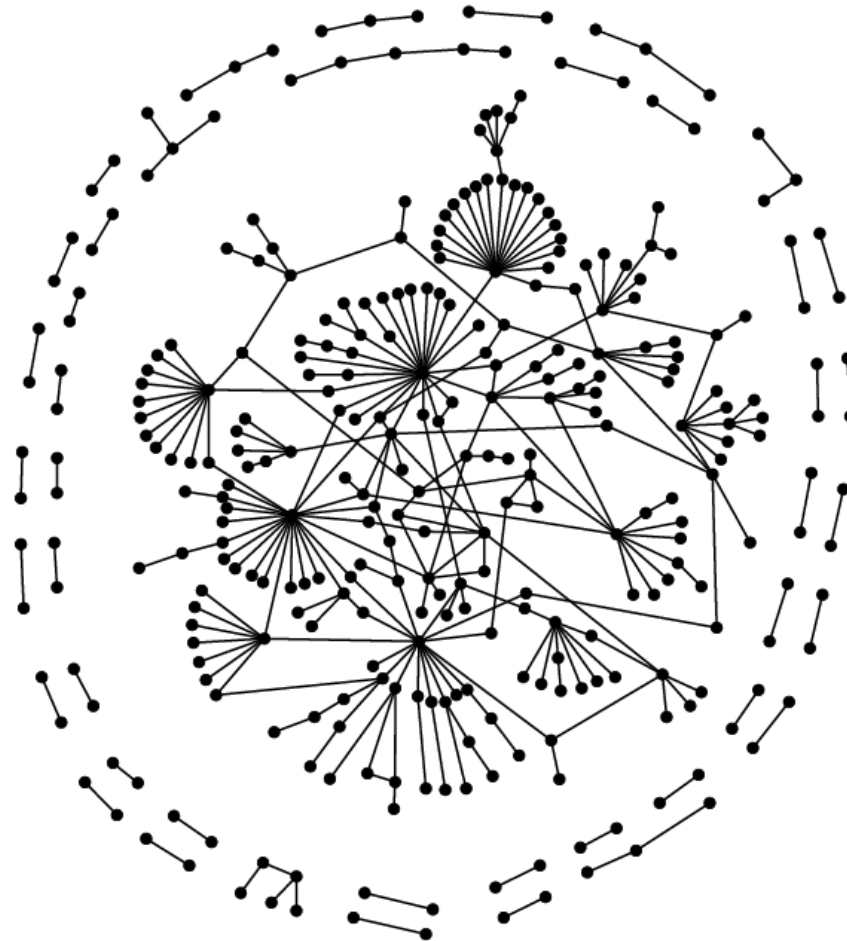
Freshwater food web: [Neo Martinez](#) and Richard Williams.

Contagion of TB



Contagion of TB, books on politics: Valdis Krebs, www.orgnet.com.

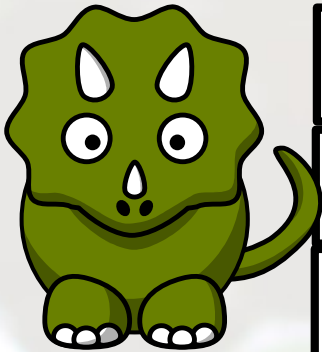
Yeast proteins



Yeast proteins: Sergei Maslov and Kim Sneppen,
Specificity and stability in topology of protein networks,
Science 296, 910-913 (2002).

What can I ask to a Graph?

