ITINERIS

Exploring meaning in data: a hands-on course in semantics and analysis for FAIR Research

Module 1: Introduction to Linked Data and semantic technologies

Martina Pulieri

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 "Education and Research" - Component 2: "From research to business" - Investment 3.1: "Fund for the realisation of an integrated system of research and innovation infrastructures"

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani

Agenda

9:00-9:30 Short introduction of the participants and their background

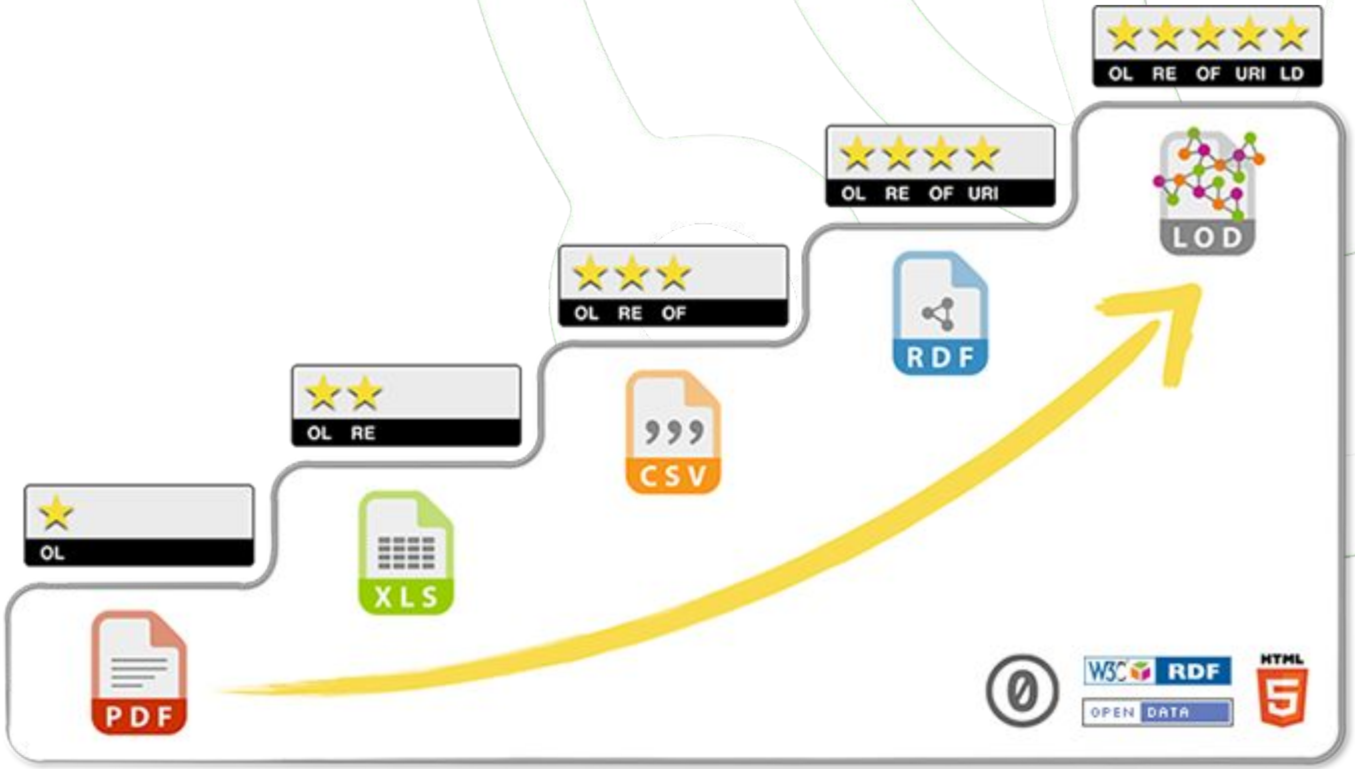
9:30-11:00 Introduction to linked data and semantic technologies