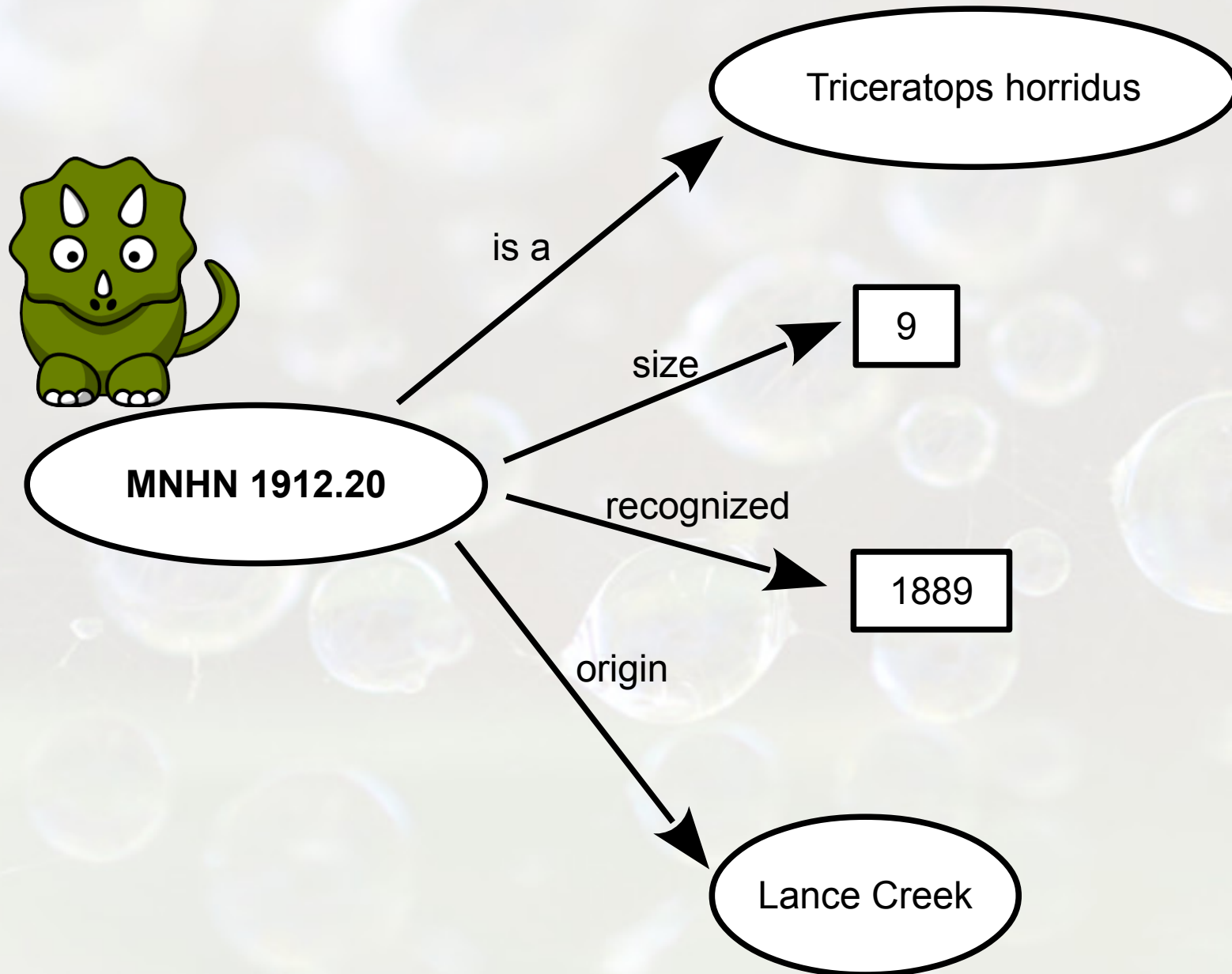
Triceratops in a Graph



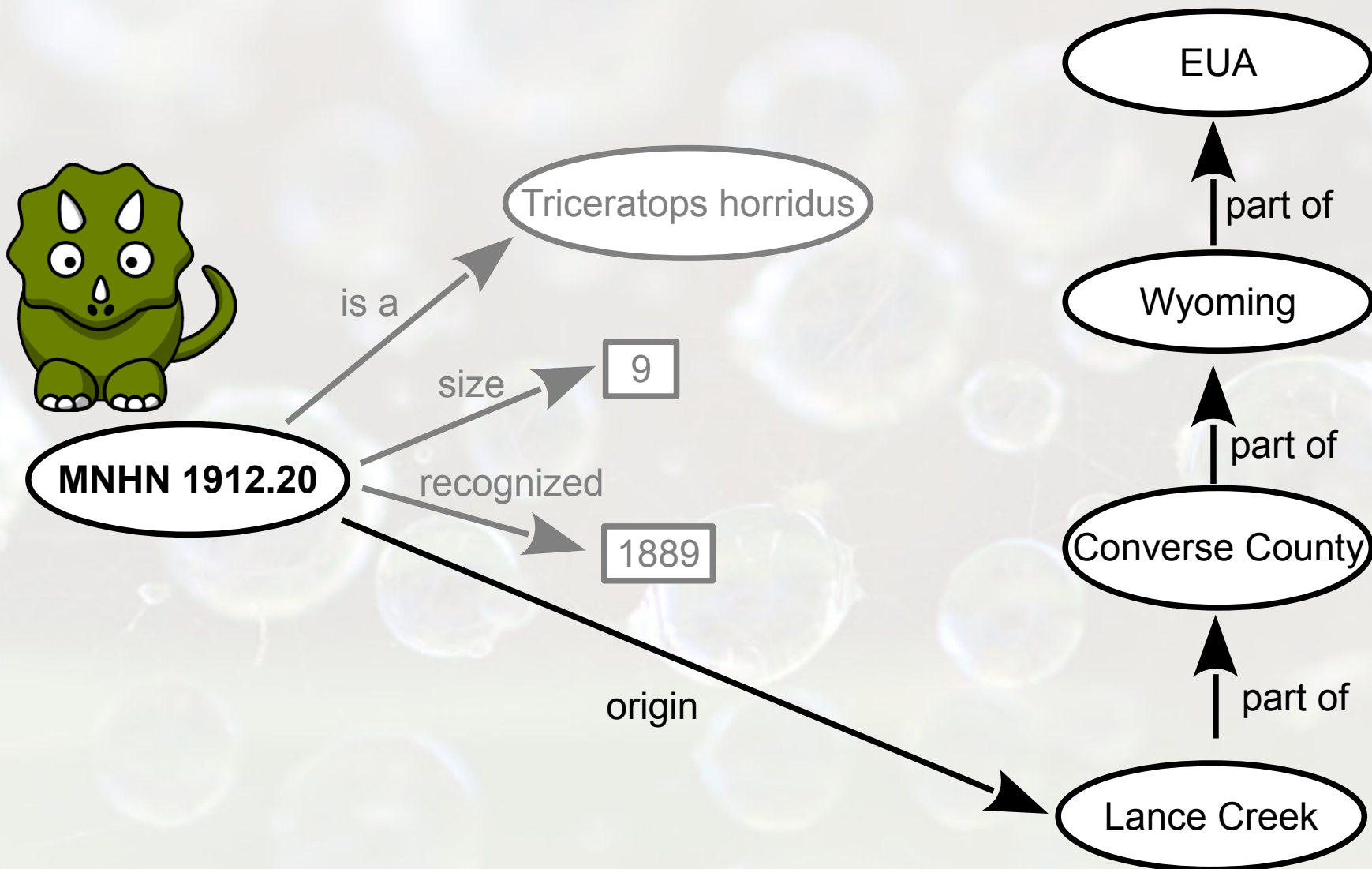
MNHN 1912.20

Is a	Triceratops horridus	
Origin	Lance Creek	EUA
Recognized	1889	
Size	9	

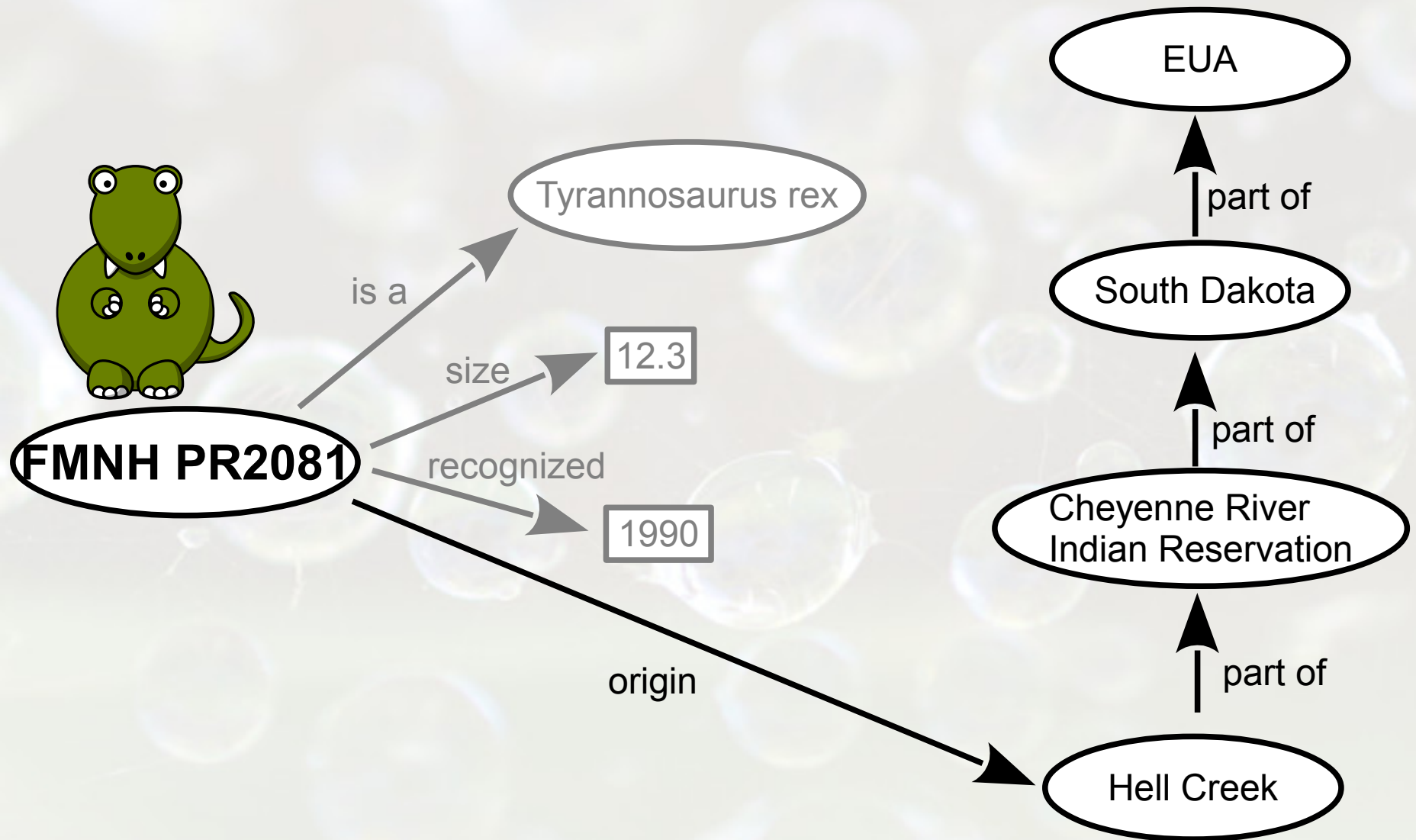
Triceratops in a Graph



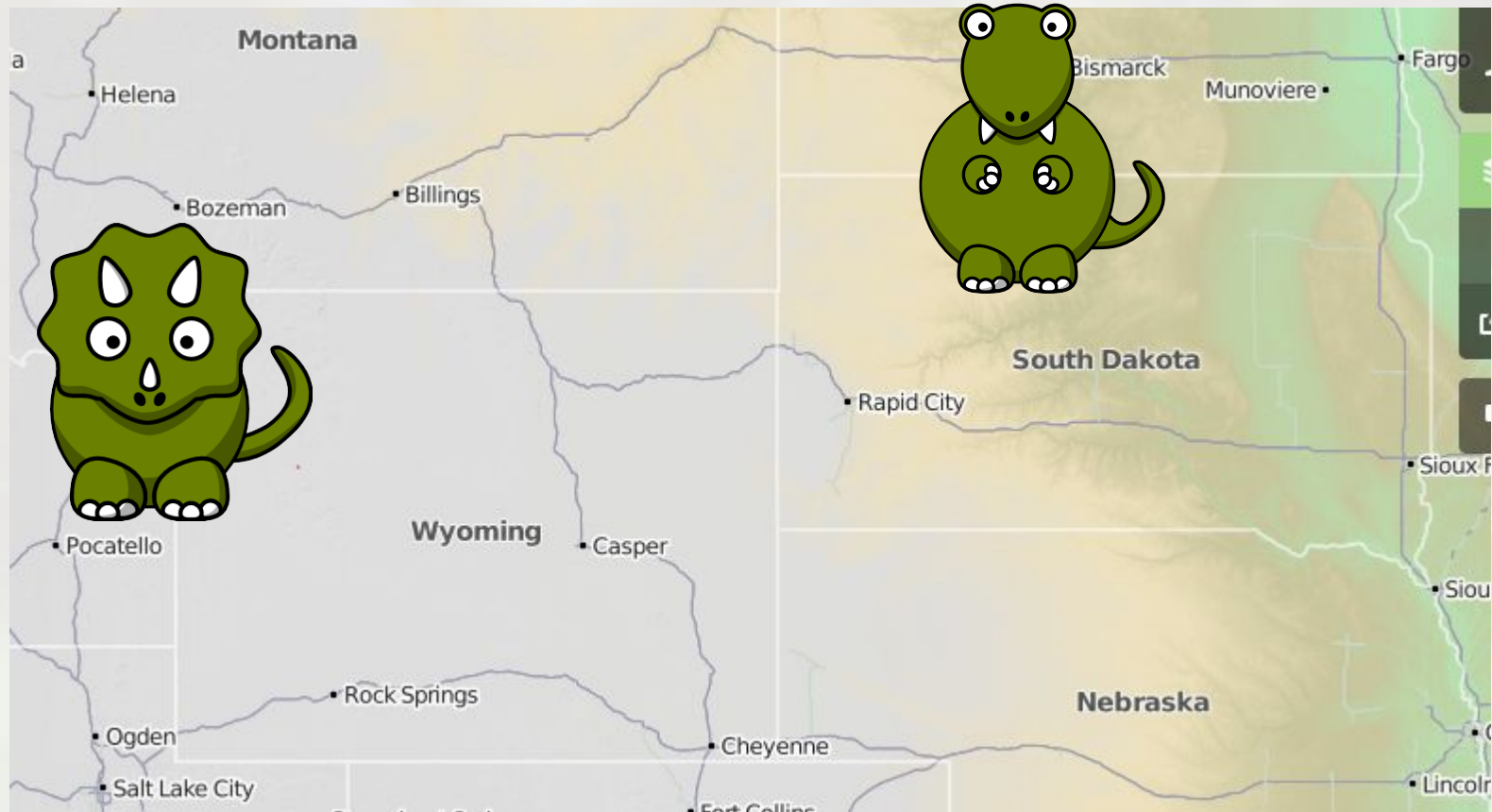
Triceratops in a Graph



Tyrannosaurus in a Graph

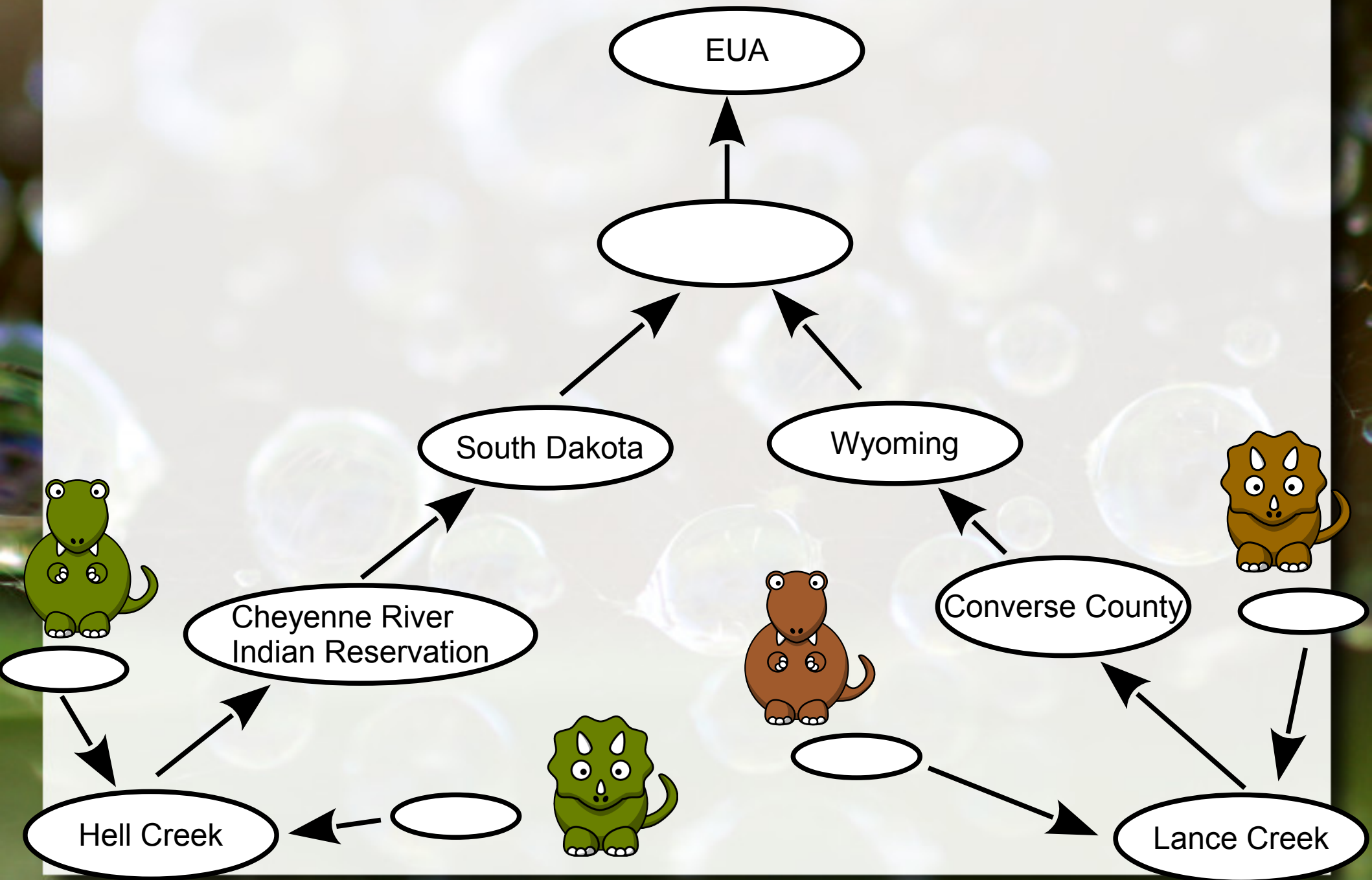


Analyzing in the Space

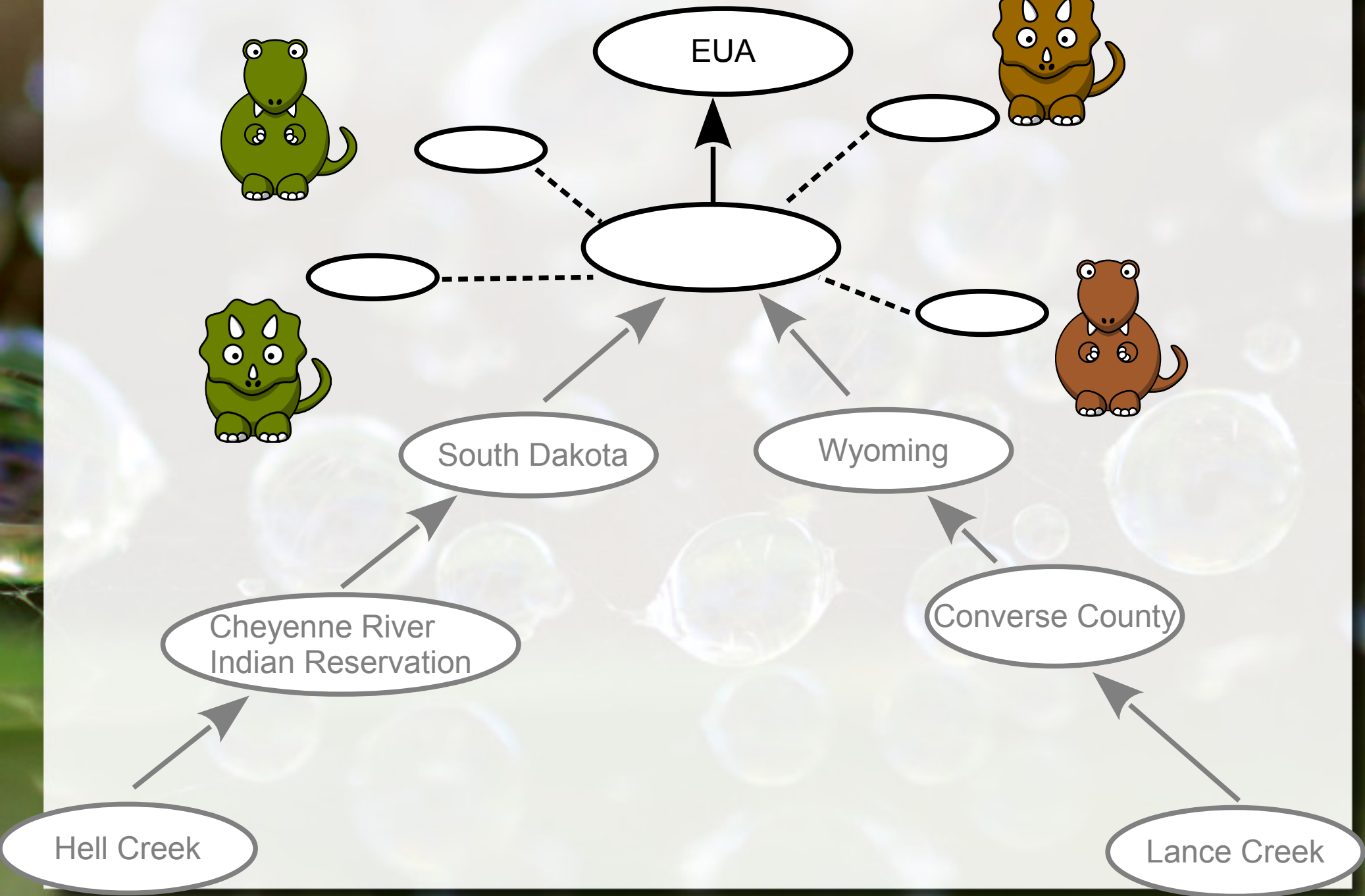


OpenStreetMap

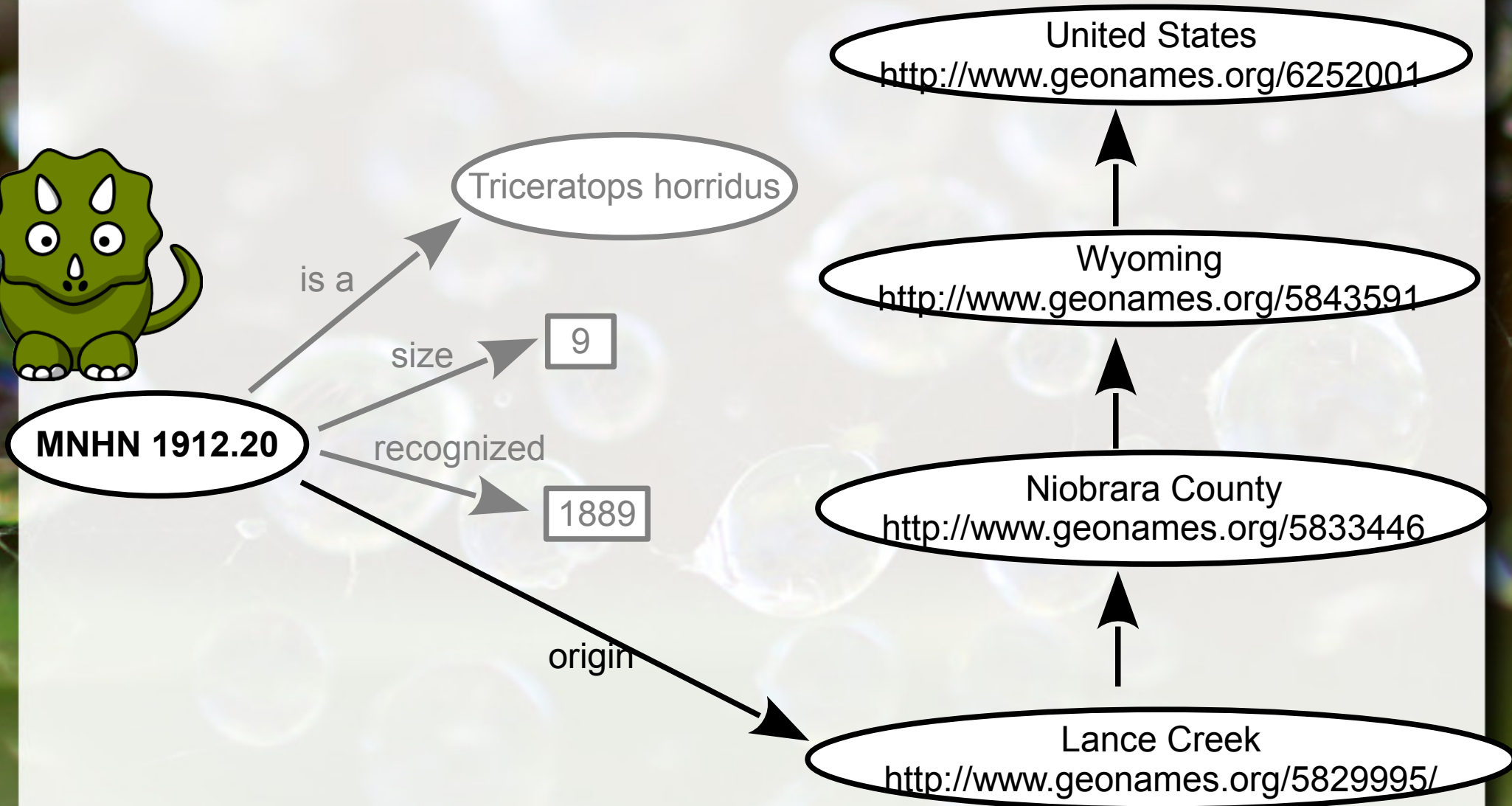
Analyzing in the Space



Analyzing in the Space



GeoNames



Processing the Query

- Process by pattern

- Find [species] whose [origin] → (0..*) part of → EUA

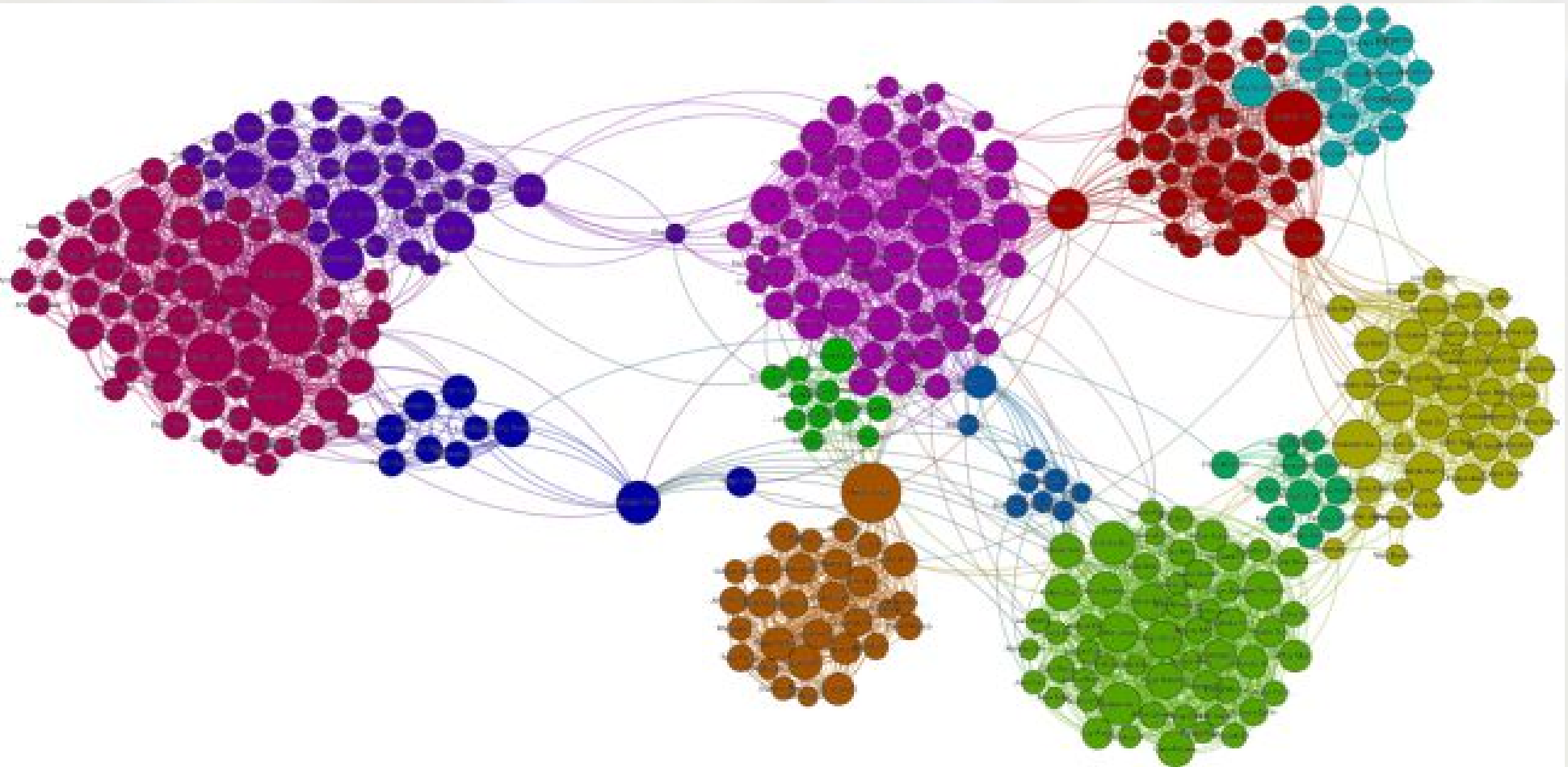
- Process by inference

- Find [species] whose [origin] → EUA

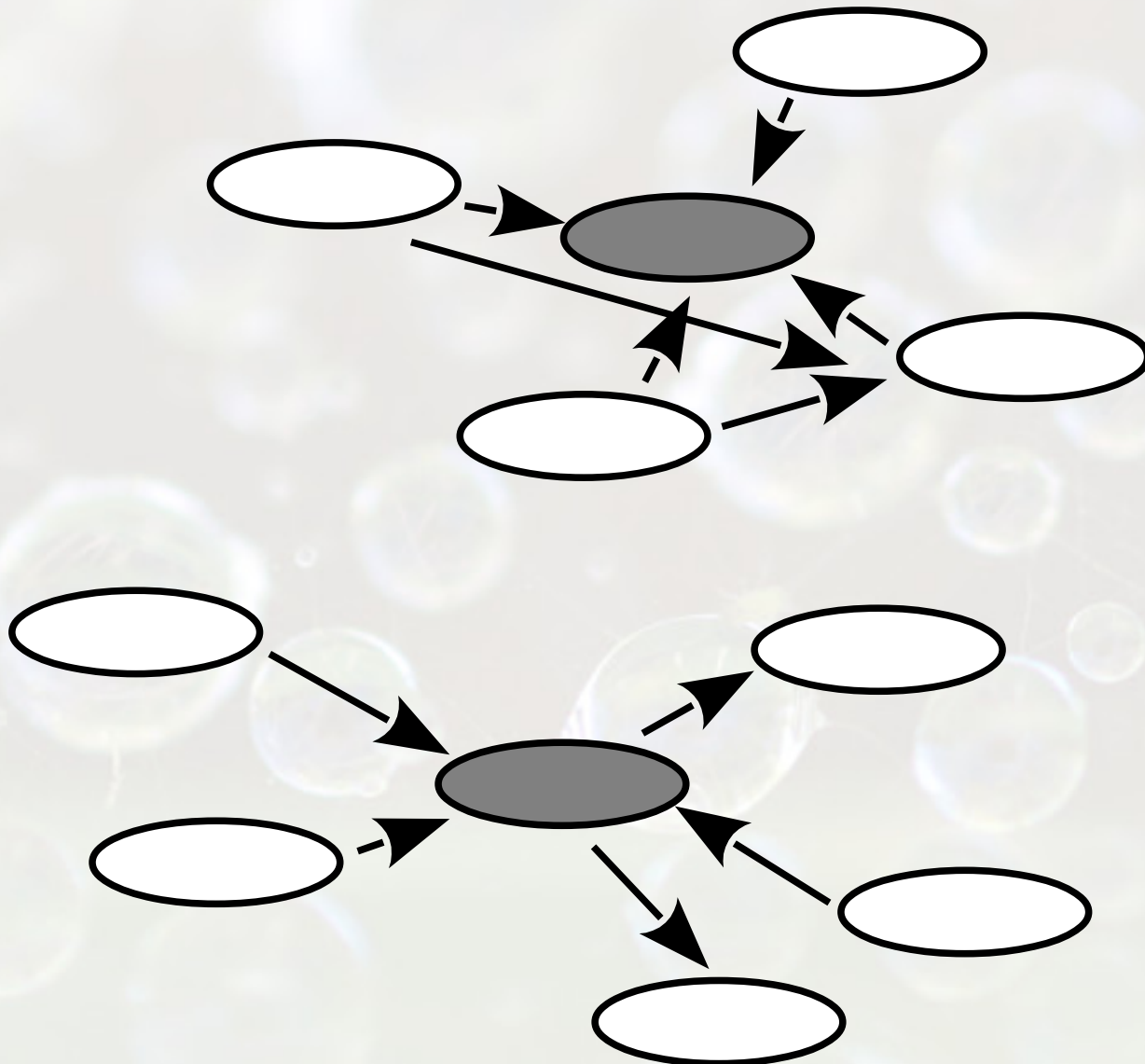
- Rules:

- If (A) origin (B) and (B) part of (C) => (A) origin (C)
- If (A) part of (B) and (B) part of (C) => (A) part of (C)

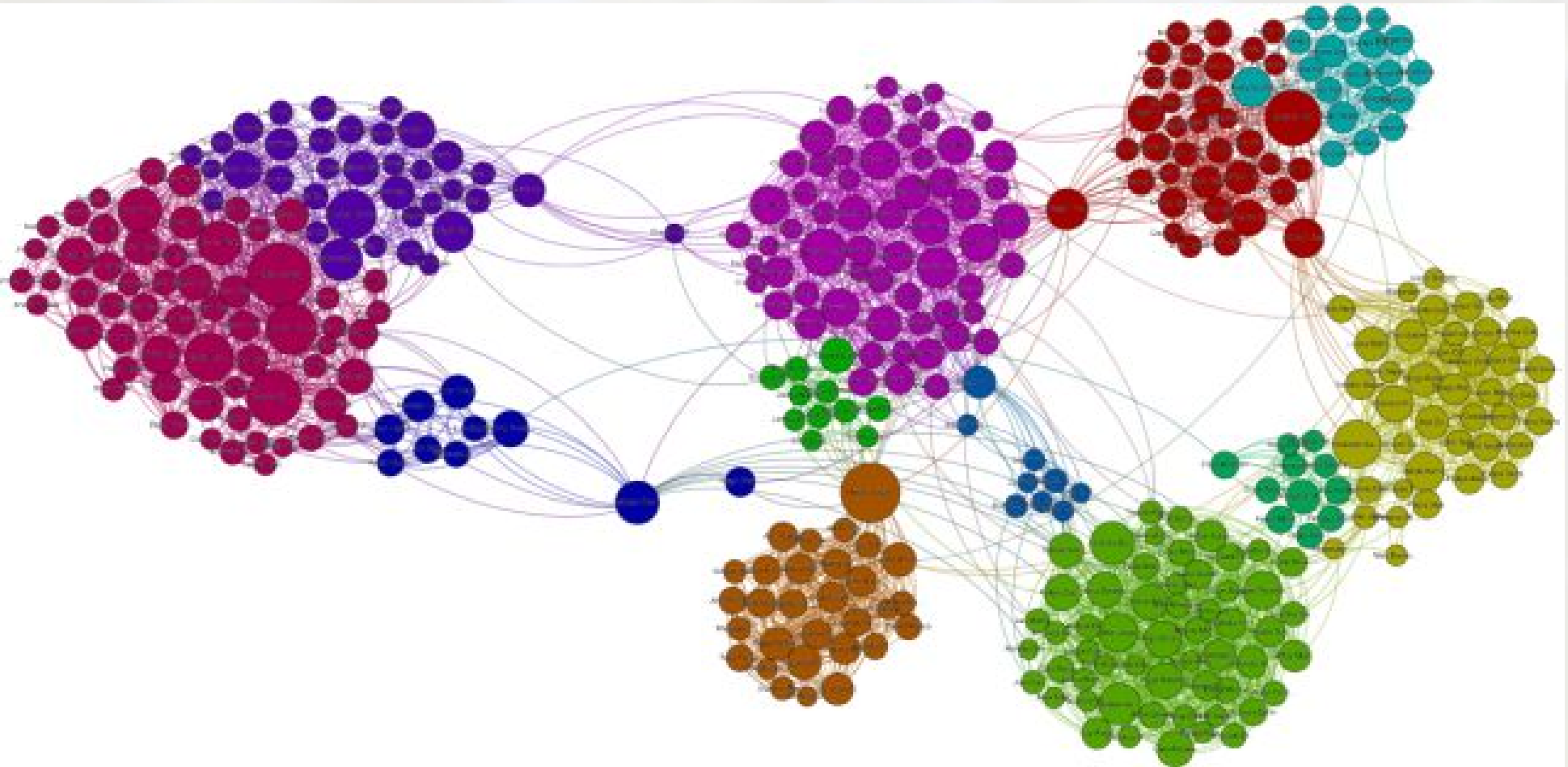
Looking the Topology



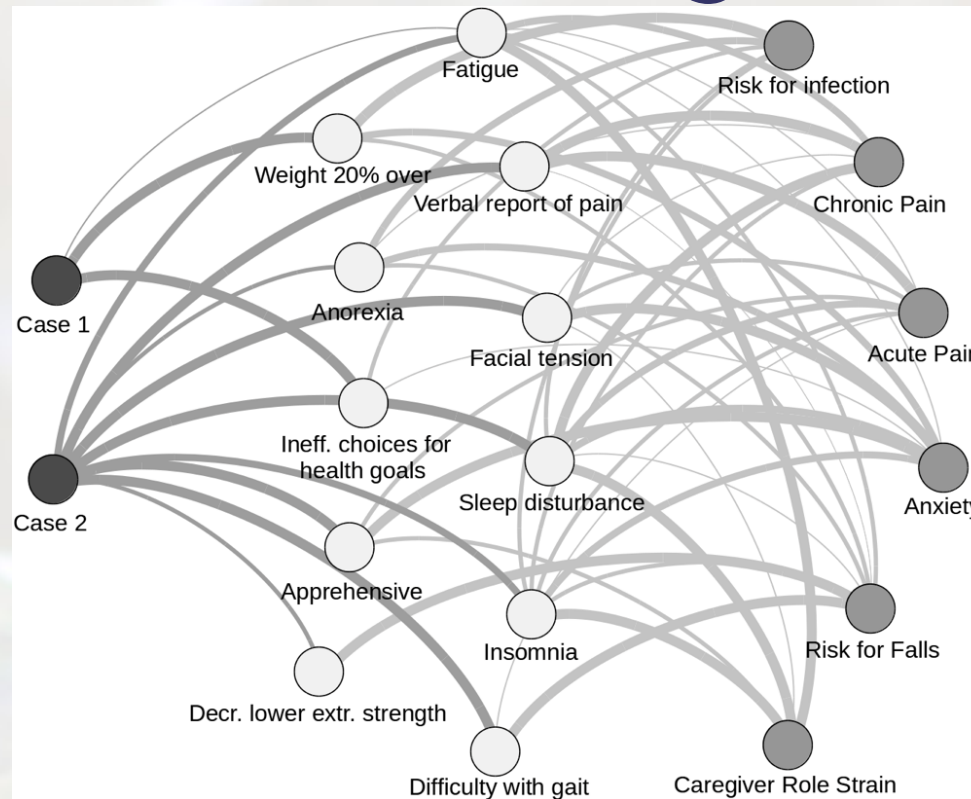
Graph Topology



Graph Topology



Patient x Diagnosis

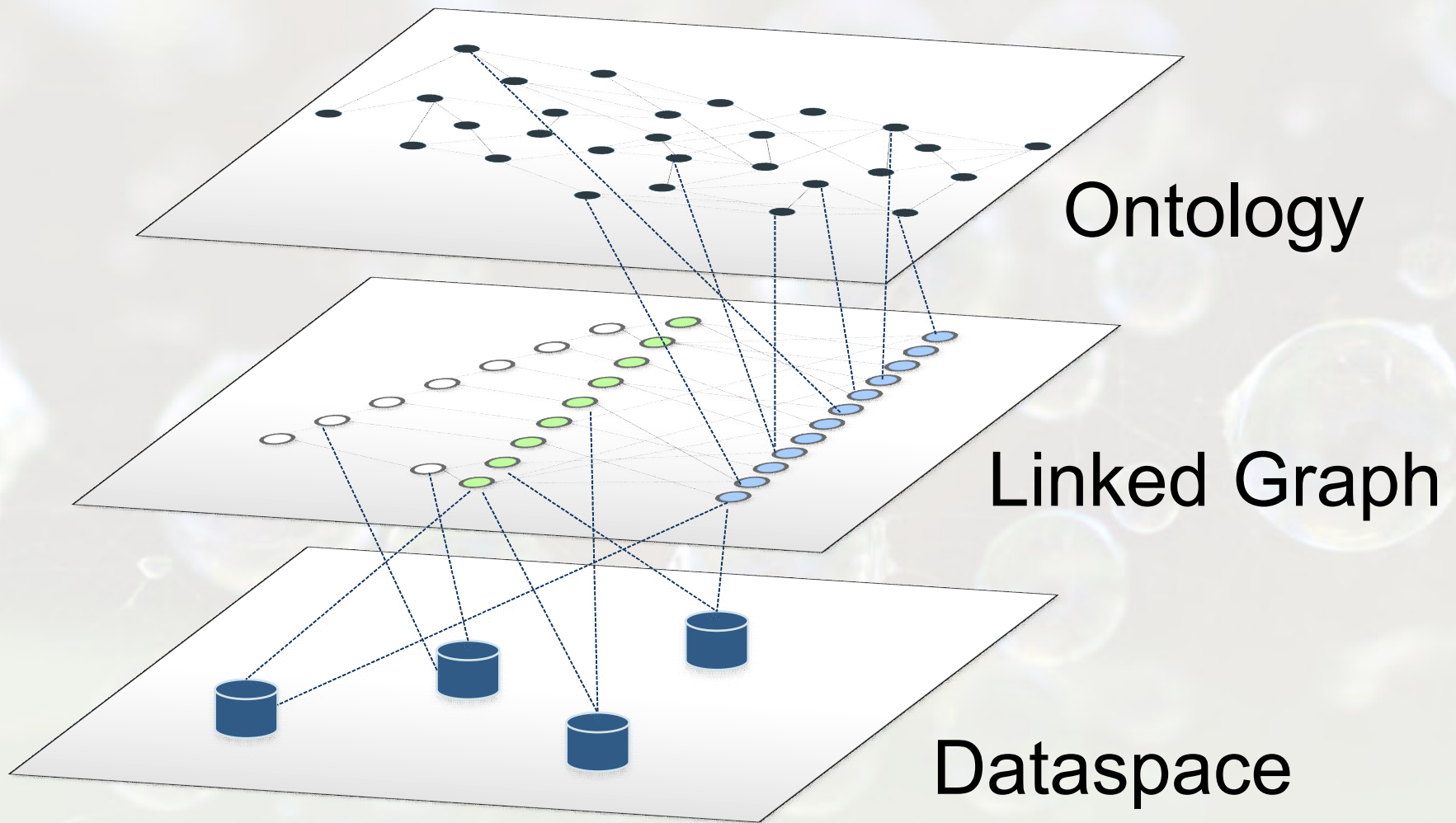


- Retrieve candidate **diagnosis** relevant to a given **patient** based on her **symptoms** and correlations with other similar patients.

(Gomes-Jr, 2013)

From Dataspaces to Ontologies

From Dataspaces to Ontologies



Building Patterns

ID ←

	A	B	C	D	E	F	G
1	Espécie	Área	Data	Hora	Isca	Amostra	coleta
23	<i>Eulaema cingulata</i>	Bom Jardim	8/7/2008	11:01	S.M.	1	2
24	<i>Eulaema meriana</i>	Bom Jardim	8/7/2008	13:01	S.M.	1	2
25	<i>Eulaema meriana</i>	Bom Jardim	8/7/2008	12:55	A.B.	1	2
26	<i>Eulaema meriana</i>	Bom Jardim	8/7/2008	13:11	S.M.	1	2
27	<i>Eulaema meriana</i>	Bom Jardim	8/7/2008	16:10	A.B.	1	2
28	<i>Eulaema meriana</i>	Bom Jardim	8/7/2008	12:09	S.M.	1	2
29	<i>Eulaema meriana</i>	Bom Jardim	8/7/2008	14:44	A.B.	1	2
30	<i>Eulaema meriana</i>	Bom Jardim	8/7/2008	11:51	A.B.	1	2
31	<i>Eulaema meriana</i>	Bom Jardim	8/7/2008	14:44	A.B.	1	2
32	<i>Eulaema meriana</i>	Bom Jardim	8/7/2008	12:04	A.B.	1	2
33	<i>Exaerete frontalis</i>	Bom Jardim	8/7/2008	11:27	S.M.	1	2
34	<i>Eufriesea flaviventris</i>	Bom Jardim	8/7/2008	12:40	S.M.	1	2
35	<i>Eufriesea flaviventris</i>	Bom Jardim	8/7/2008	13:25	S.M.	1	2
36	<i>Euglossa cognata</i>	Bom Jardim	8/7/2008	9:20	S.M.	1	2
37	<i>Euglossa ignita</i>	Bom Jardim	8/7/2008	11:08	A.B.	1	2
38	<i>Euglossa ignita</i>	Bom Jardim	8/7/2008	10:42	A.B.	1	2
39	<i>Euglossa imperialis</i>	Bom Jardim	8/7/2008	14:32	S.M.	1	2

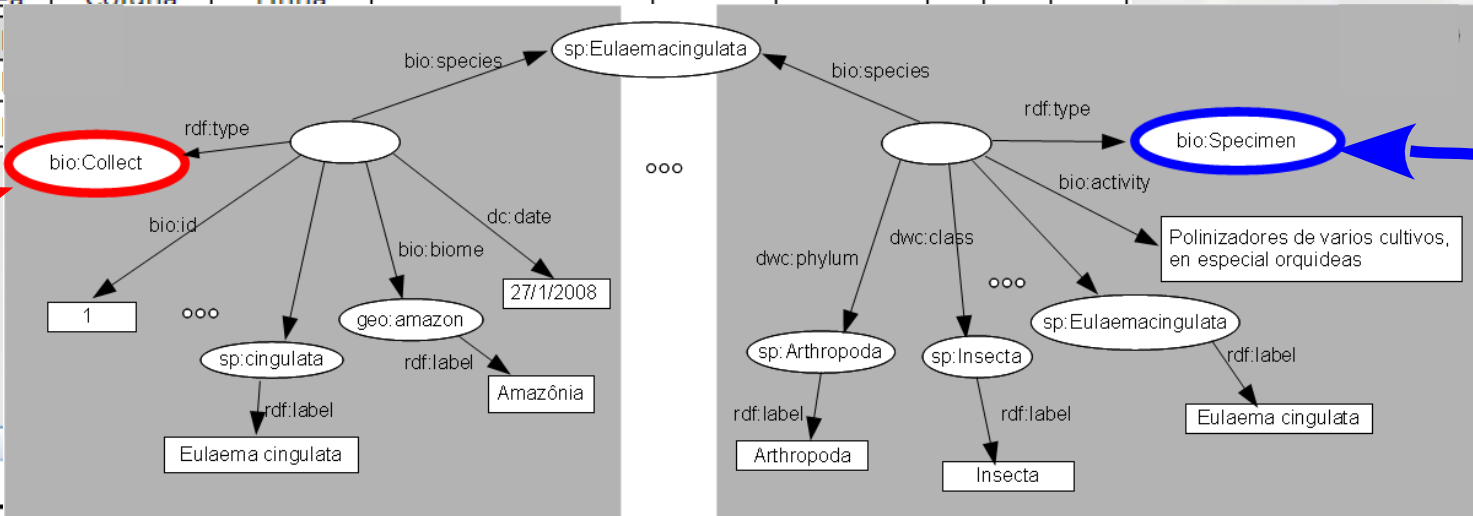
Schema →

Instances →

when what where what

	Data	N. da Estação	Latitude	Longitude	Prof.	Classificação
3						
4			Graus Minutos	Graus Minutos	(metros)	Larsonneur et al.(1982)
5	14/12/1997	6644	25 45.80'	45 11.77'	485	Litoclástico
6	14/12/1997	6645	25 44.09'	45 13.93'	256	Litobioclástico
7	14/12/1997	6646	25 43.78'	45 16.06'	198	Biolitoclástico

	A	B	C	D	E	H	I	M	N	O
1	Planilha Geral dos Dados da Biota									
2	Arquivo de origem: Abelhas 2o. Período									
3		Número da unidade de coleta **			Espécie	Grupo	Bioma ***	dia	mês	ano
4	ID *									
5		Área	Coluna	Linha						
6	1	A1M								
7	2	A1M								
8	3	A1M								



	C										
	Filo	Subfilo	Clase	Orden	Superfamilia	Familia	Subfamilia	Género	Especie	Etapa de Vida	Función / Acción
73	Arthropod	Hexapoda	Insecta	Hymenoptera	Apoidea	Apidae	Apinae	<i>Eulaema</i>	<i>Eulaema bombiformis</i>	Adulto	Polinizadores de varios
74	Arthropod	Hexapoda	Insecta	Hymenoptera	Apoidea	Apidae	Apinae	<i>Eulaema</i>	<i>Eulaema cingulata</i>	Adulto	Polinizadores de varios
75	Arthropod	Hexapoda	Insecta	Hymenoptera	Apoidea	Apidae	Apinae	<i>Eulaema</i>	<i>Eulaema meriana</i>	Adulto	Polinizadores de varios

Data Exploration Prototype

BioSpread Project - LIS IC UNICAMP - Mozilla Firefox

localhost/biospreadPortal/v2/?task=pages/bxnavigator

BioSpread Project

Integrating data from spreadsheet abroad

- Home
- About
- Contact
- Examples
- Data
- Playground
- SPARQL Endpoint
- License

A Taxonomy Navigator

A JavaScript navigator for the extracted data. Navigate through the select boxes and retrieve more information about species

animalia | chordata | aves | passeriformes | icteridae | icterus | icterus galbula

Taxonomy Path: animalia> chordata> aves> passeriformes> icteridae> icterus> icterus galbula

About: "*Icterus galbula*"

Associated Events: Records(200)

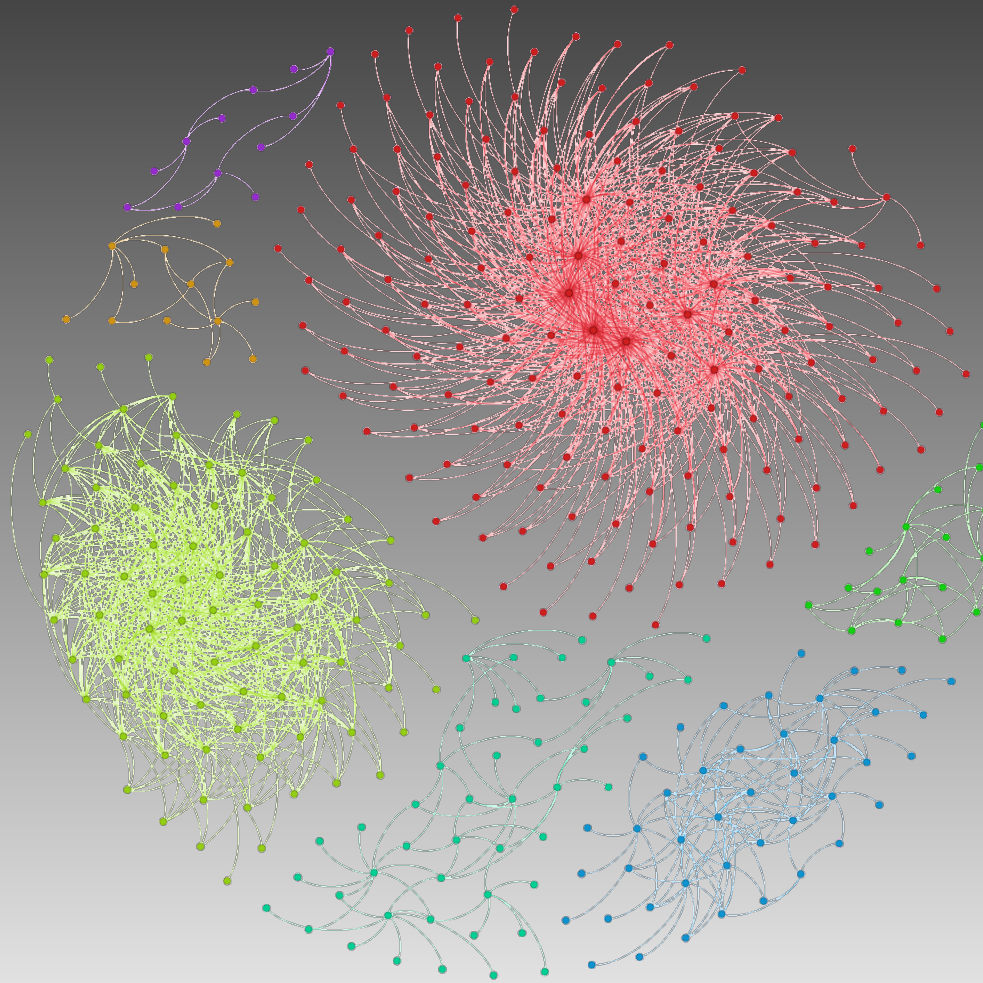
Item uri: <http://purl.org/biospread/resource/collectionitem/30707>
About the collected item | About the species
Location Description: *mato grosso general carneiro*
Country: Brazil

Find: paola | Previous | Next | Highlight all | Match case

Integrating phenotypes

7 distinct morphological descriptions:

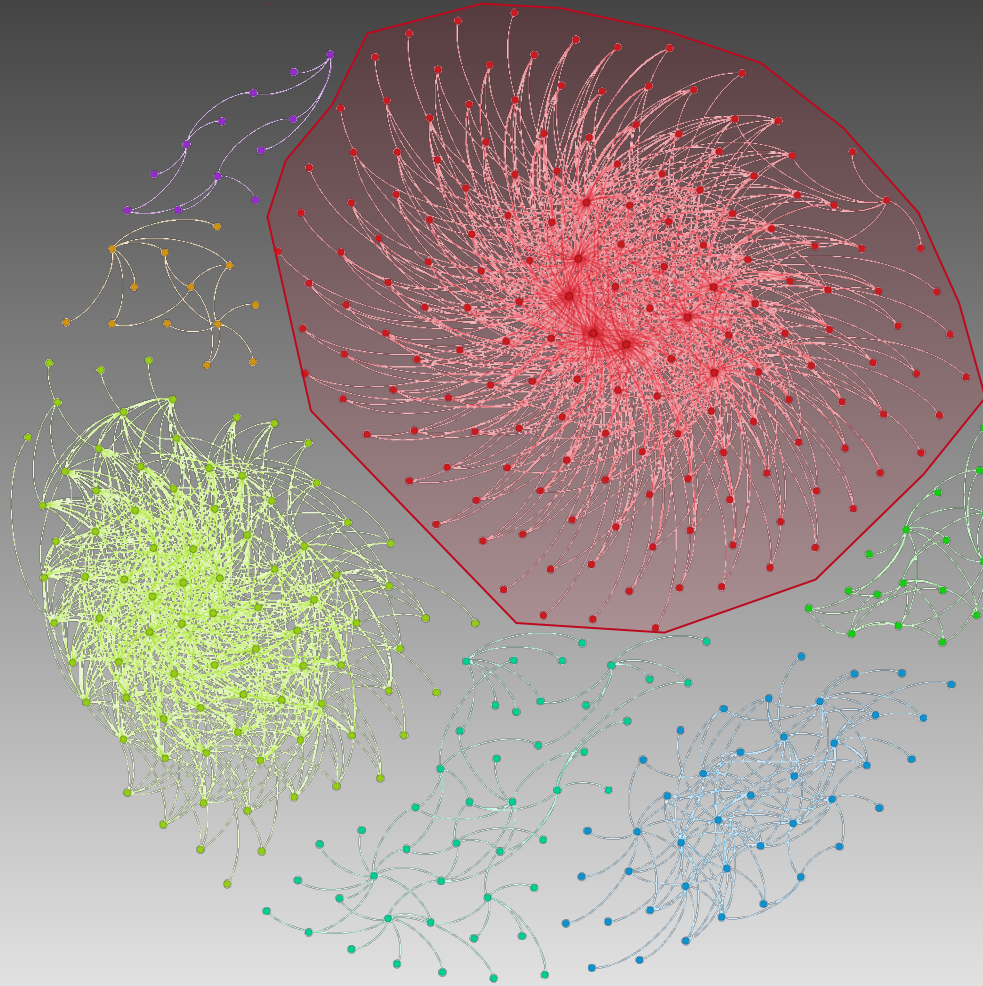
- genus *Varanus*
- *Varanus gouldii*
- *Varanus indicus*
- *Varanus prasinus*
- *Varanus salvator*
- *Varanus spiny*
- *Varanus timorensis*



Our approach

7 distinct morphological descriptions:

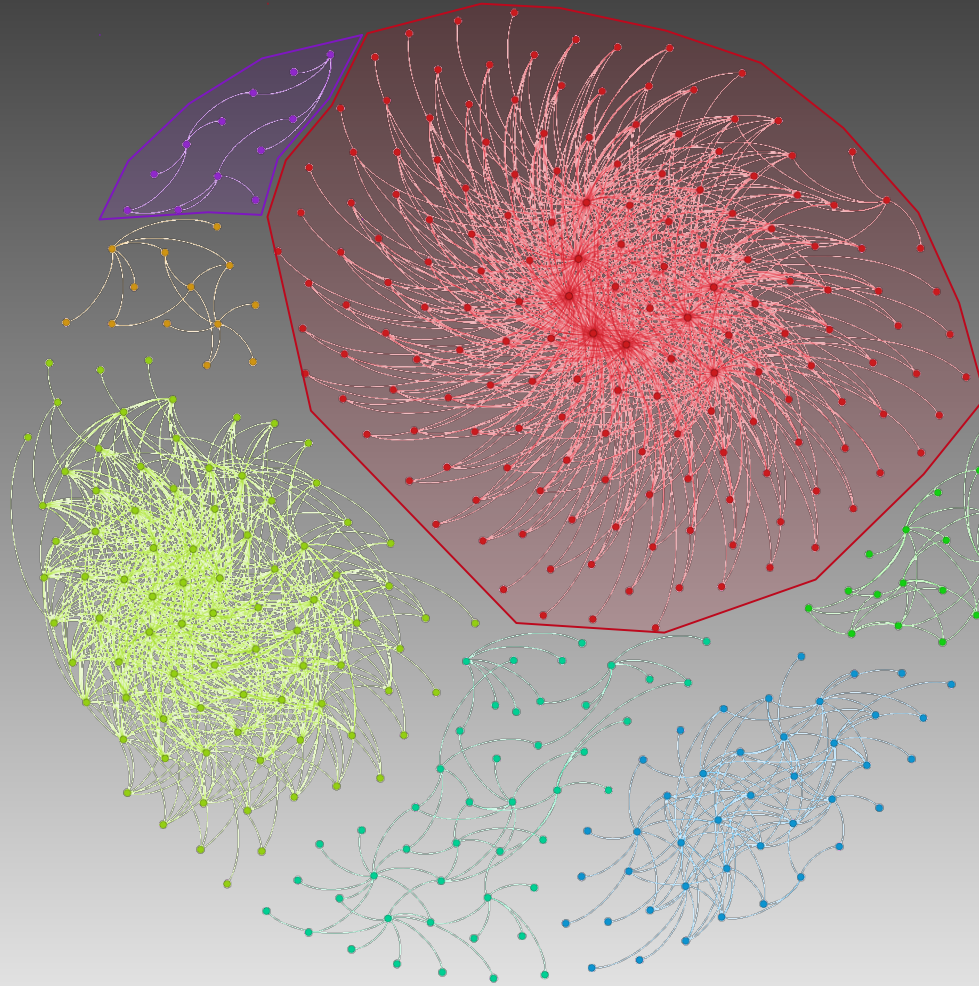
- genus *Varanus*
- *Varanus gouldii*
- *Varanus indicus*
- *Varanus prasinus*
- *Varanus salvator*
- *Varanus spiny*
- *Varanus timorensis*



Our approach

7 distinct morphological descriptions:

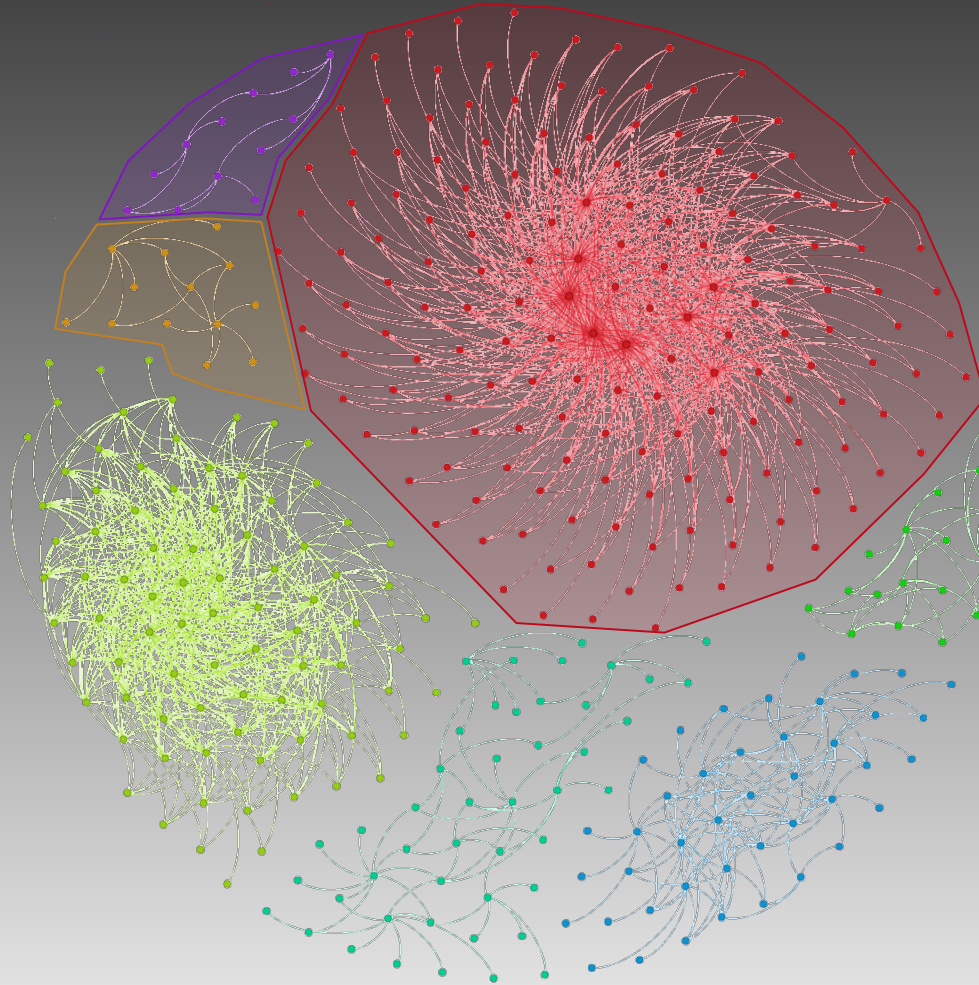
- genus *Varanus*
- *Varanus gouldii*
- *Varanus indicus*
- *Varanus prasinus*
- *Varanus salvator*
- *Varanus spiny*
- *Varanus timorensis*



Our approach

7 distinct morphological descriptions:

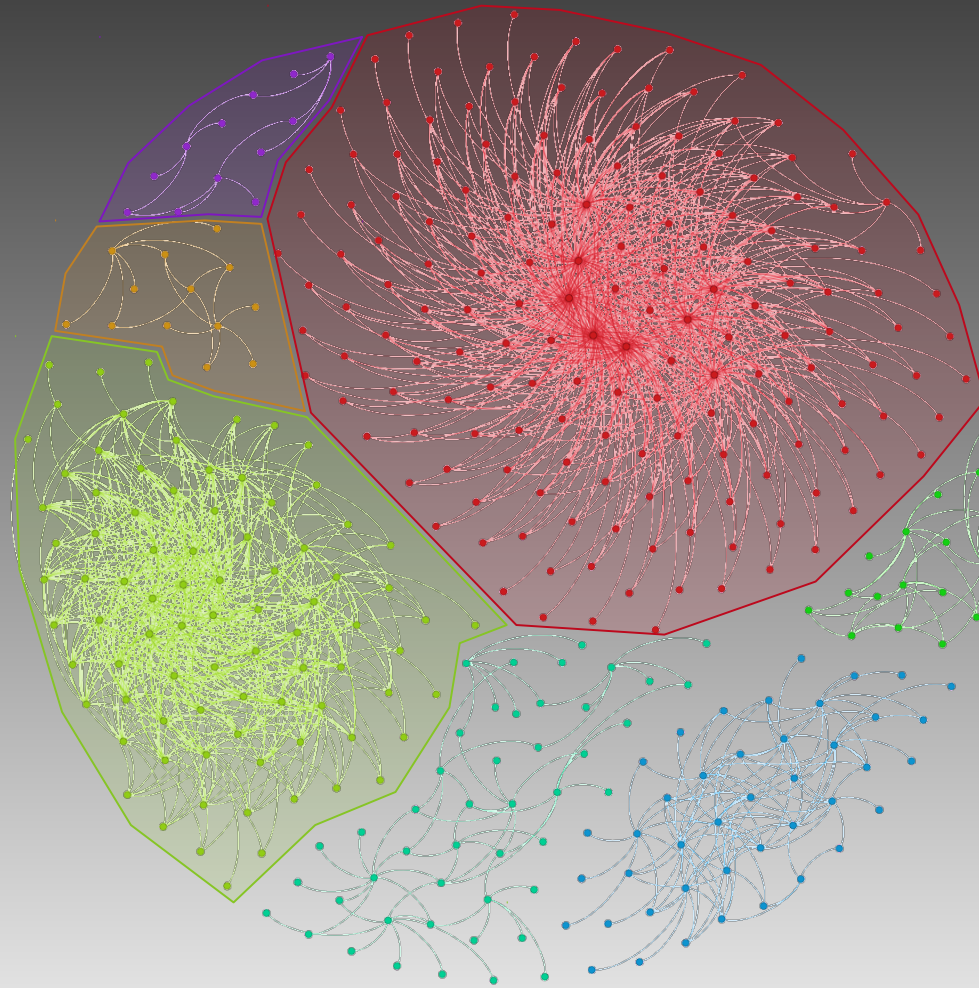
- genus *Varanus*
- *Varanus gouldii*
- *Varanus indicus*
- *Varanus prasinus*
- *Varanus salvator*
- *Varanus spiny*
- *Varanus timorensis*



Our approach

7 distinct morphological descriptions:

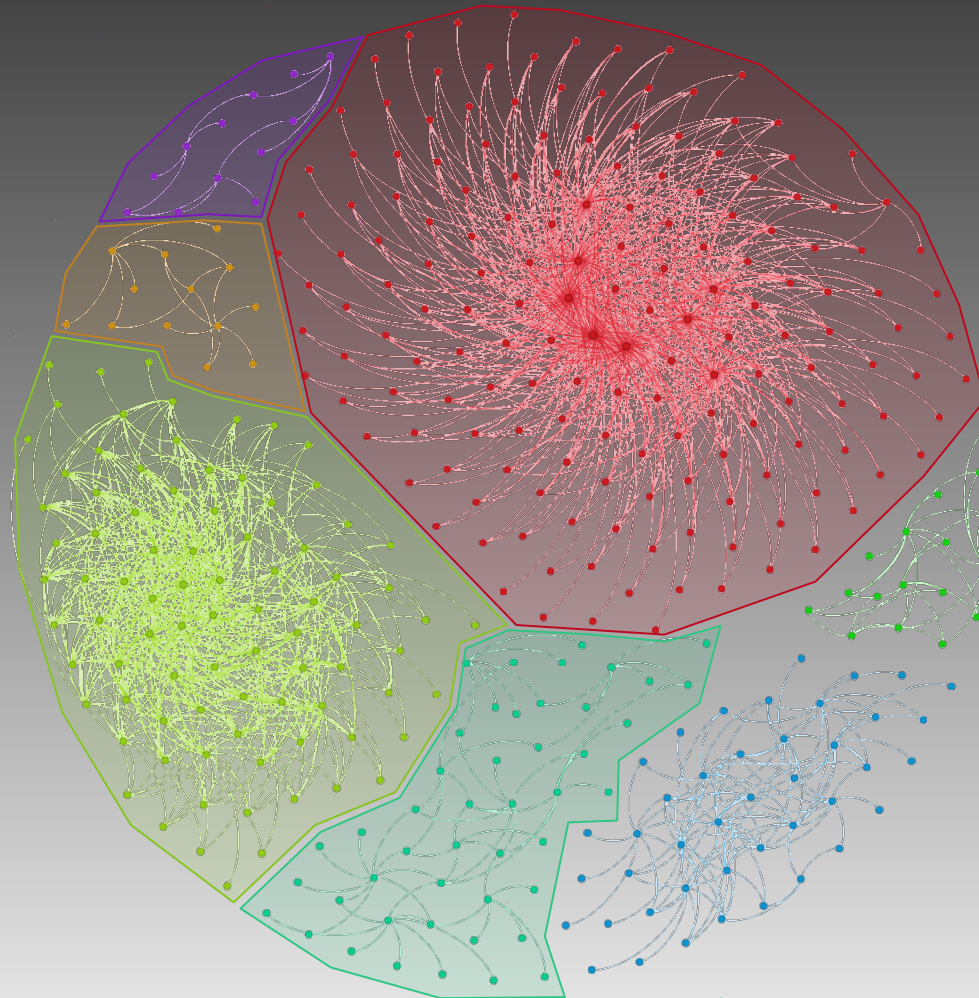
- genus *Varanus*
- *Varanus gouldii*
- *Varanus indicus*
- *Varanus prasinus*
- *Varanus salvator*
- *Varanus spiny*
- *Varanus timorensis*



Our approach

7 distinct morphological descriptions:

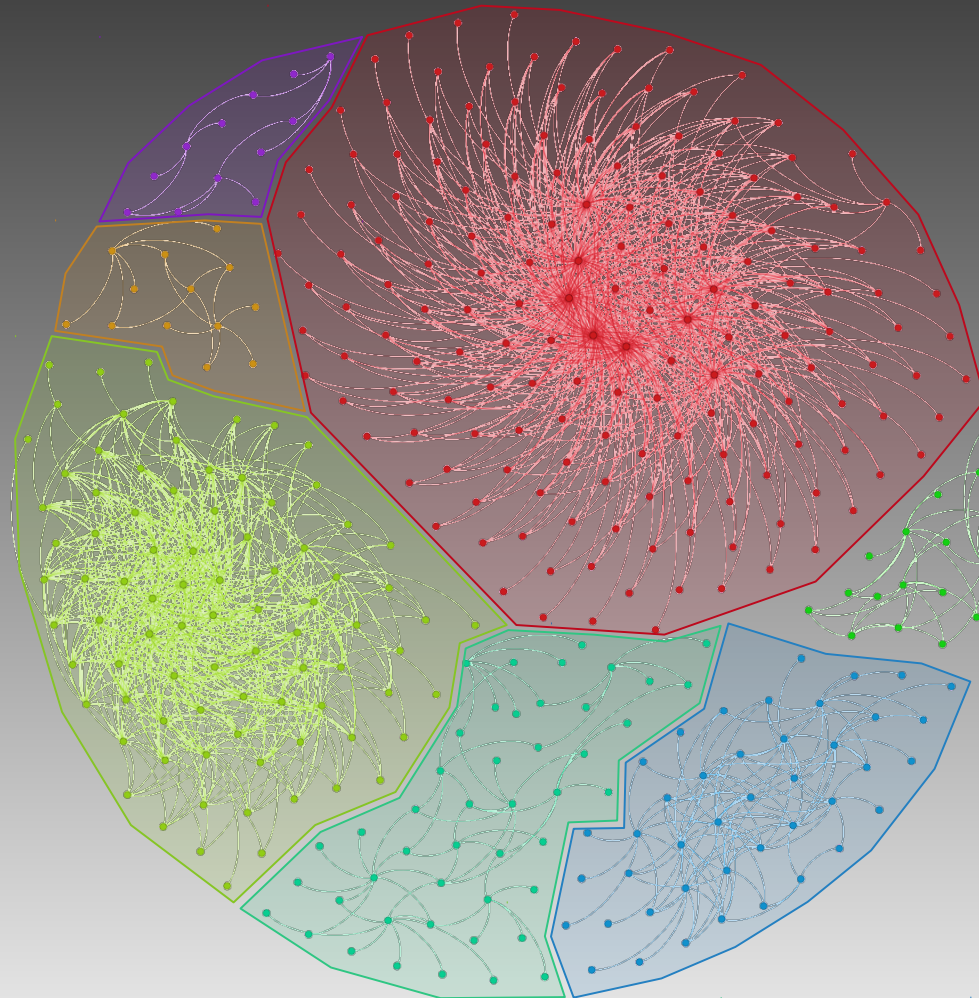
- genus *Varanus*
- *Varanus gouldii*
- *Varanus indicus*
- *Varanus prasinus*
- *Varanus salvator*
- *Varanus spiny*
- *Varanus timorensis*



Our approach

7 distinct morphological descriptions:

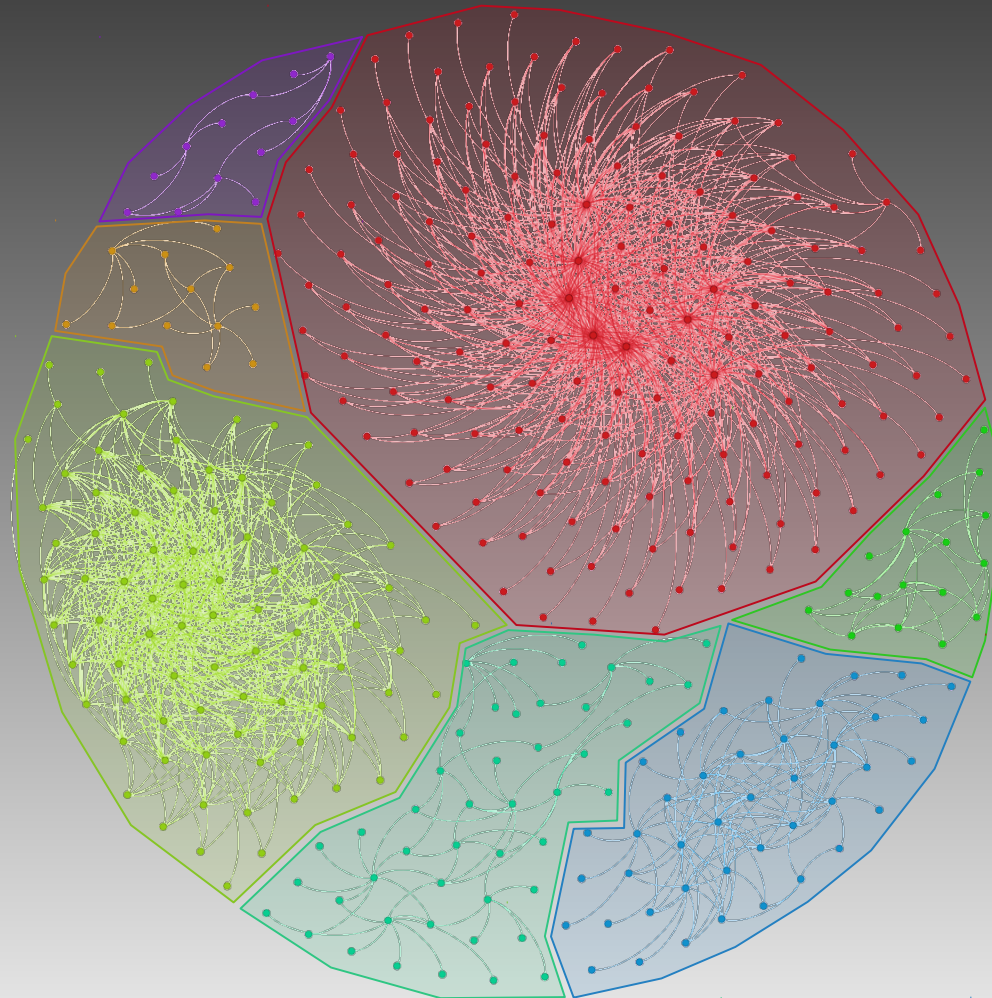
- genus *Varanus*
- *Varanus gouldii*
- *Varanus indicus*
- *Varanus prasinus*
- *Varanus salvator*
- *Varanus spiny*
- *Varanus timorensis*



Our approach

7 distinct morphological descriptions:

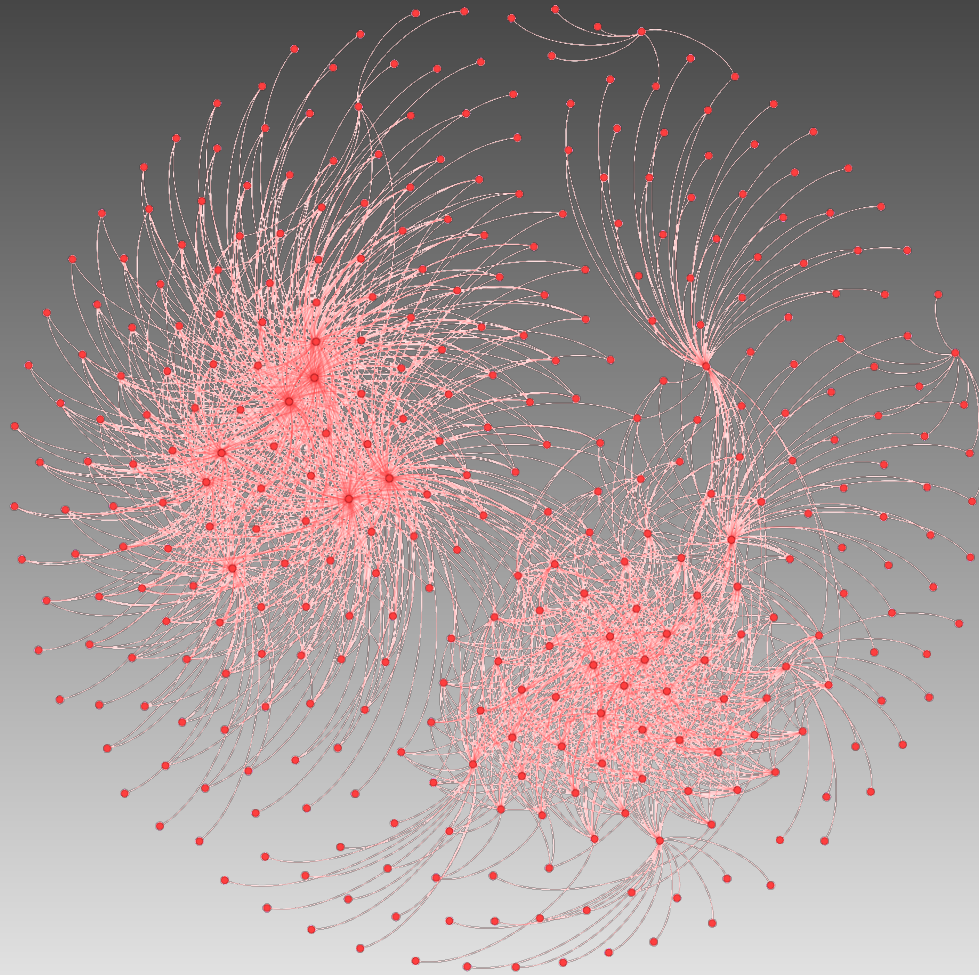
- genus *Varanus*
- *Varanus gouldii*
- *Varanus indicus*
- *Varanus prasinus*
- *Varanus salvator*
- *Varanus spiny*
- *Varanus timorensis*



Our approach

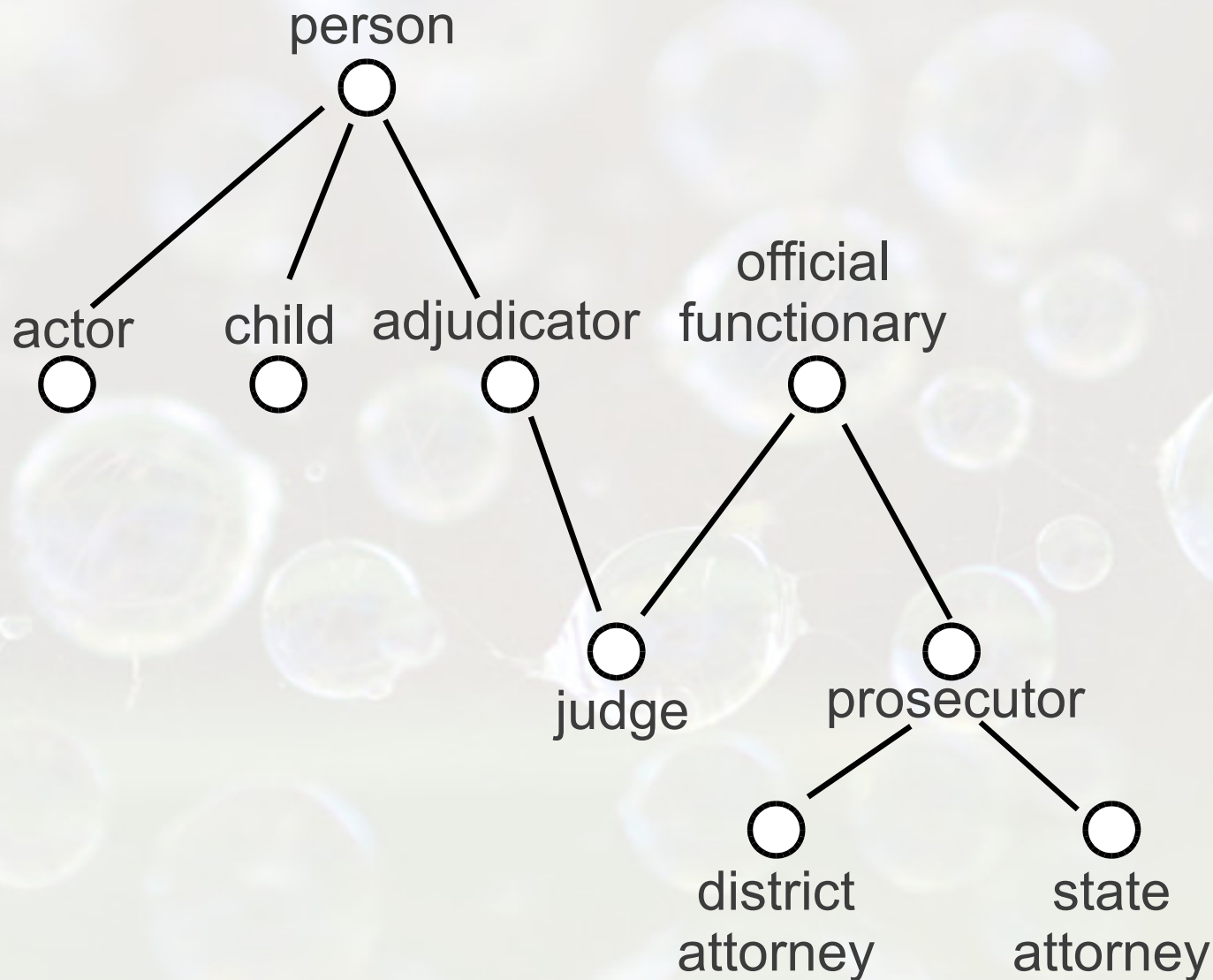
7 distinct morphological descriptions:

- genus *Varanus*
- *Varanus gouldii*
- *Varanus indicus*
- *Varanus prasinus*
- *Varanus salvator*
- *Varanus spiny*
- *Varanus timorensis*