11:00-11:20 Coffee break

11:20-13:00 Key standards and technologies: RDF, OWL, SKOS

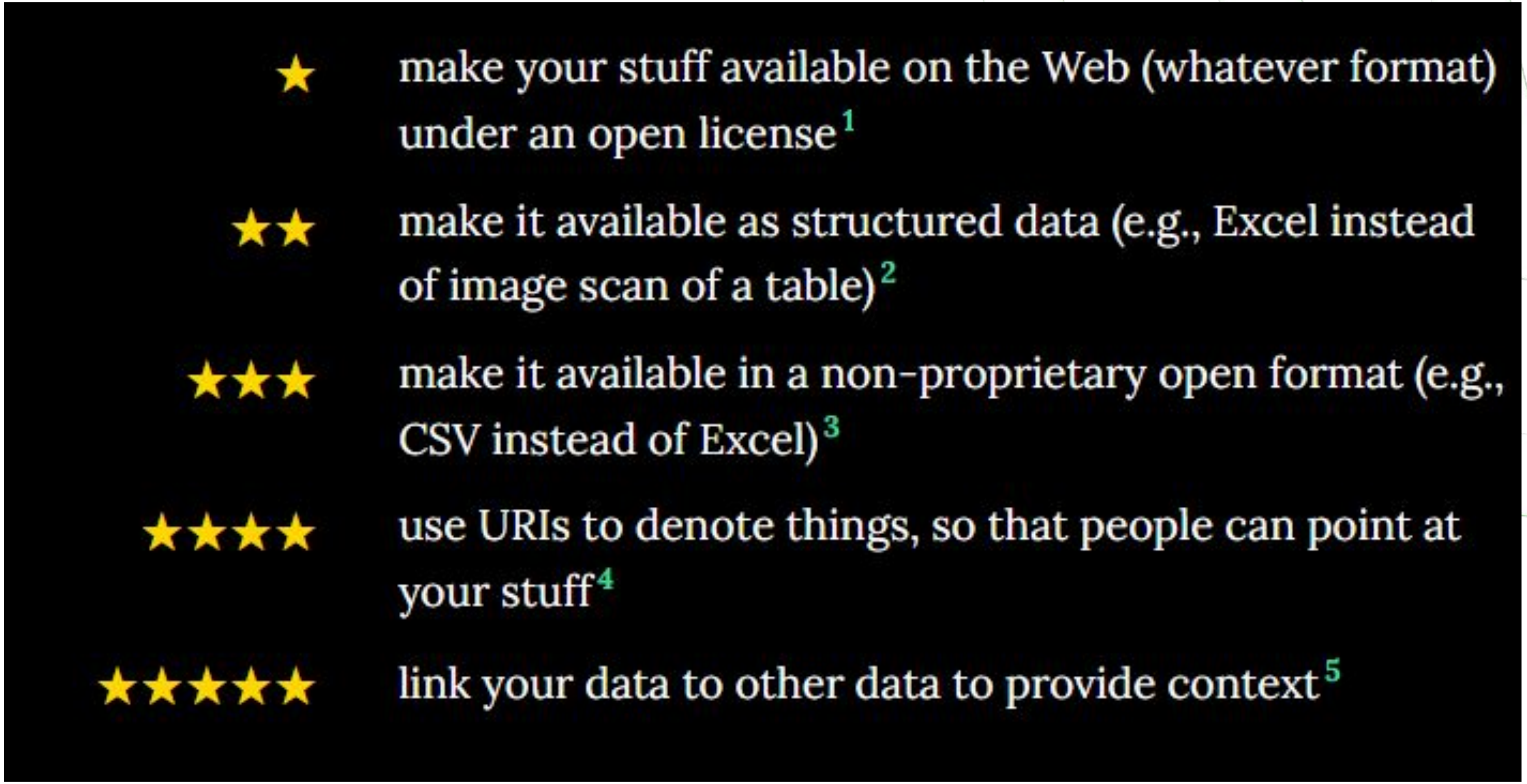
5 star data

Tim Berners-Lee, the inventor of the Web and Linked Data initiator, suggested a 5-star deployment scheme for Open Data.



https://5stardata.info/en/

5 star data



★ make your stuff available on the Web (whatever format) under an open license[1]

★★ make it available as structured data (e.g., Excel instead of image scan of a table)[2]

★★★ make it available in a non-proprietary open format (e.g., CSV instead of Excel)[3]

★★★★ use URIs to denote things, so that people can point at your stuff[4]

★★★★★ link your data to other data to provide context[5]

https://www.youtube.com/watch?v=OM6XIICm_qo

What is Linked Data?