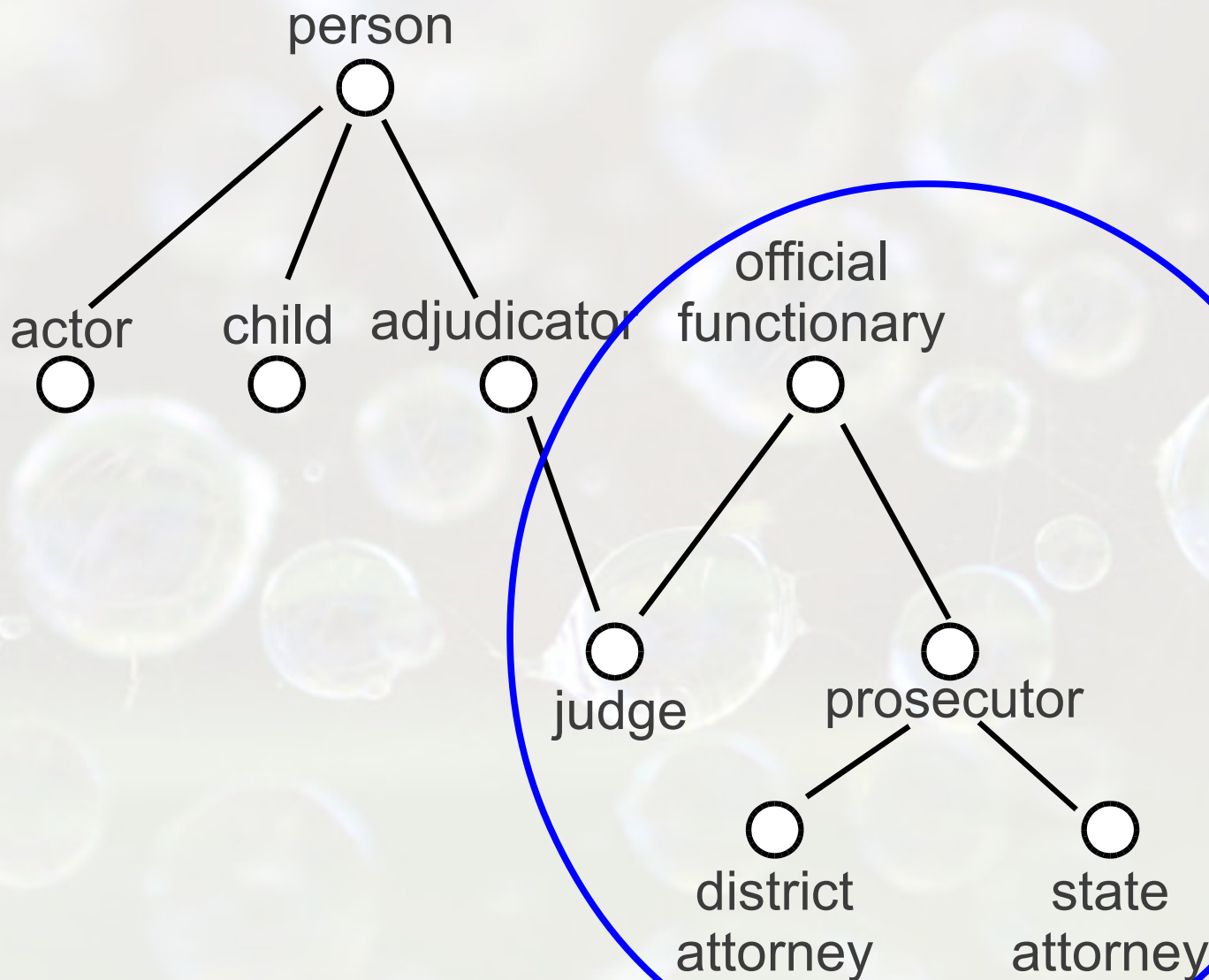


Evaluating Similarity

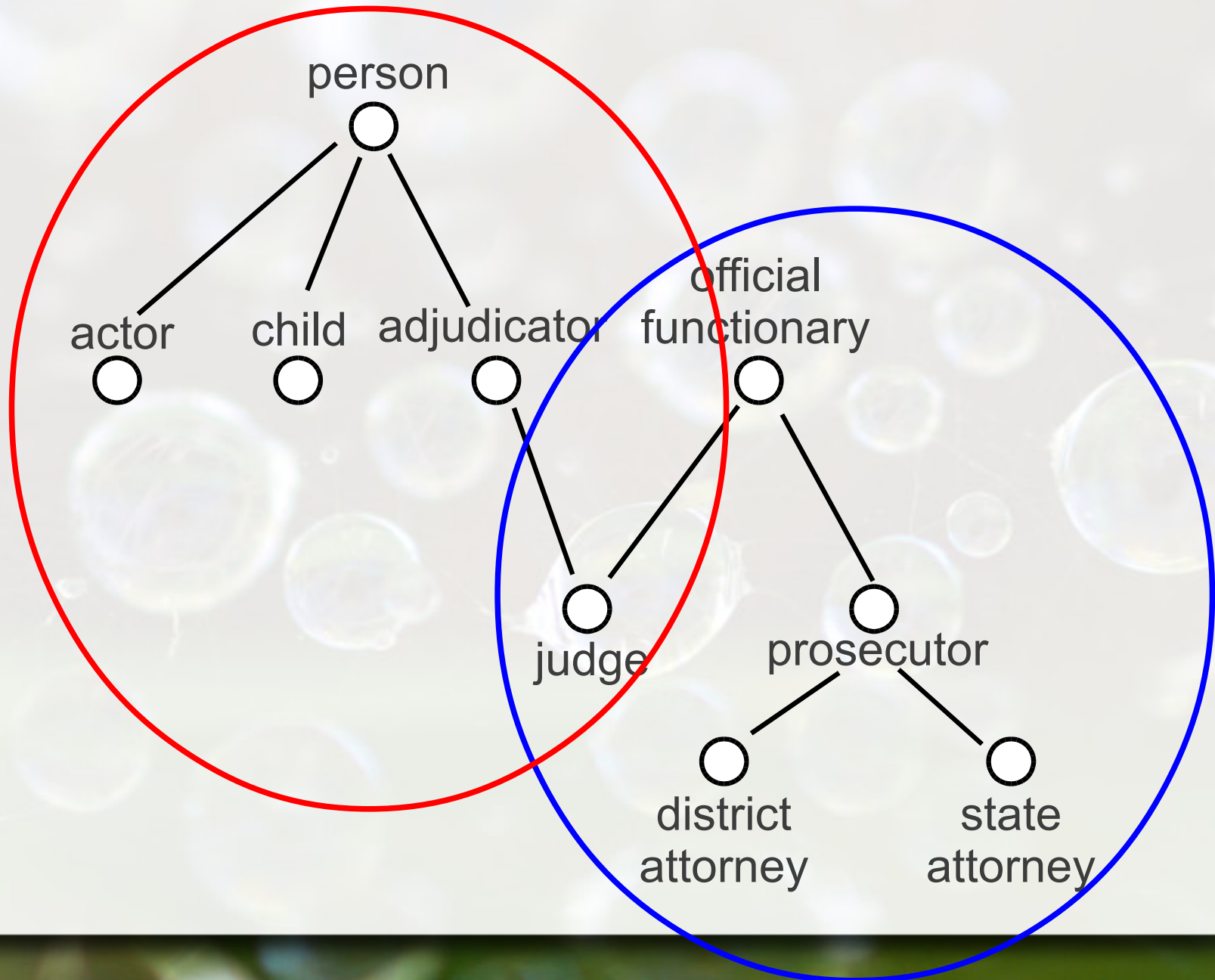
Size of the Shared Universe



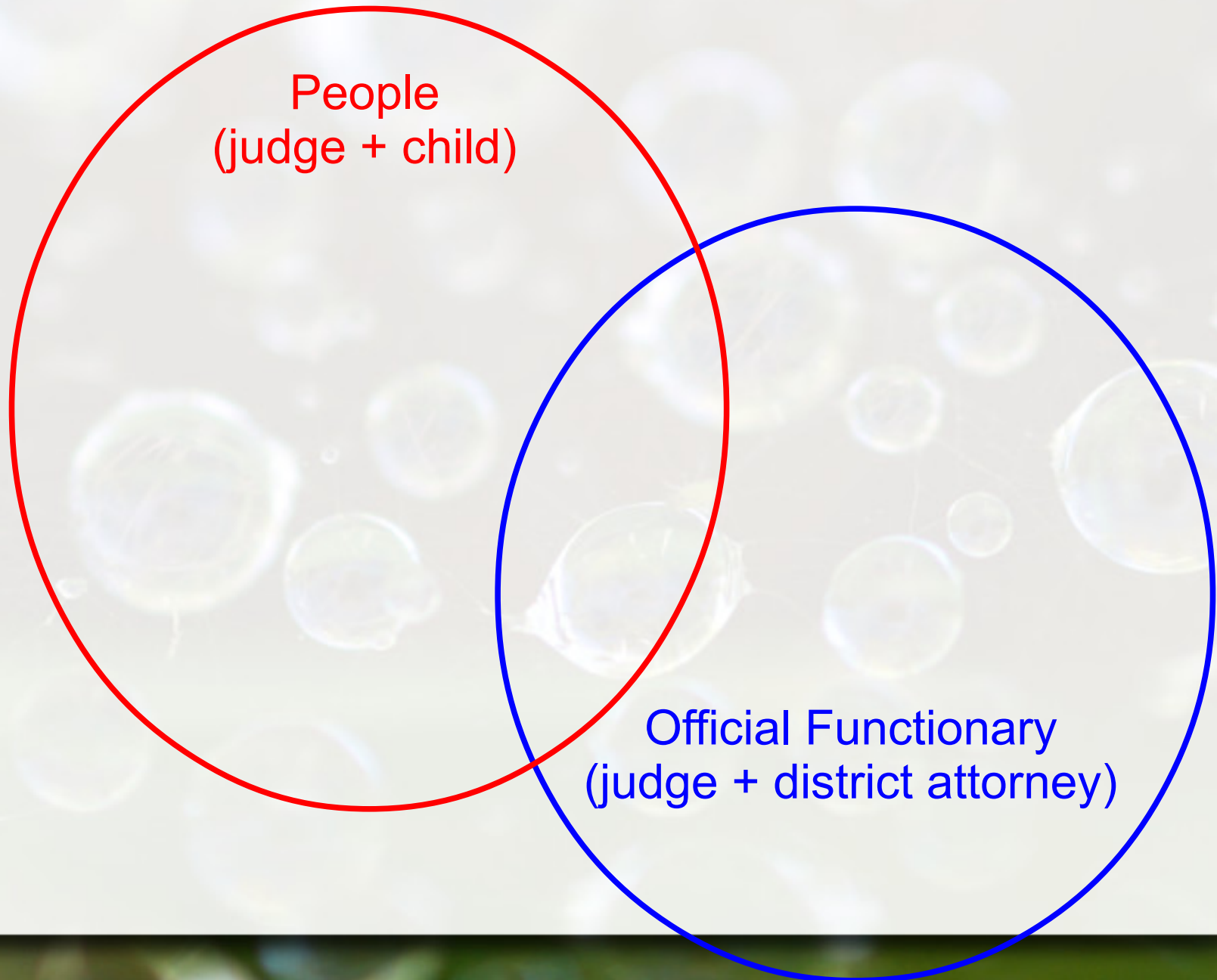
Size of the Shared Universe



Size of the Shared Universe



Size of the Shared Universe



Size of the Shared Universe



People
(judge + child)

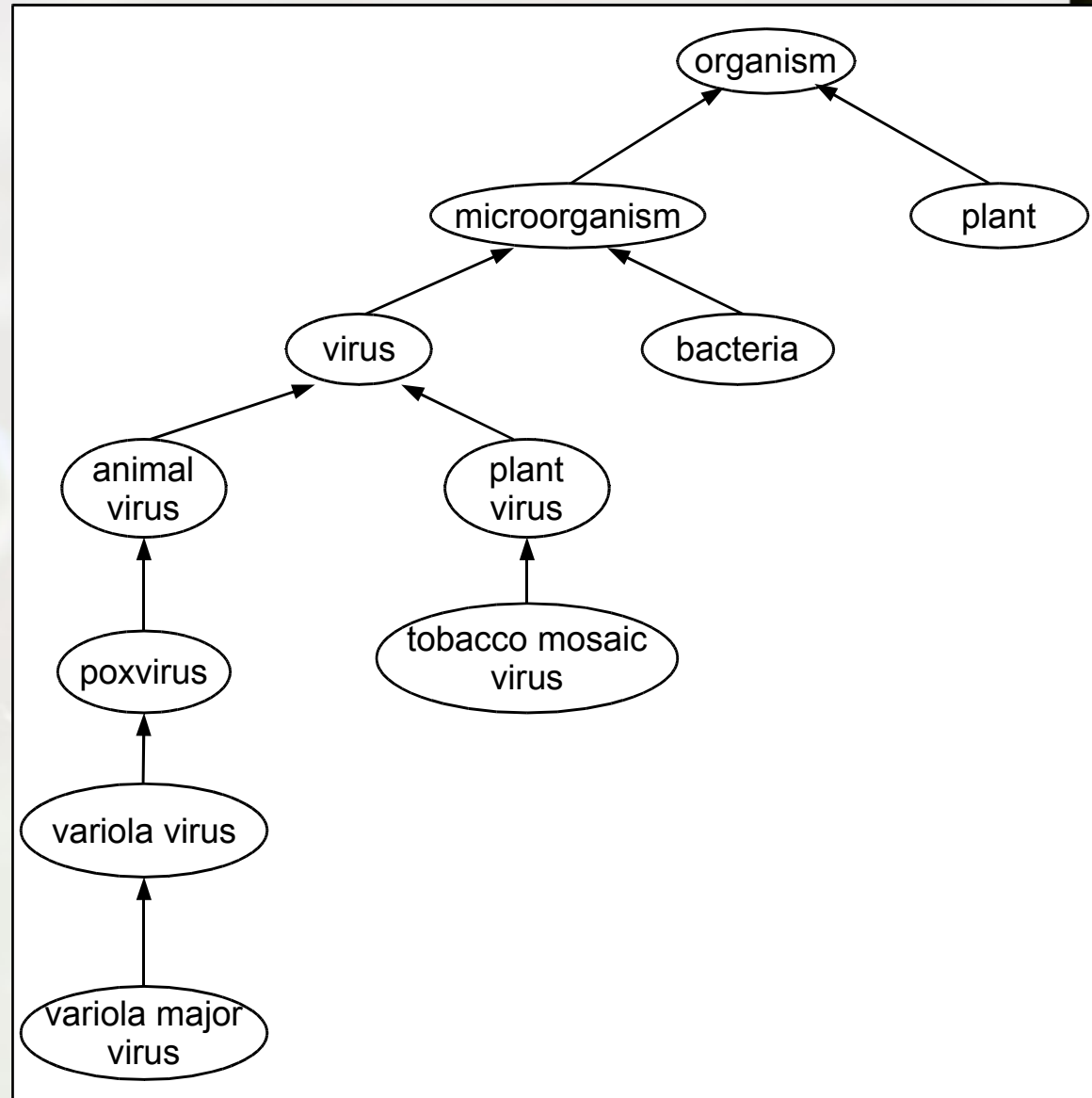
Official Functionary
(judge + district attorney)

Similarity

- What is more similar of a tobacco mosaic virus?

a) variola major virus

b) plant



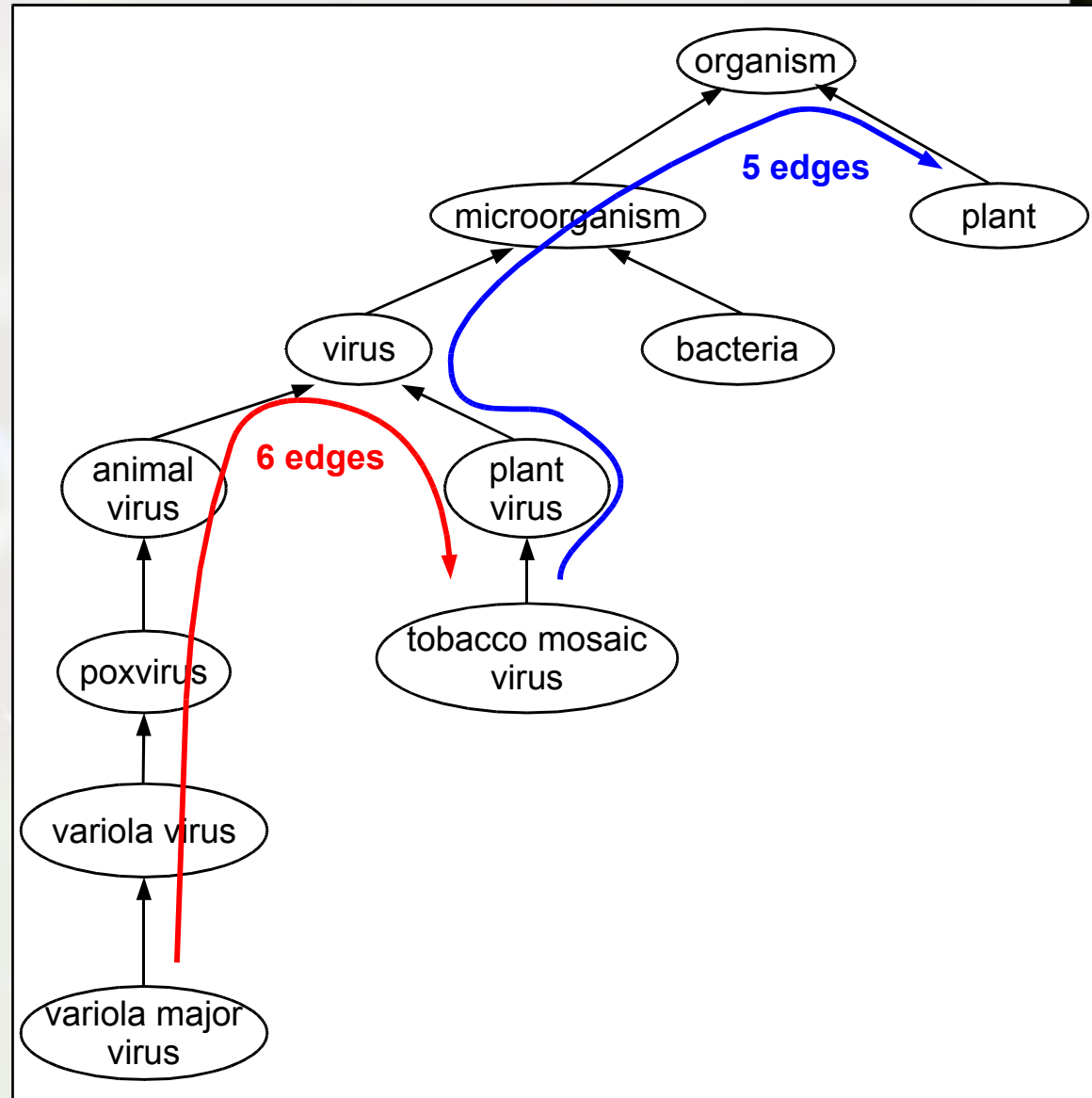
Similarity

Minimum path length

- What is more similar of a tobacco mosaic virus?

a) variola major virus

b) plant (?)



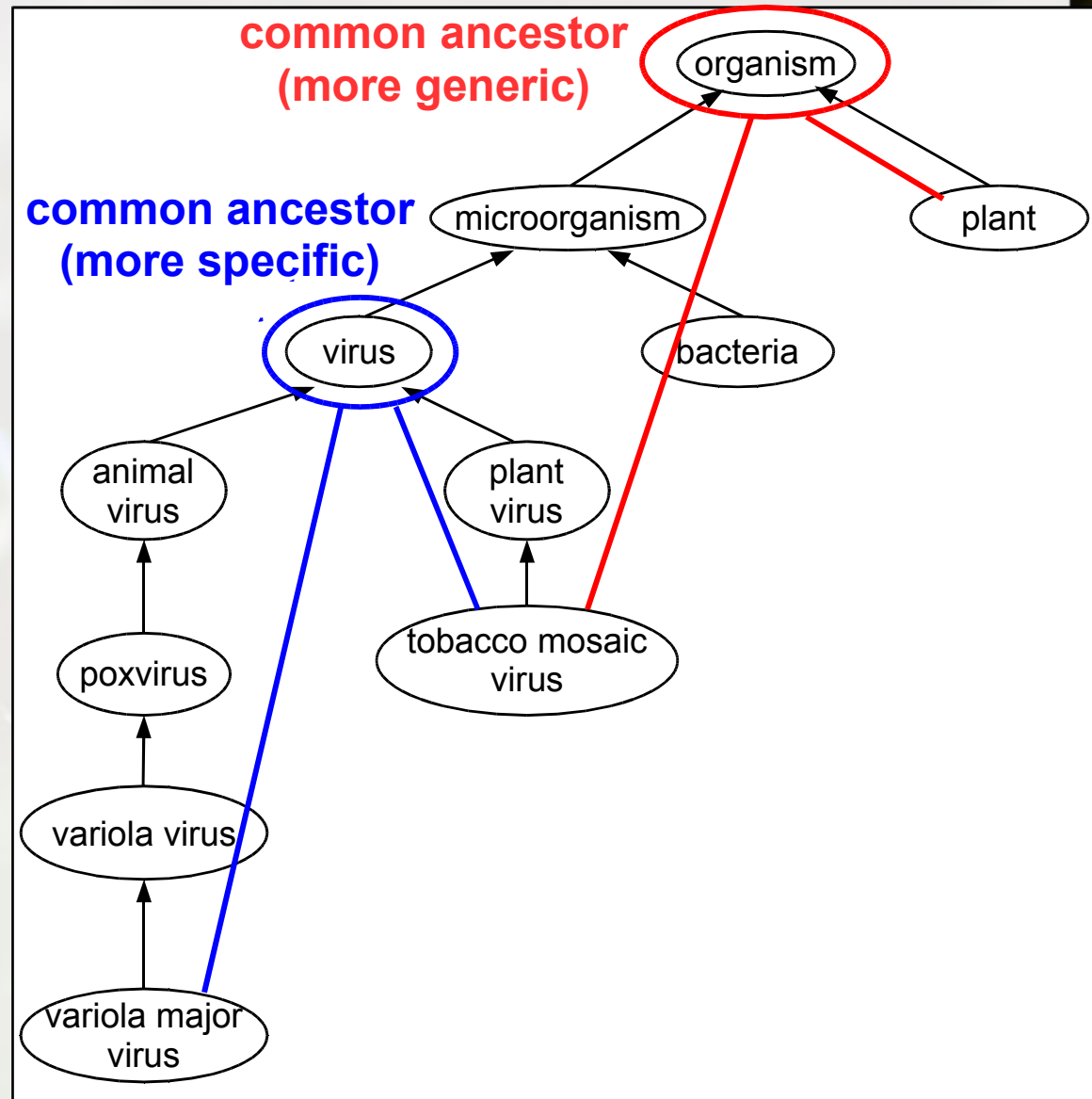
Similarity

More specialized common ancestor

- What is more similar of a tobacco mosaic virus?

a) variola major virus

b) plant



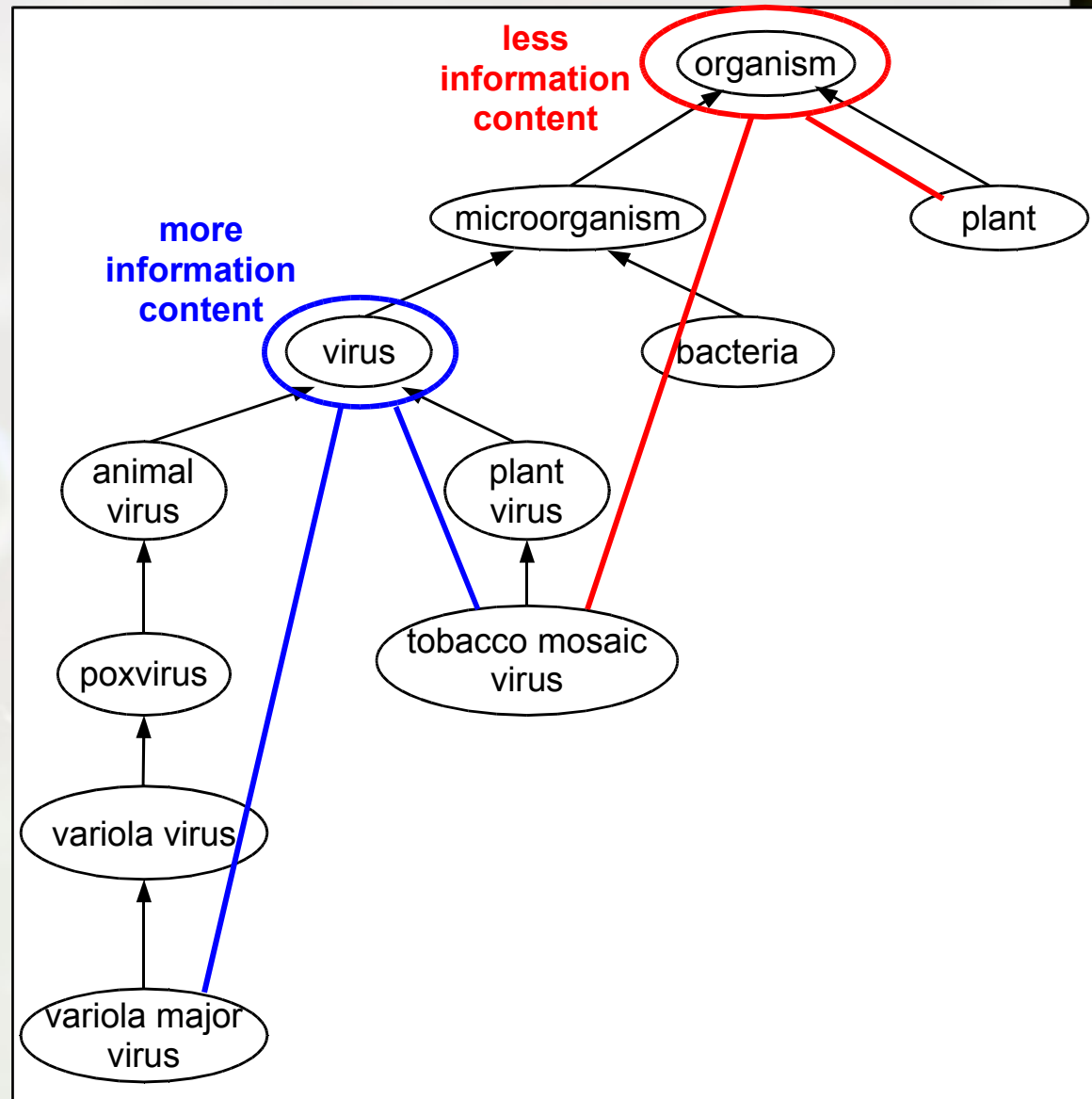
Similarity

Maximum information content

- What is more similar of a tobacco mosaic virus?

a) **variola major virus**

b) **plant**



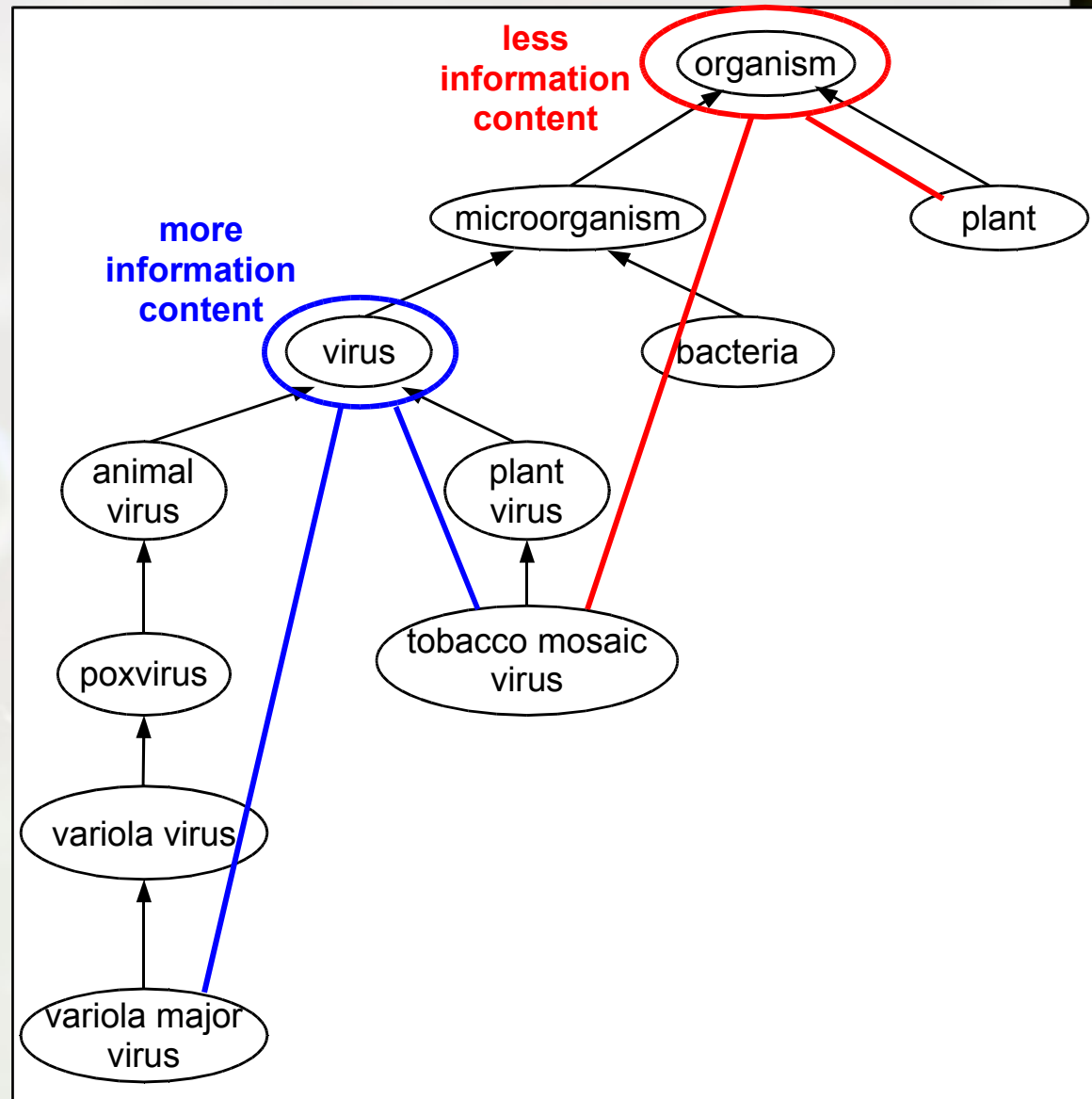
Similarity

Maximum information content

- What is more similar of a tobacco mosaic virus?

a) **variola major virus**

b) **plant**



Graph Databases

Graph Model

Graph Model

RDF Graph

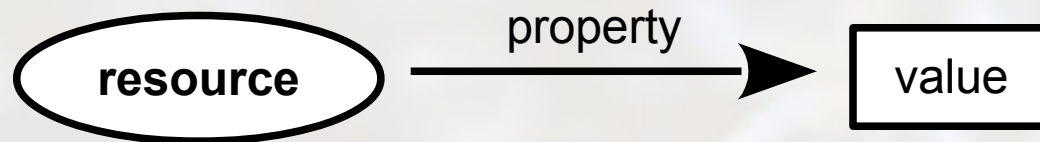
RDF Graph

- Triple model: resource, property, value

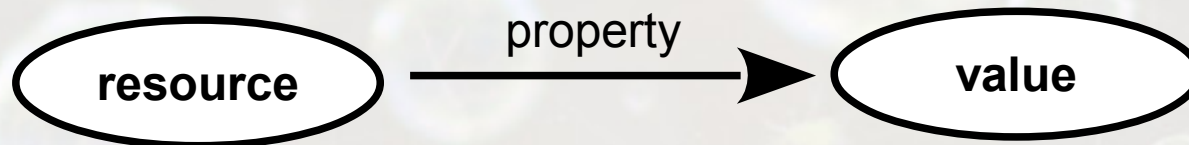


RDF Graph

- Triple model: resource, property, value



- The value can be another resource

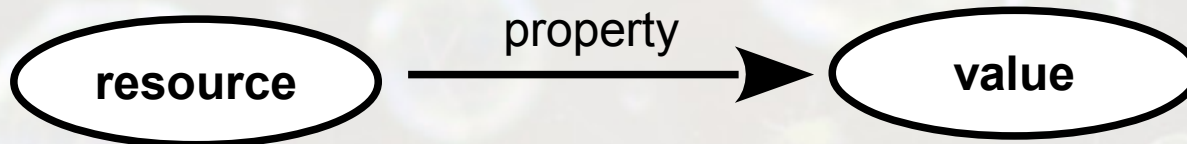


RDF Graph

- Triple model: resource, property, value



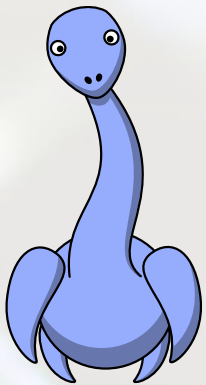
- The value can be another resource



- ... described by a property, value



Plesiosaurus dolichodeirus in RDF



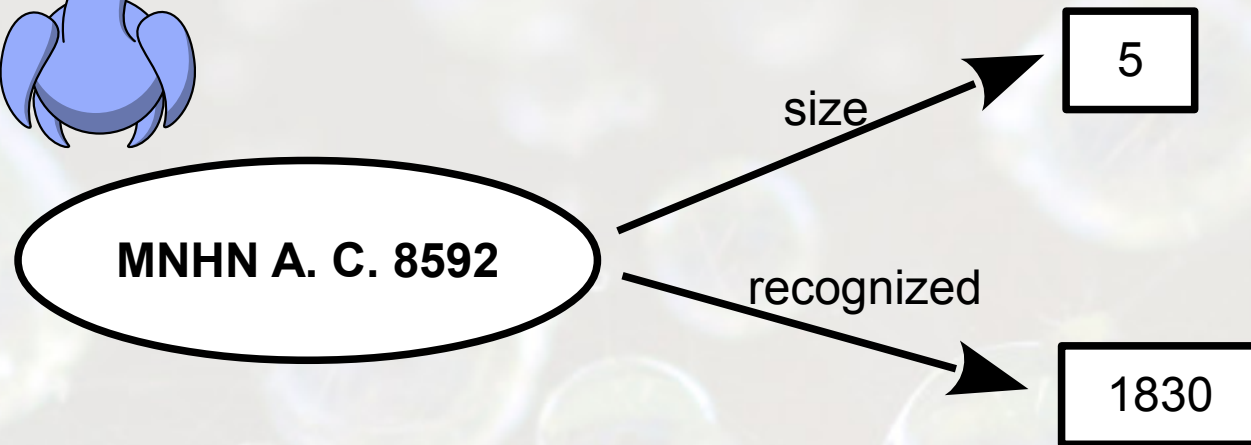
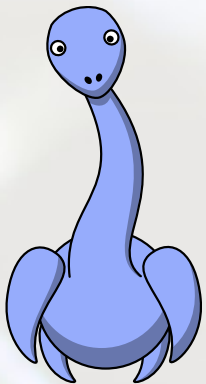
MNHN A. C. 8592

size

5

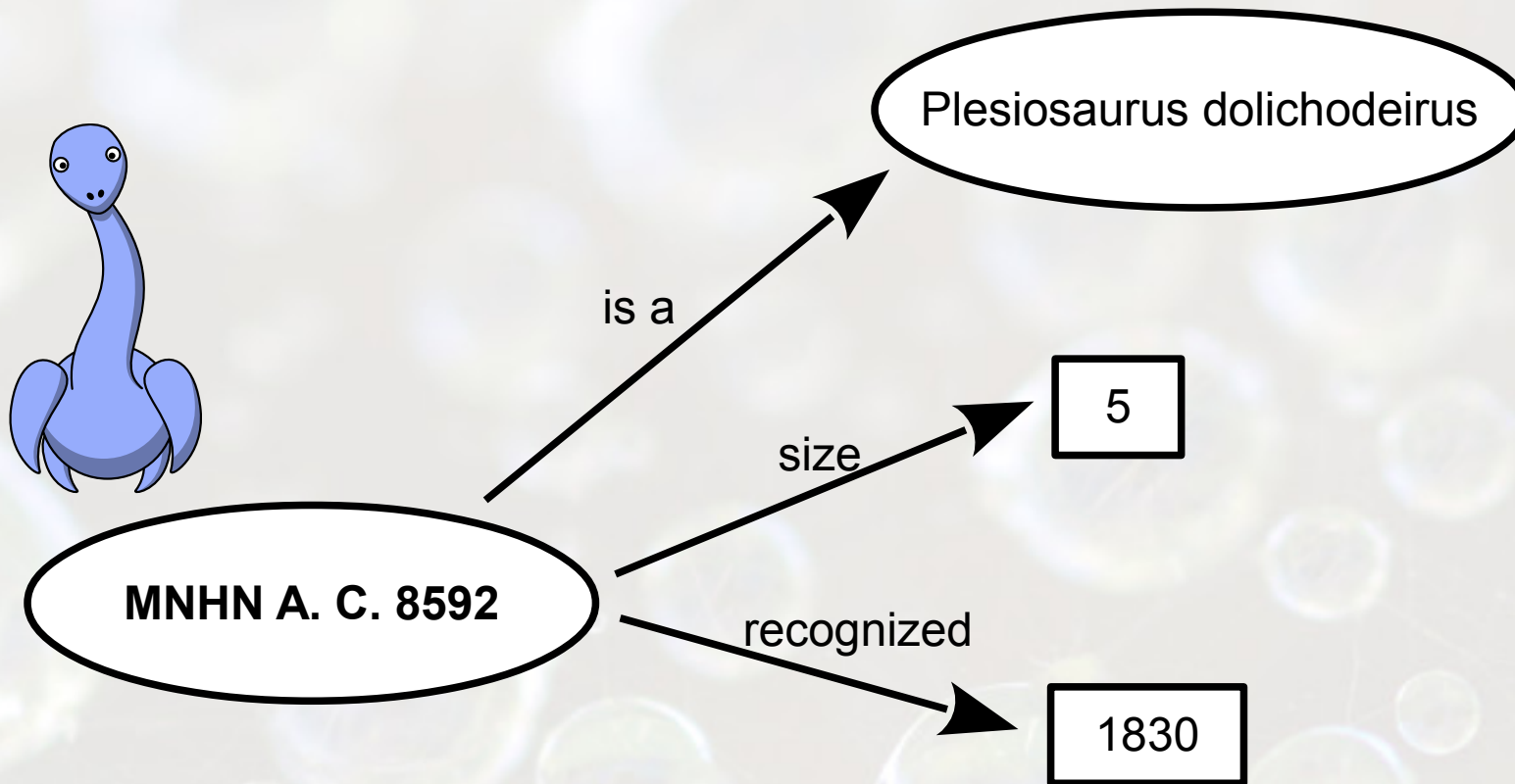
- Triple: <MNHN A. C. 8592>, size, 5

Plesiosaurus dolichodeirus in RDF



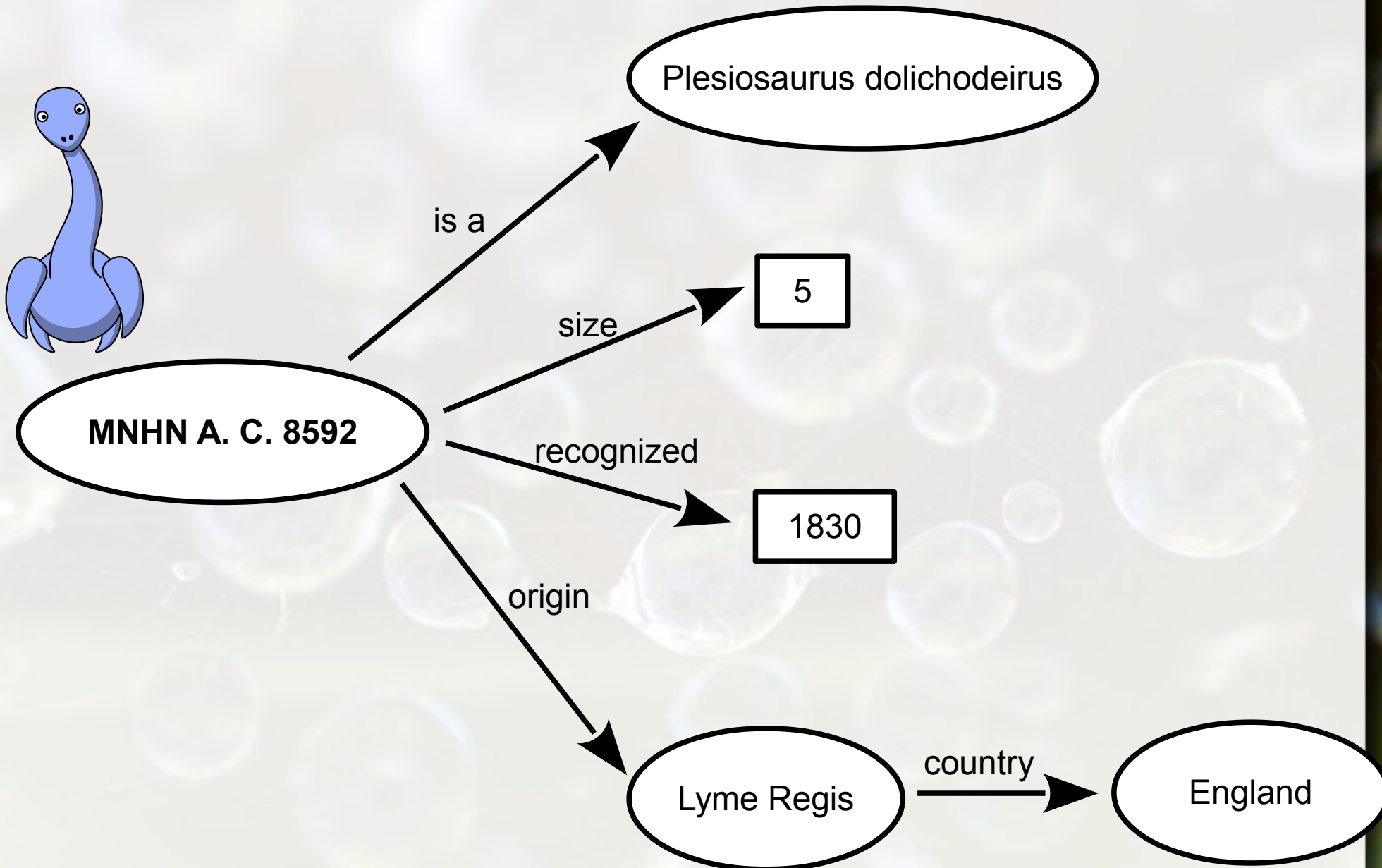
- Triple: <MNHN A. C. 8592>, size, 5
- Triple: <MNHN A. C. 8592>, recognized, 1830

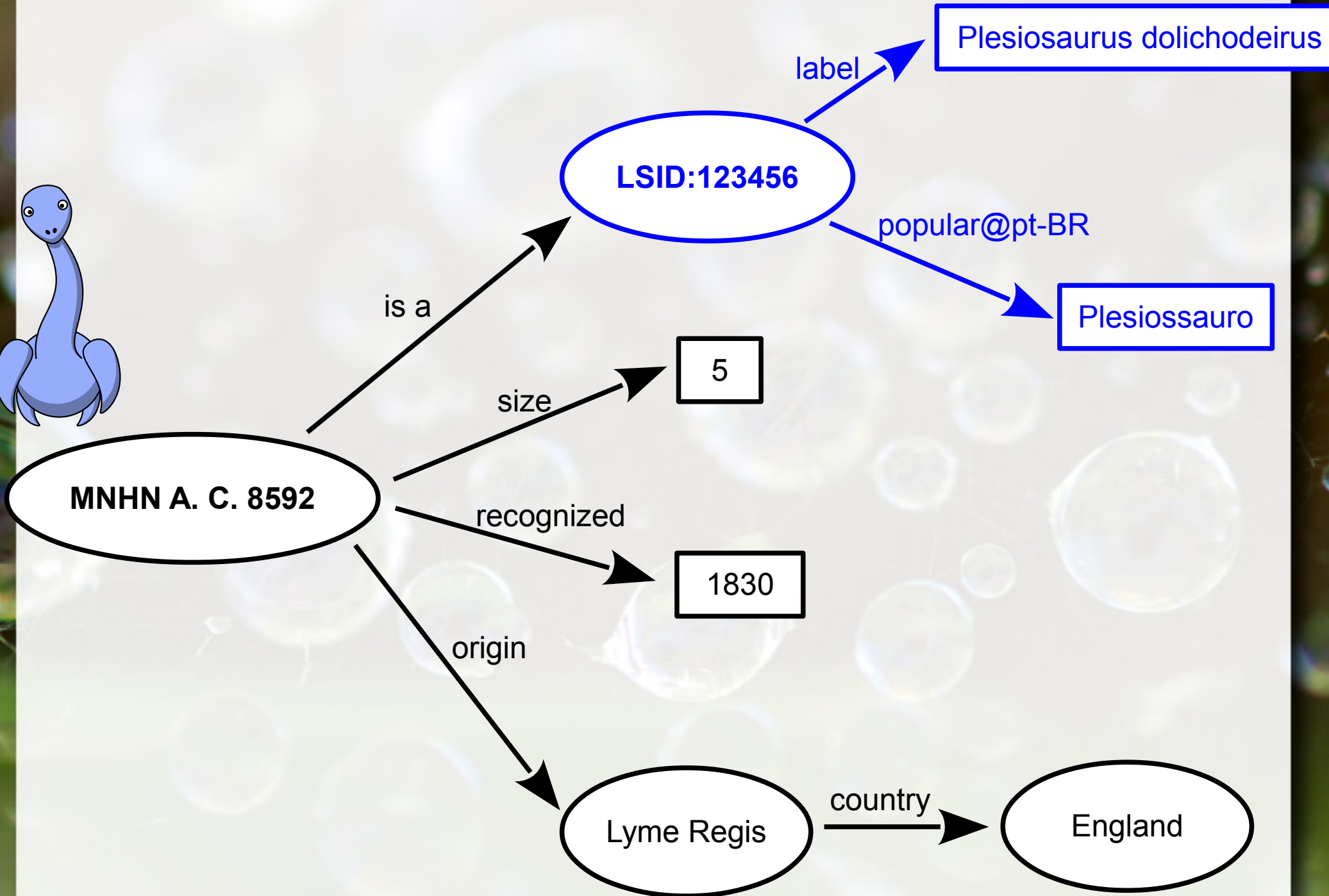
Plesiosaurus dolichodeirus in RDF



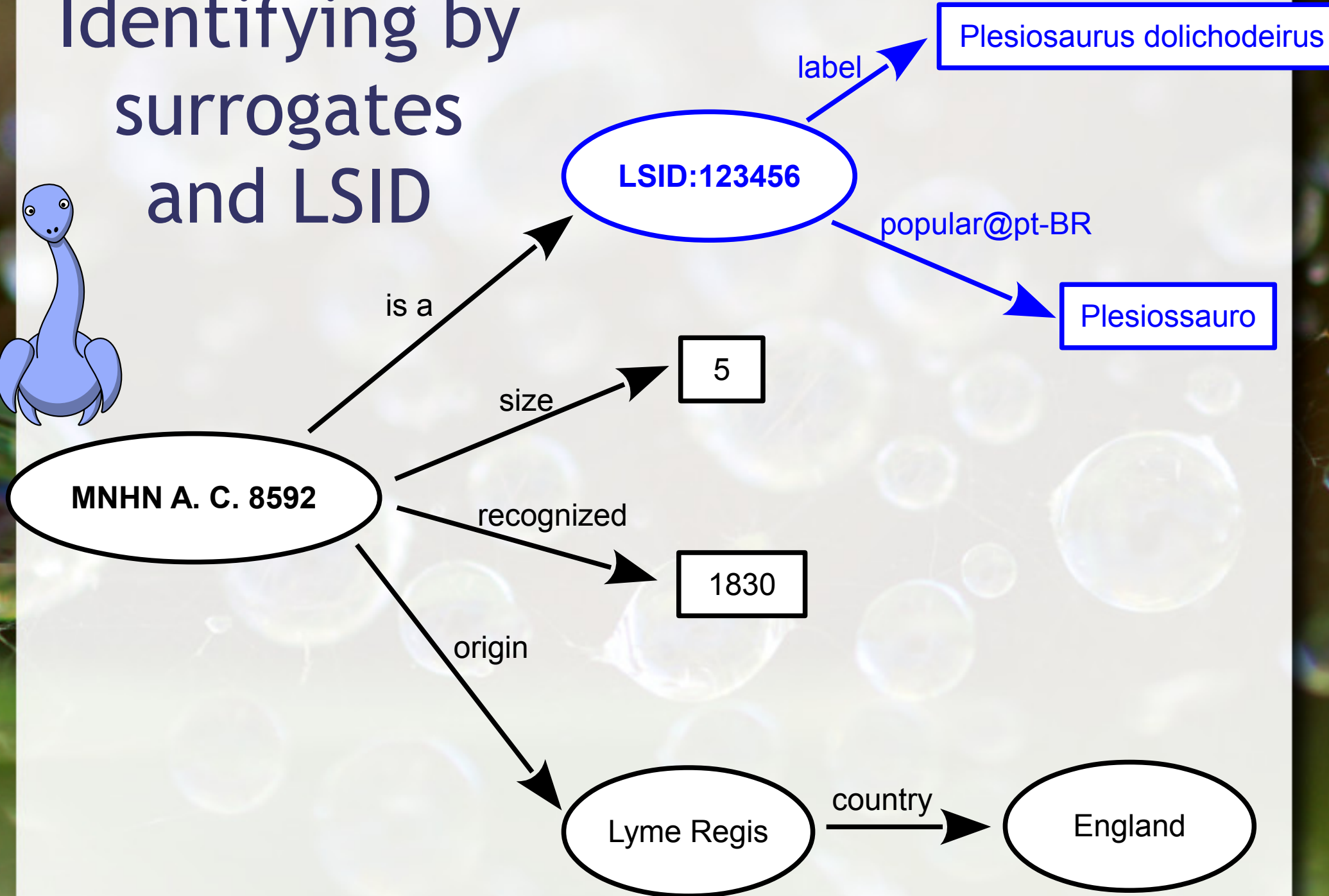
- `<MNHN A. C. 8592>`, `is_a`, `<Plesiosaurus dolichodeirus>`