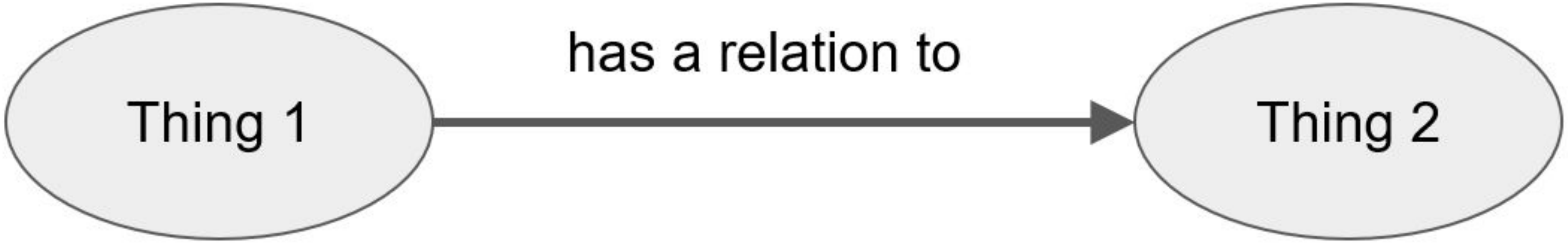
“Linked Data is a method of publishing structured data using standard Web technologies such as HTTP, RDF and URIs” - Tim Berners-Lee

Linked Data principles:

- Use URLs to name (identify) things
- Provide useful information about a thing when it's looked up
- Refer to other things (using their URL) when publishing data on the Web

What is RDF?

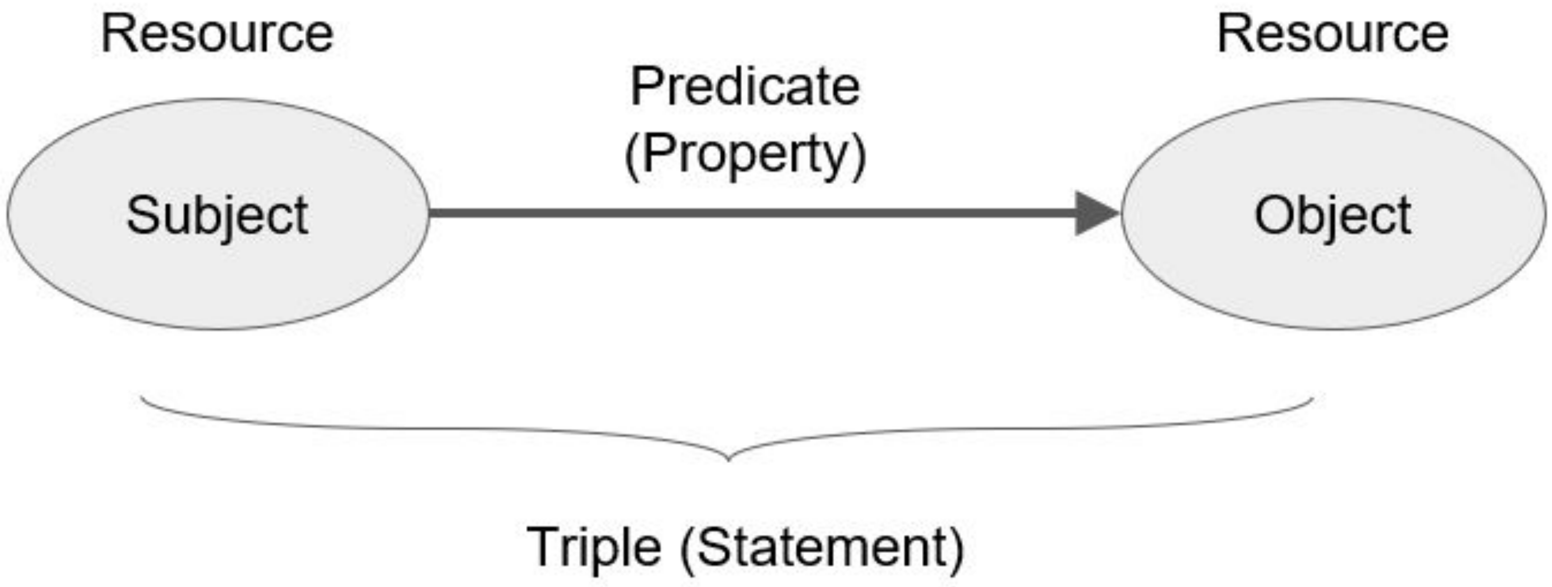
"The Resource Description Framework (RDF) is a standard model for data interchange on the Web"



https://www.w3.org/TR/rdf11-concepts/

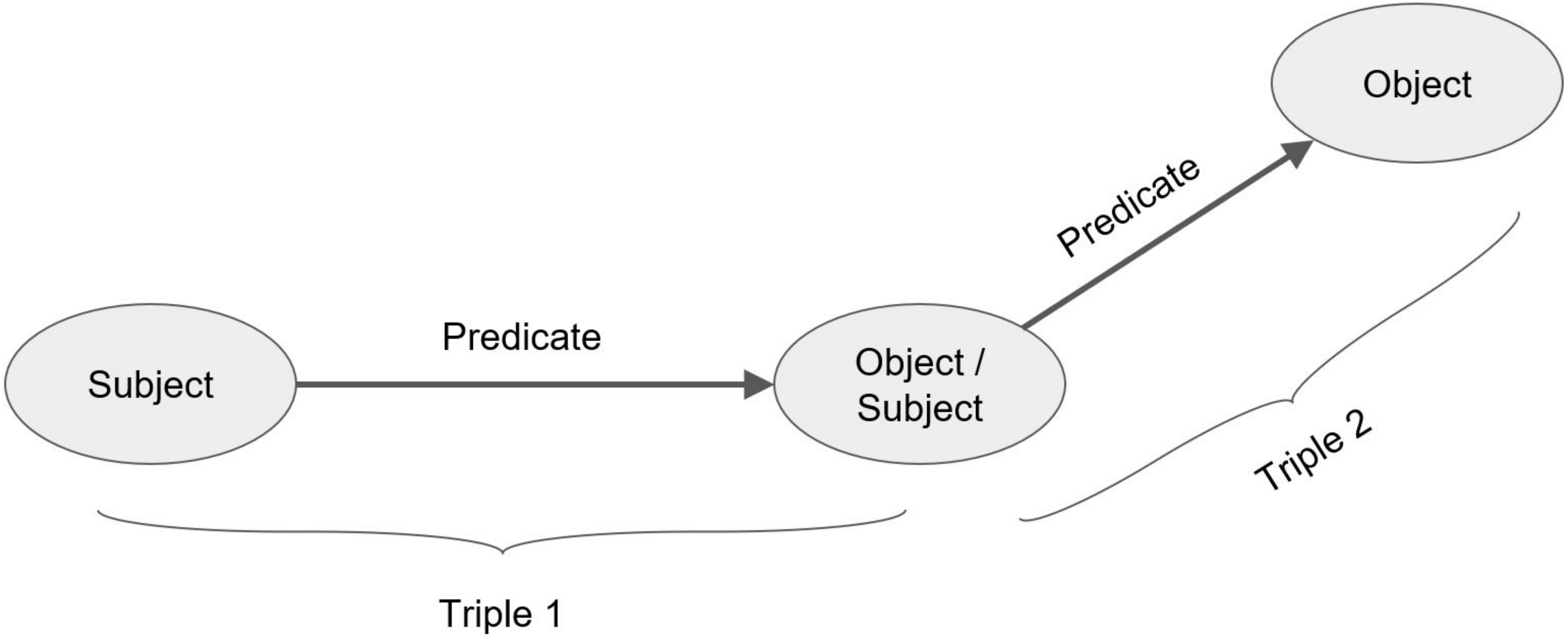
Triples

In RDF all data is modeled as a triple



RDF graphs

Multiple triples form a graph

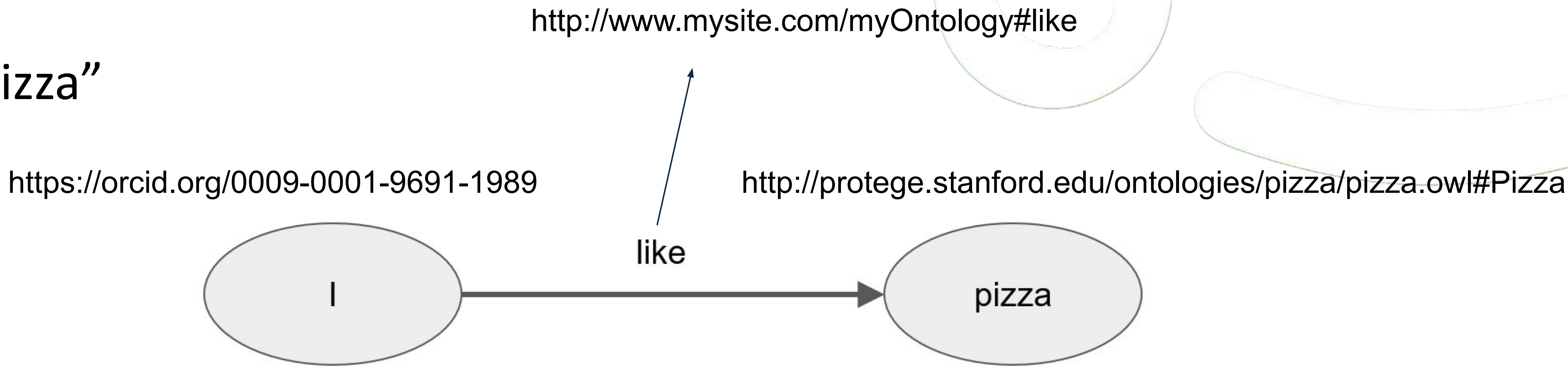


How do I publish my data as Linked Data?

1. Use URLs to name (identify) things
2. Provide useful information about a thing when it's looked up
3. Refer to other things (using their URL) when publishing data on the Web

e.g.