Plesiosaurus dolichodeirus in RDF

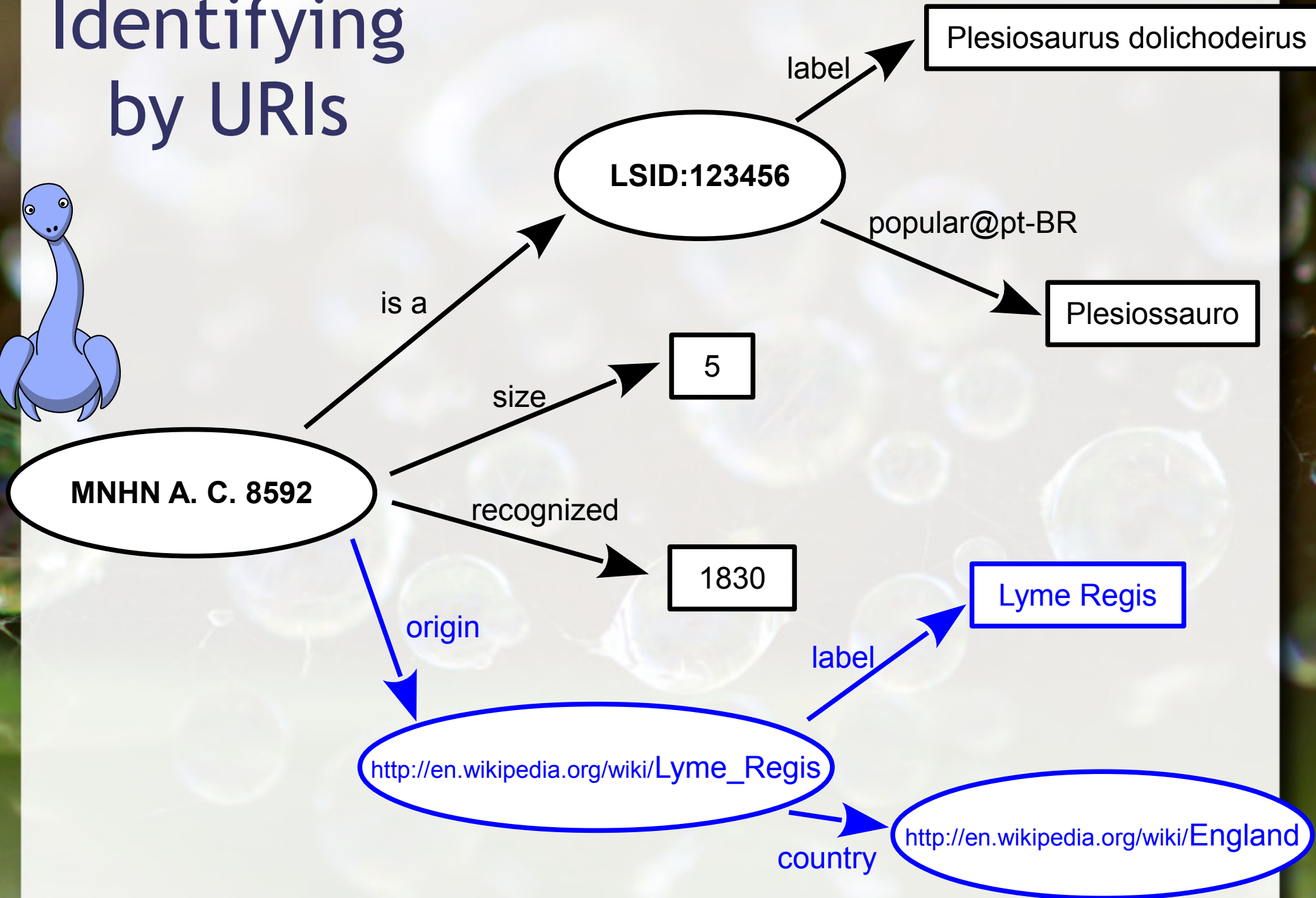




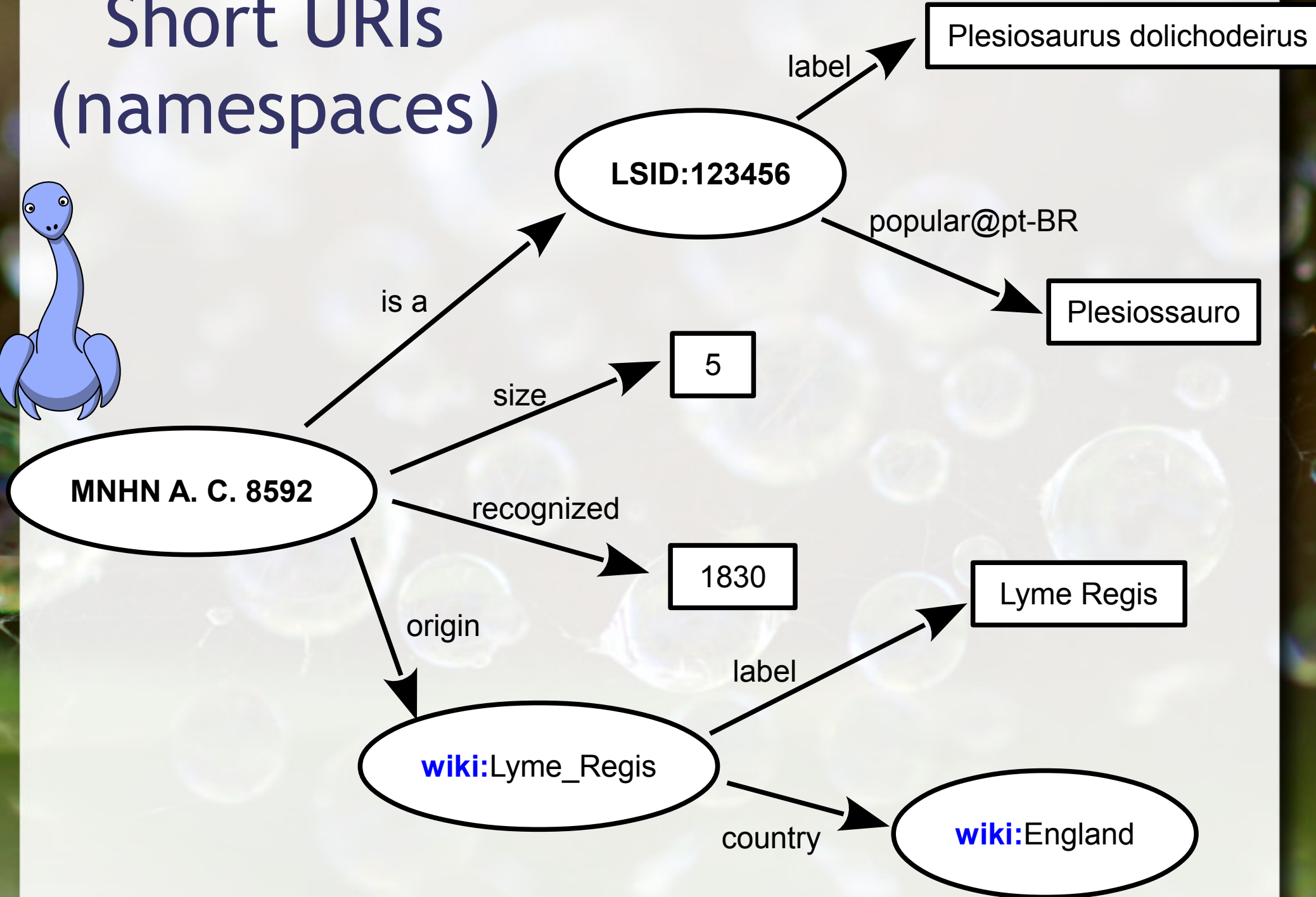
Identifying by surrogates and LSID



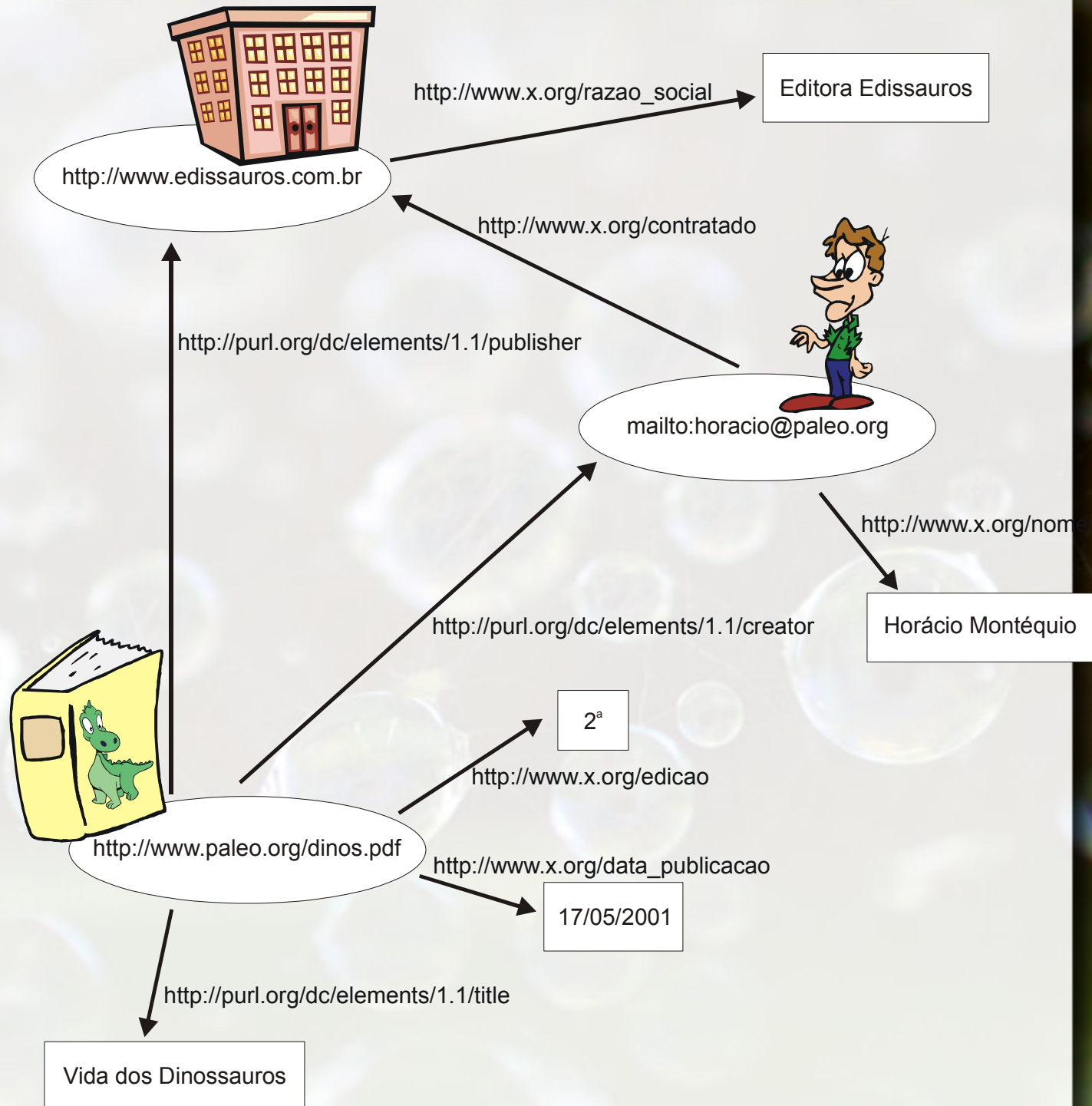
Identifying by URIs



Short URIs (namespaces)



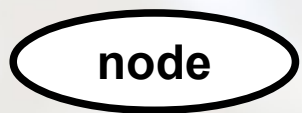
RDF Graph




Graph Model Property Graph

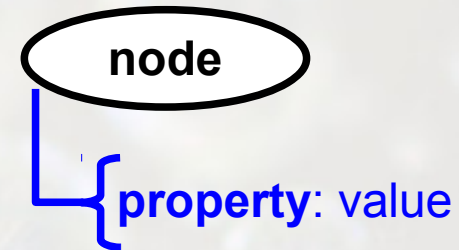
Property Graph

- (node)




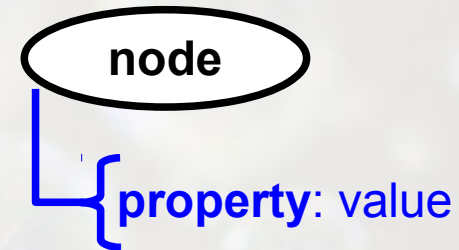
Property Graph

- (node) 
 - Nodes can have properties
(node { property: value })

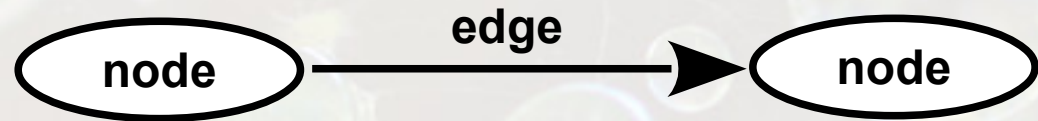


Property Graph


- (node) 
 - Nodes can have properties
(node { property: value })

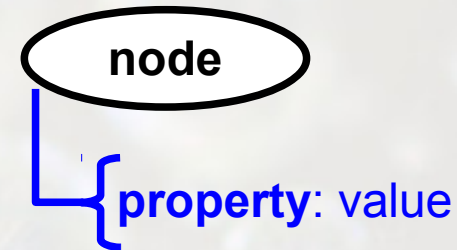


- (node)-[edge]->(node)

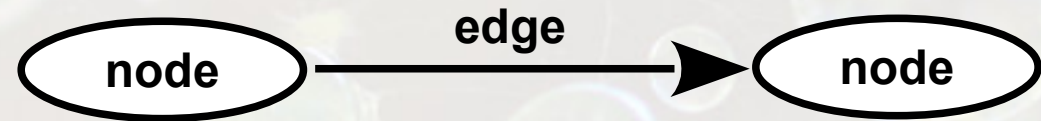


Property Graph

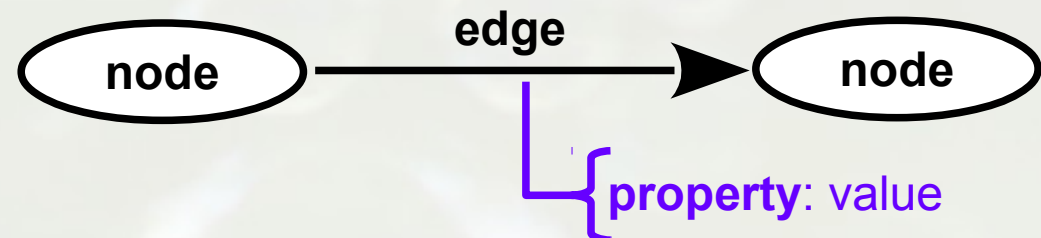
- (node) 
 - Nodes can have properties
(node { property: value })



- (node)-[edge]->(node)



- Edges can have properties
(node)-[edge {property:value}]->(node)



Plesiosaurus in a Property Graph



MNHN A. C. 8592

Plesiosaurus in a Property Graph

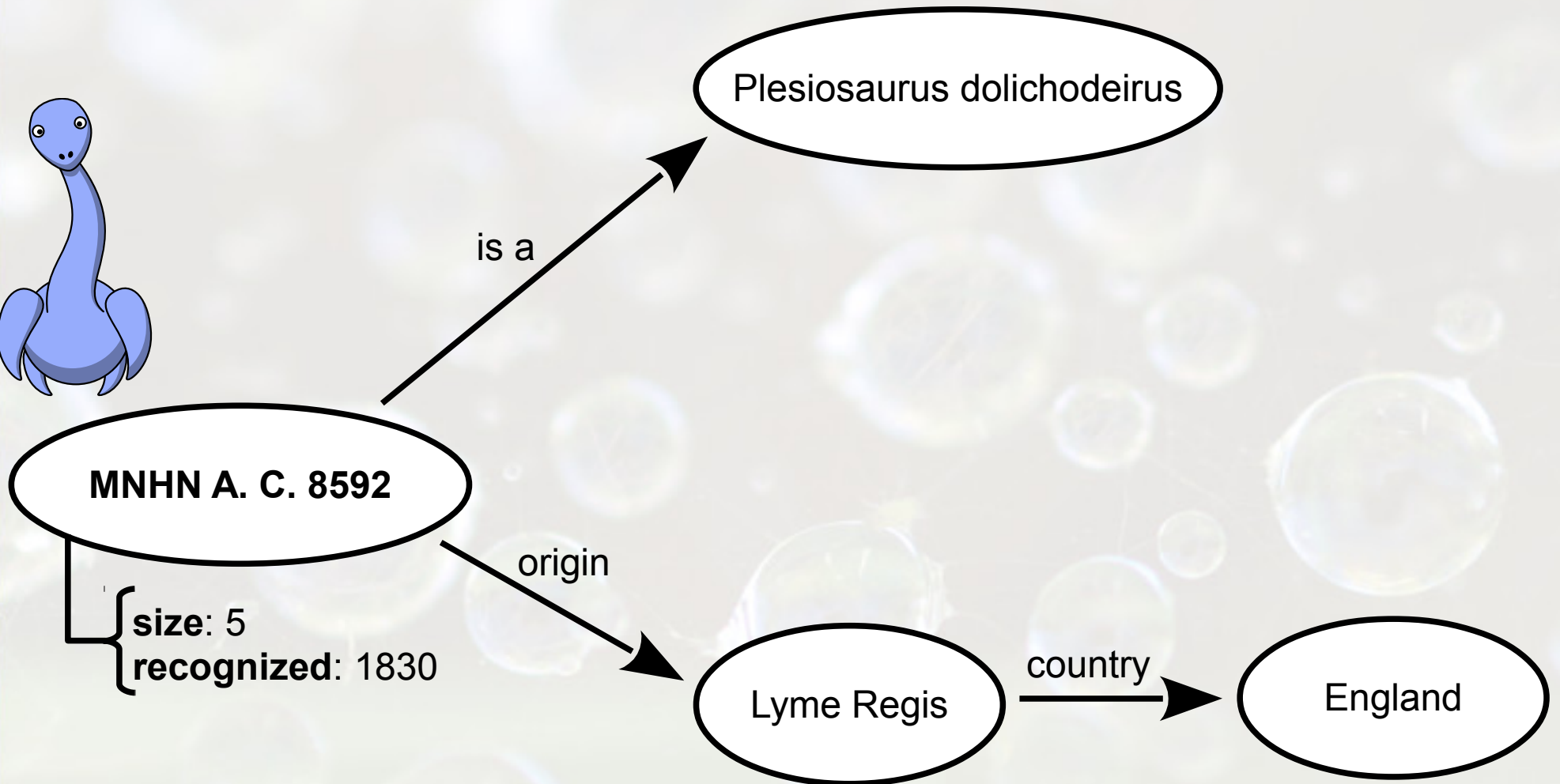


MNHN A. C. 8592

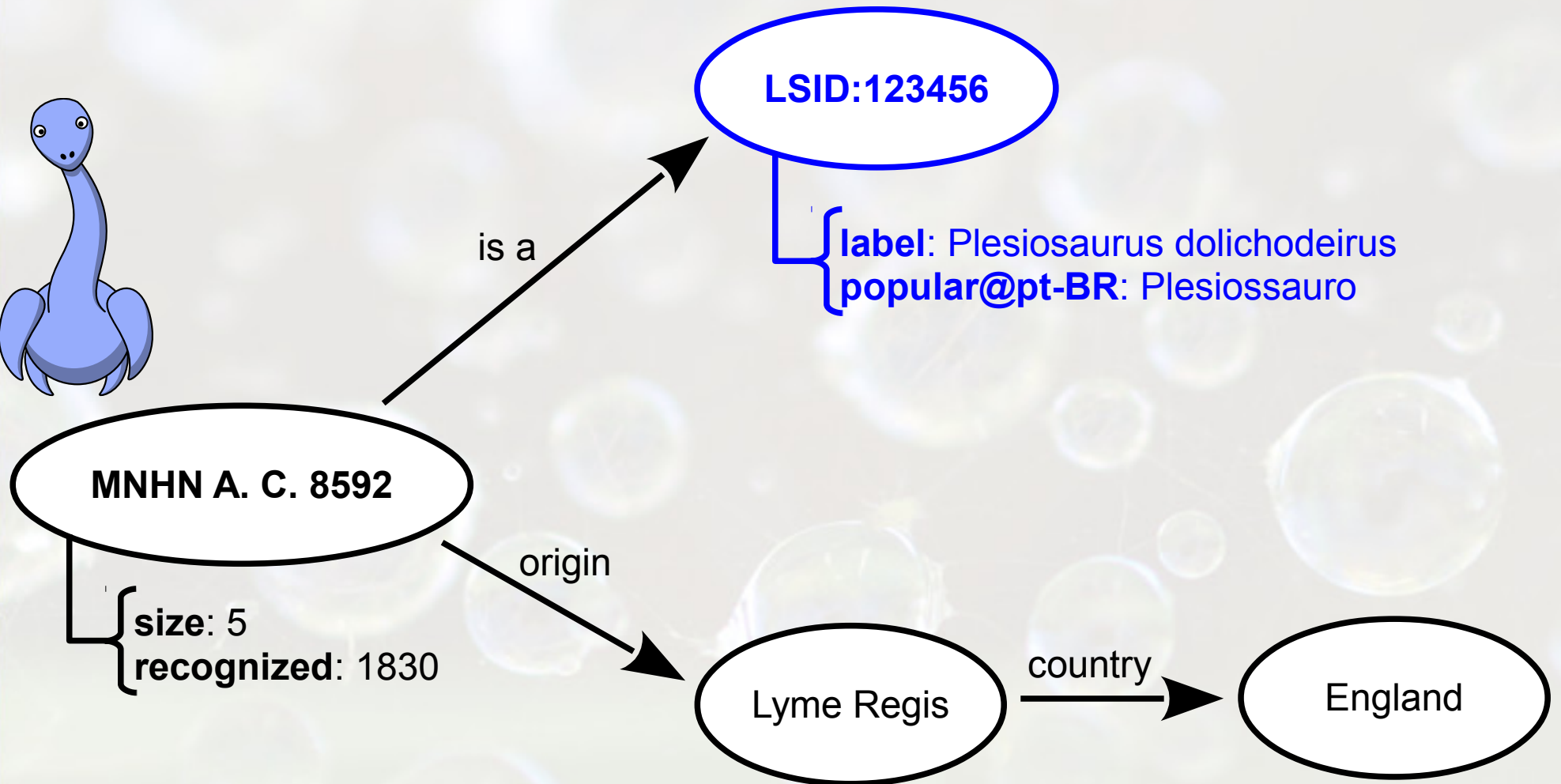
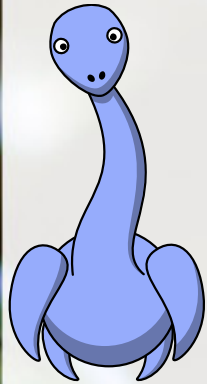
size: 5

recognized: 1830

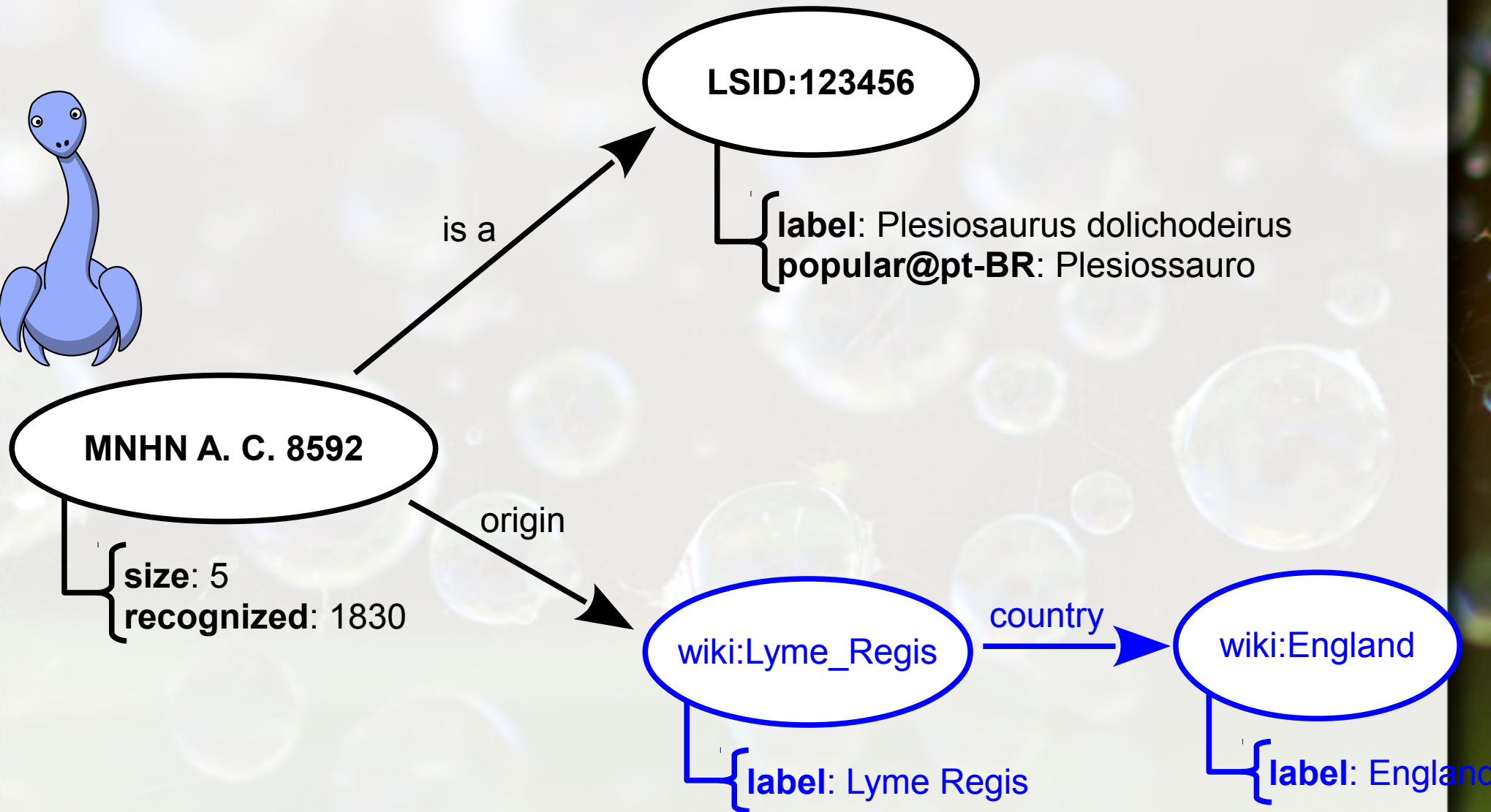
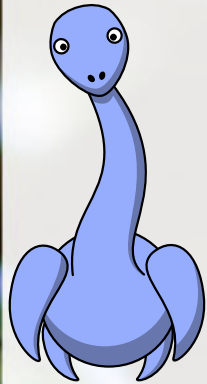
Plesiosaurus in a Property Graph



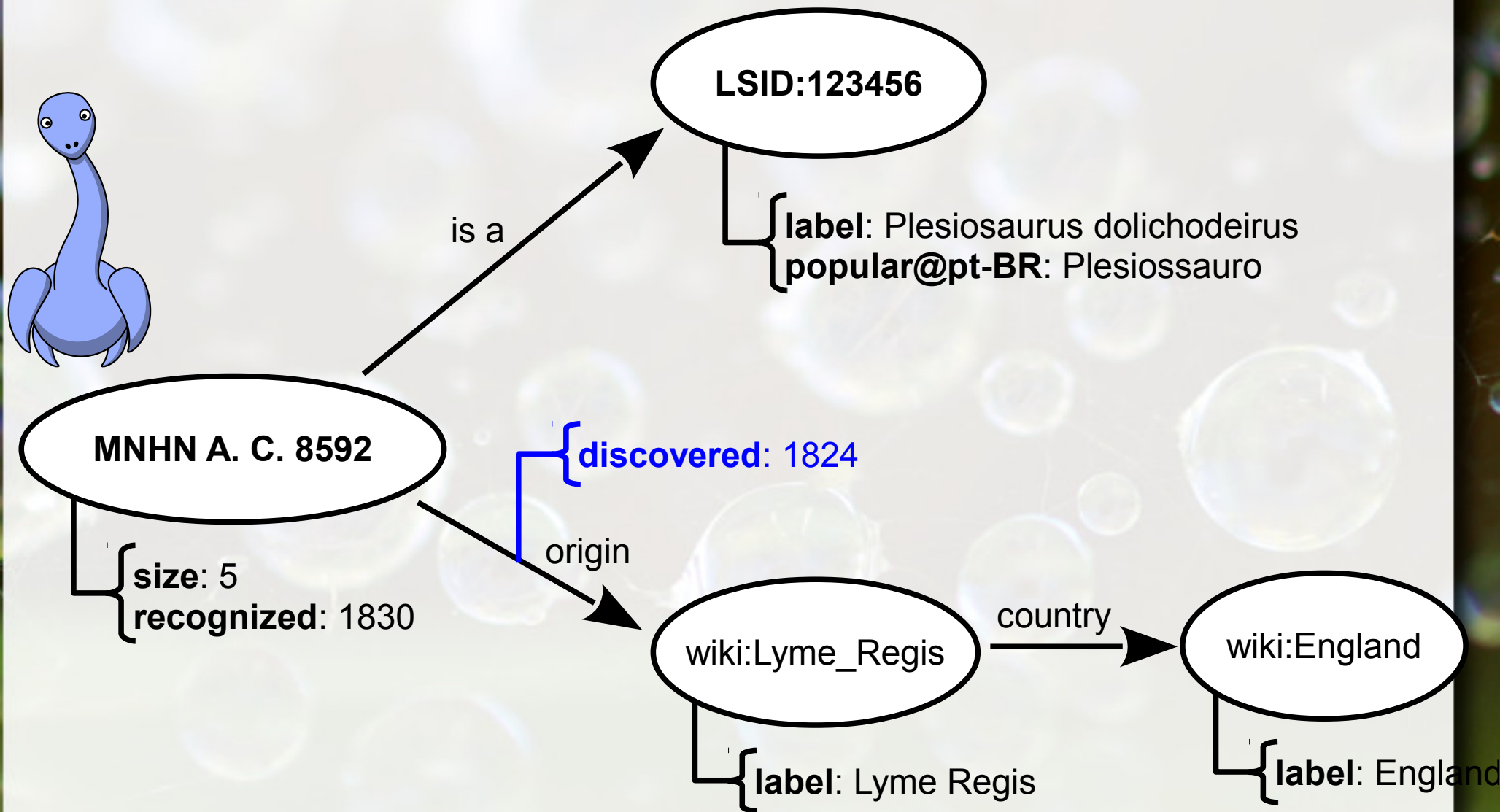
LSID Surrogate



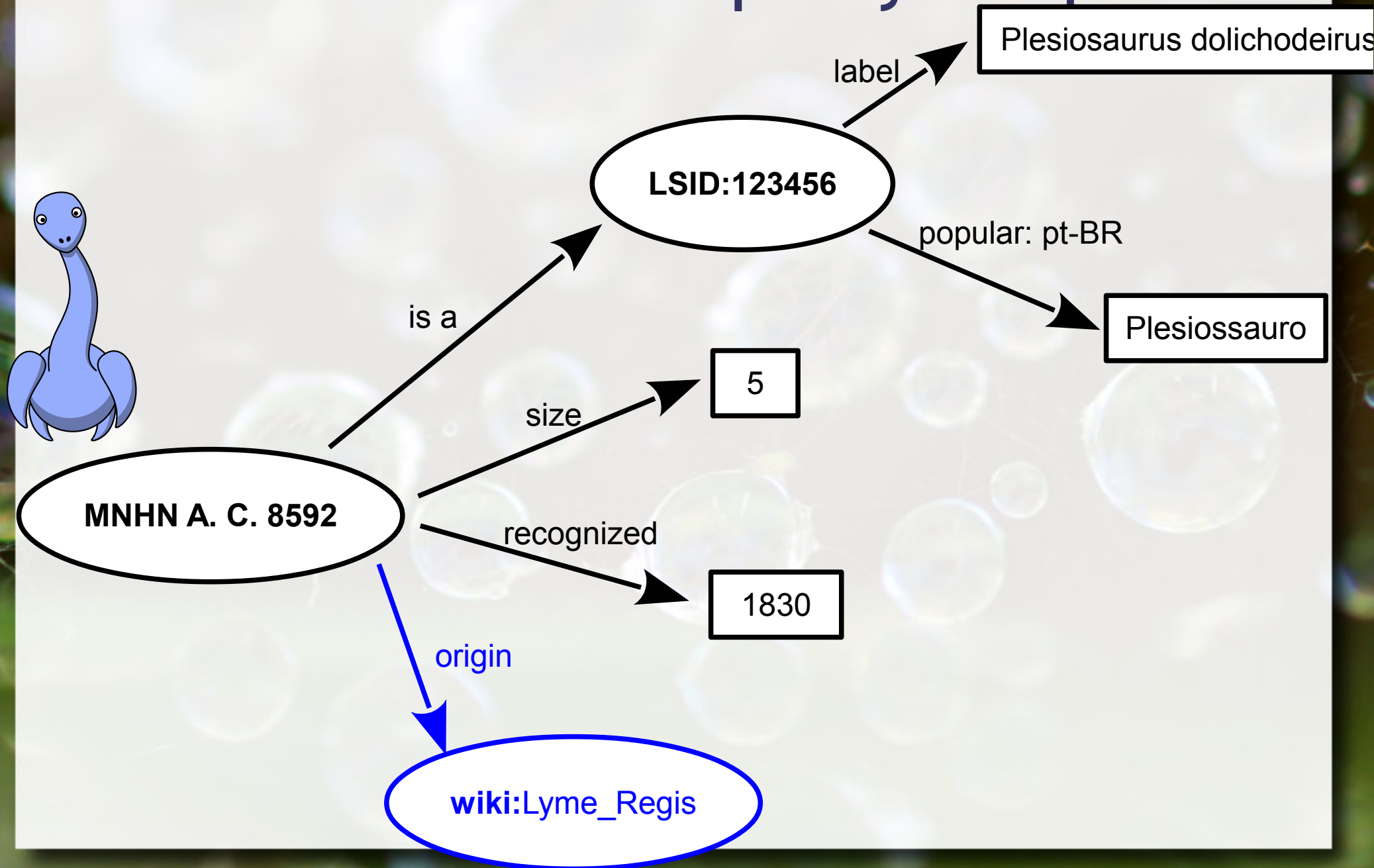
URI Identifiers



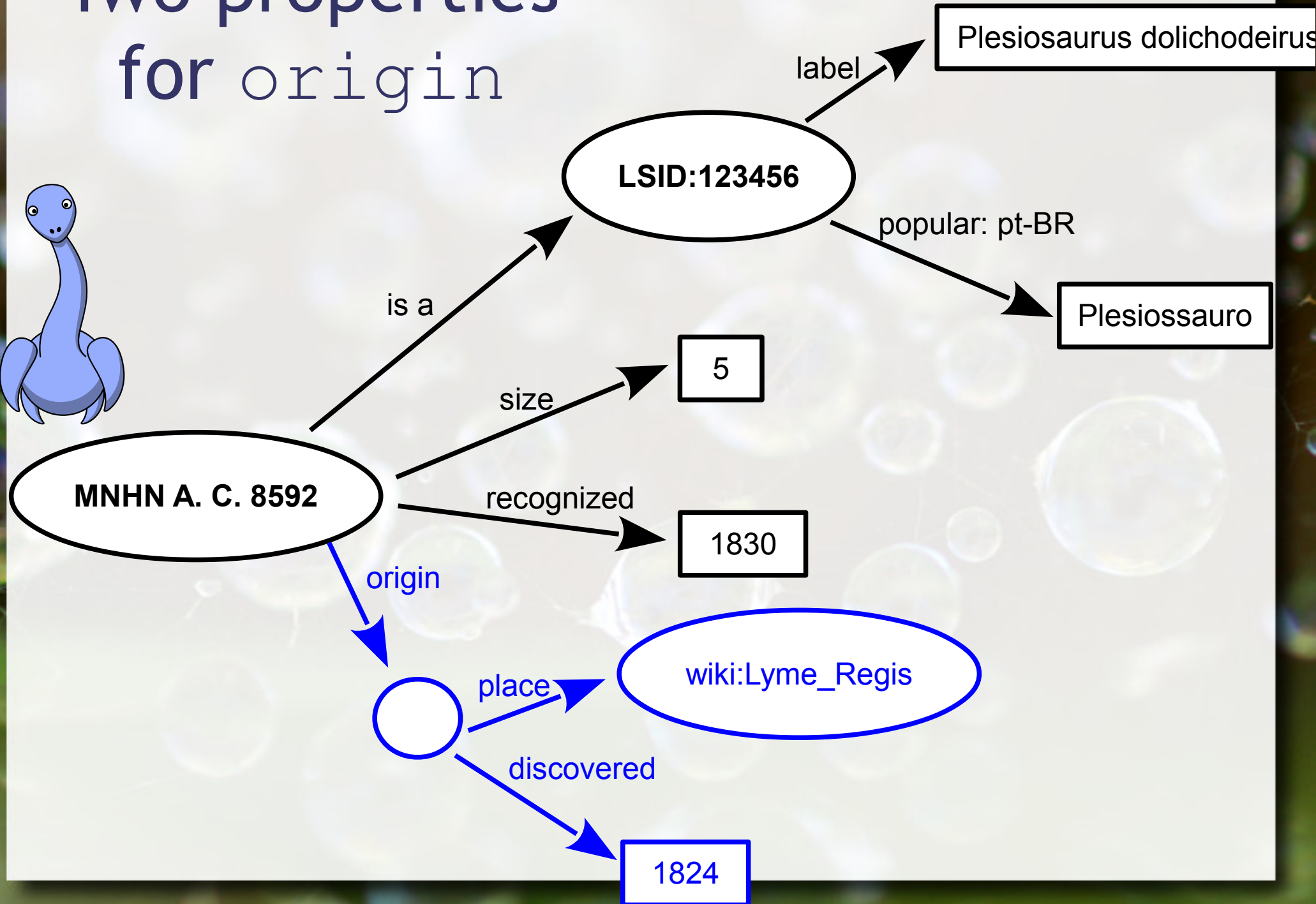
Property in the `origin` edge



Back to the Property Graph



Two properties for origin



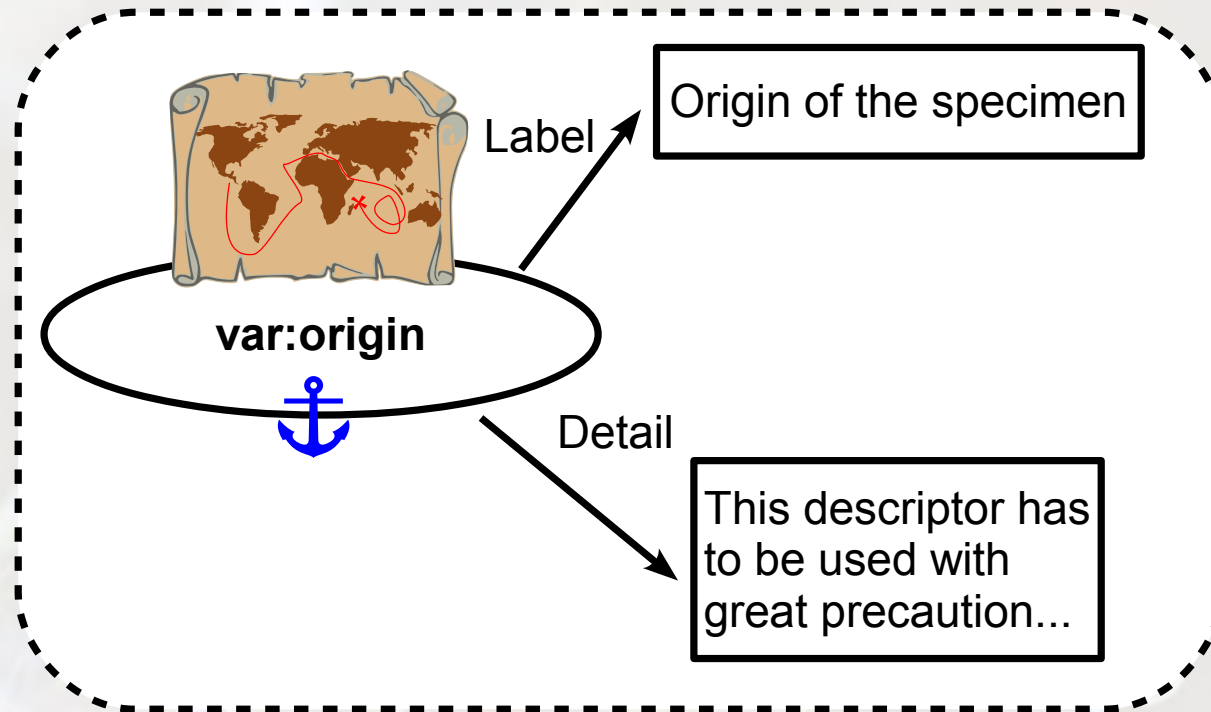
References

- GOMES-JR, L. C.; JENSEN, R.; SANTANCHÈ, A. (2013) Towards Query Model Integration: Topology-aware, IR-inspired Metrics for Declarative Graph Querying, II Workshop on Querying Graph Structured Data 2013, Genoa.

André Santanchè

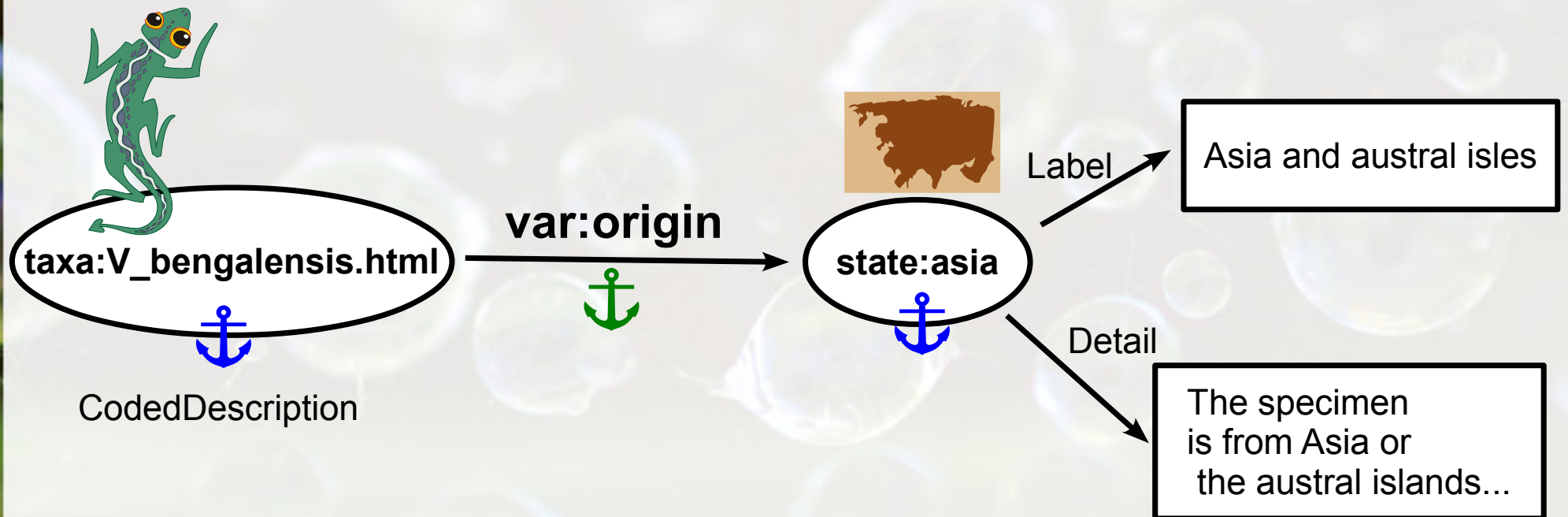
<http://www.ic.unicamp.br/~santanche>

Property

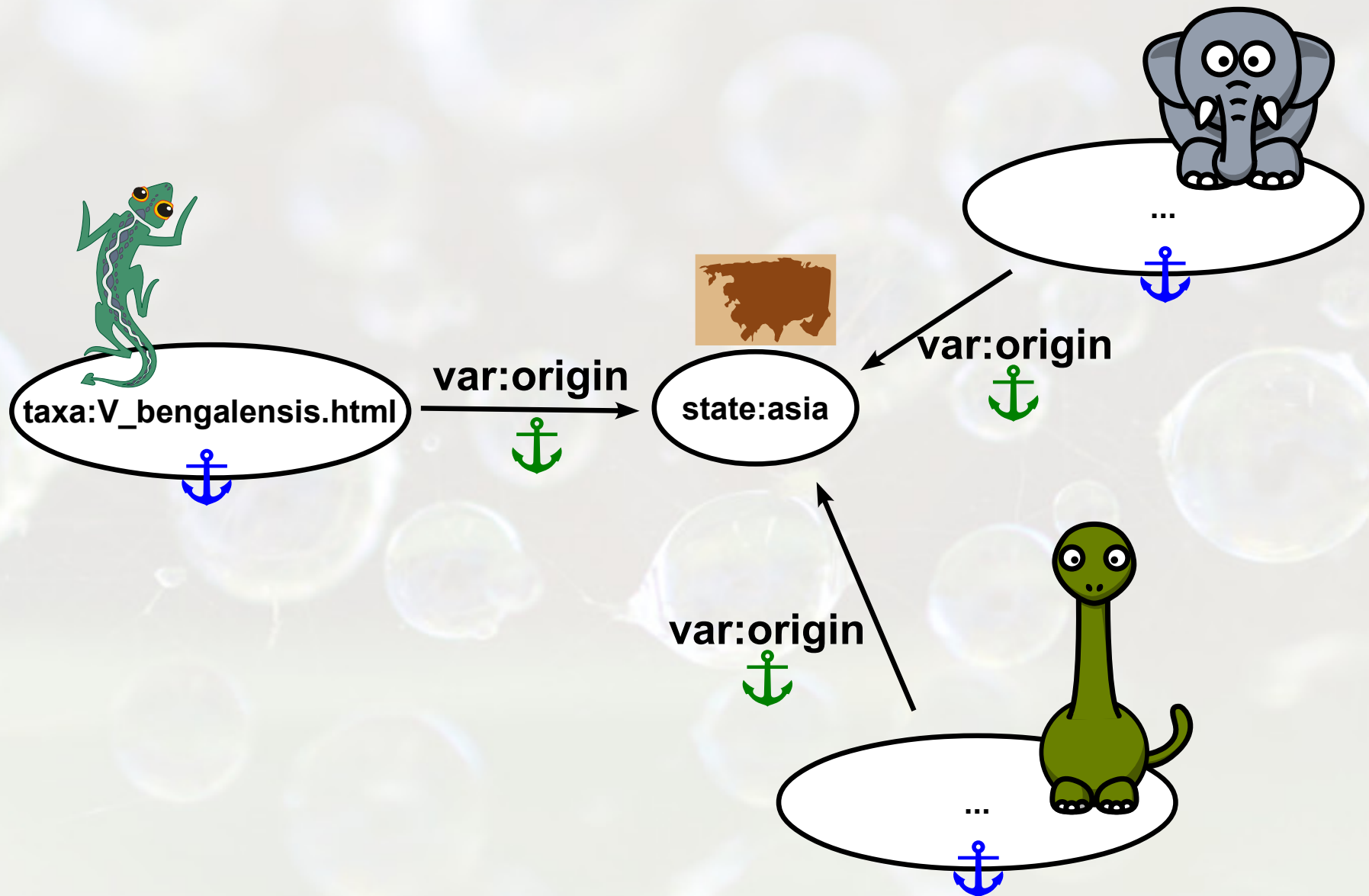


var:origin
⚓

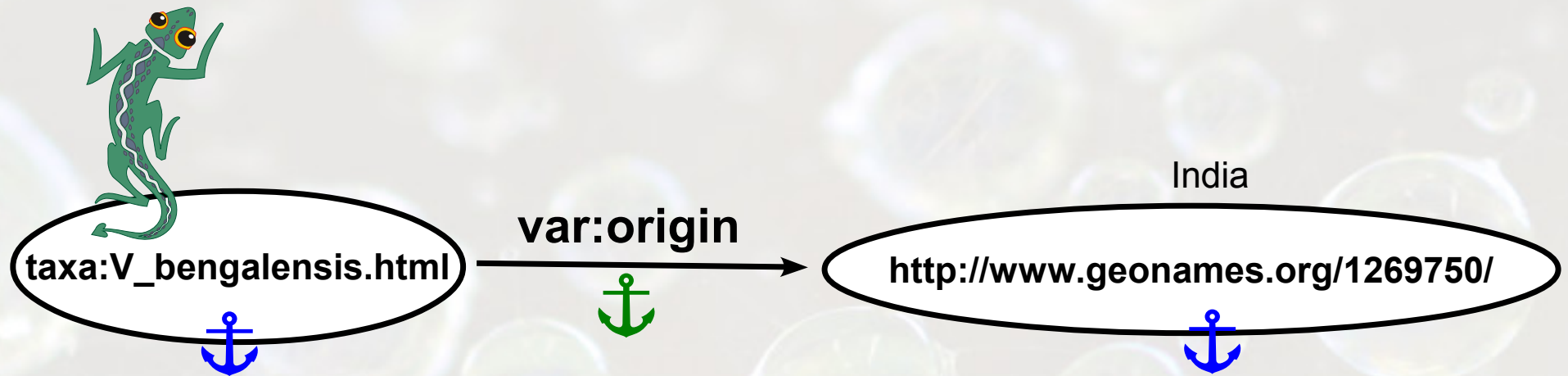
Using the Property



Connected Graph



Origin in GeoNames



Geo Tree

Asia

<http://www.geonames.org/6255147/>



gn:parentFeature

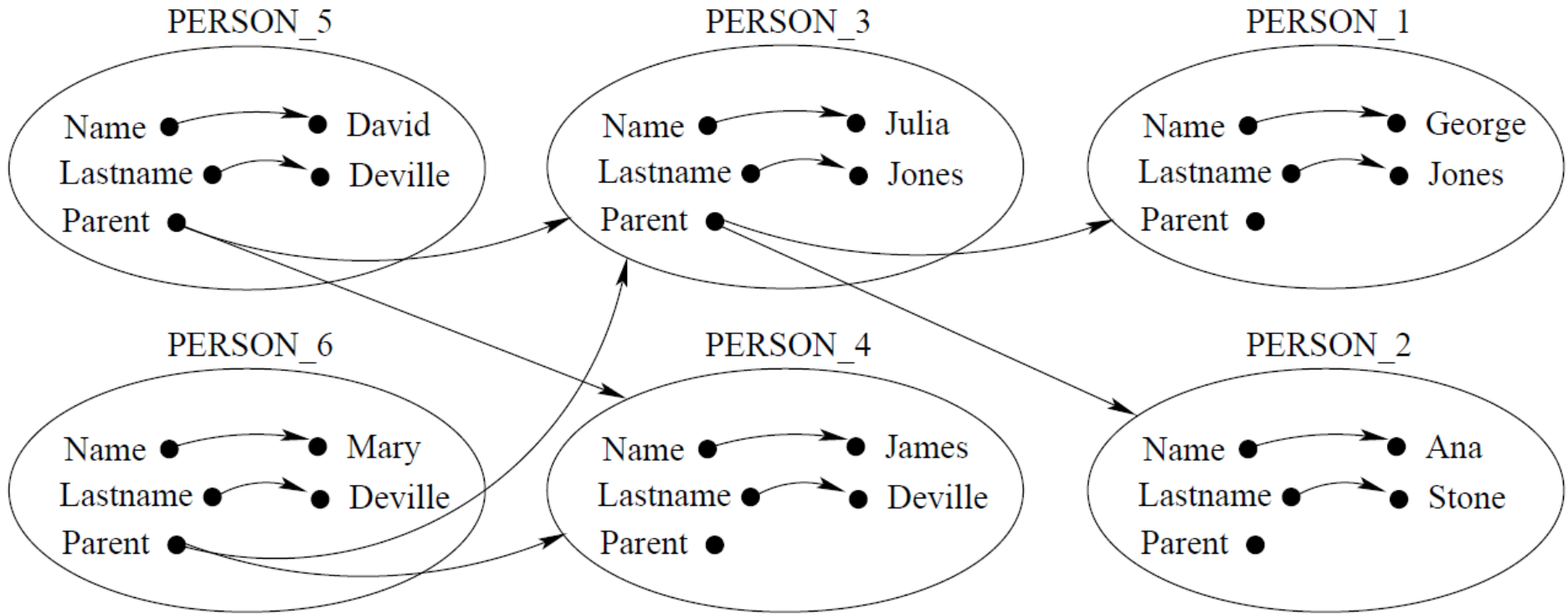


India

<http://www.geonames.org/1269750/>



Hypergraph



(Jaudete Daltio, 2013)

Licença

- Estes slides são concedidos sob uma Licença Creative Commons. Sob as seguintes condições: Atribuição, Uso Não-Comercial e Compartilhamento pela mesma Licença.
- Mais detalhes sobre a referida licença Creative Commons veja no link:
<http://creativecommons.org/licenses/by-nc-sa/3.0/>
- Fotografia de capa e fundos: web-drops por Jeremy Hiebert [<http://www.flickr.com/photos/jeremyhiebert/>] disponível em <http://www.flickr.com/photos/jeremyhiebert/6081389428/>