“I like pizza”



http://www.mysite.com/myOntology#like

https://orcid.org/0009-0001-9691-1989

http://protege.stanford.edu/ontologies/pizza/pizza.owl#Pizza

How do I publish my data as Linked Data?

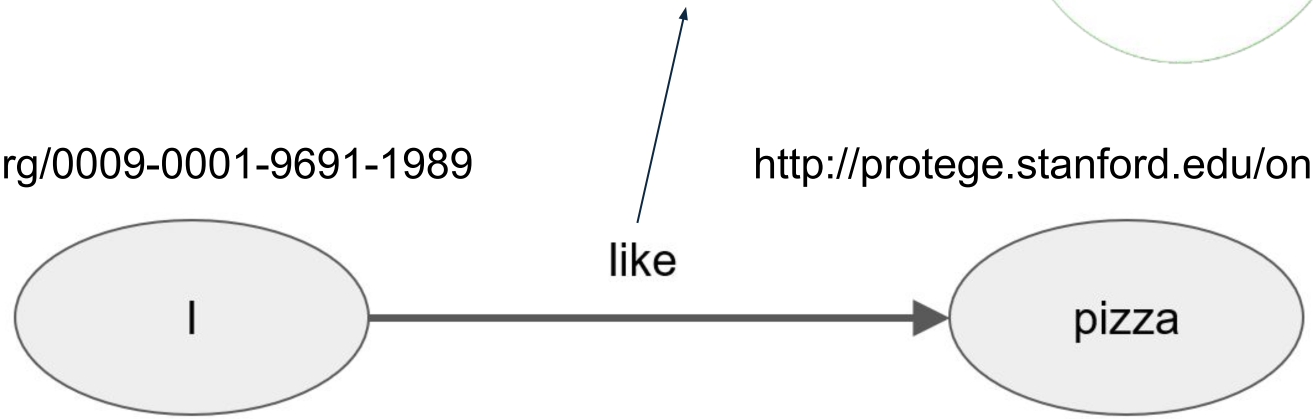
1. Use URLs to name (identify) things
2. Provide useful information about a thing when it's looked up
3. Refer to other things (using their URL) when publishing data on the Web

What do these URLs resolve to?

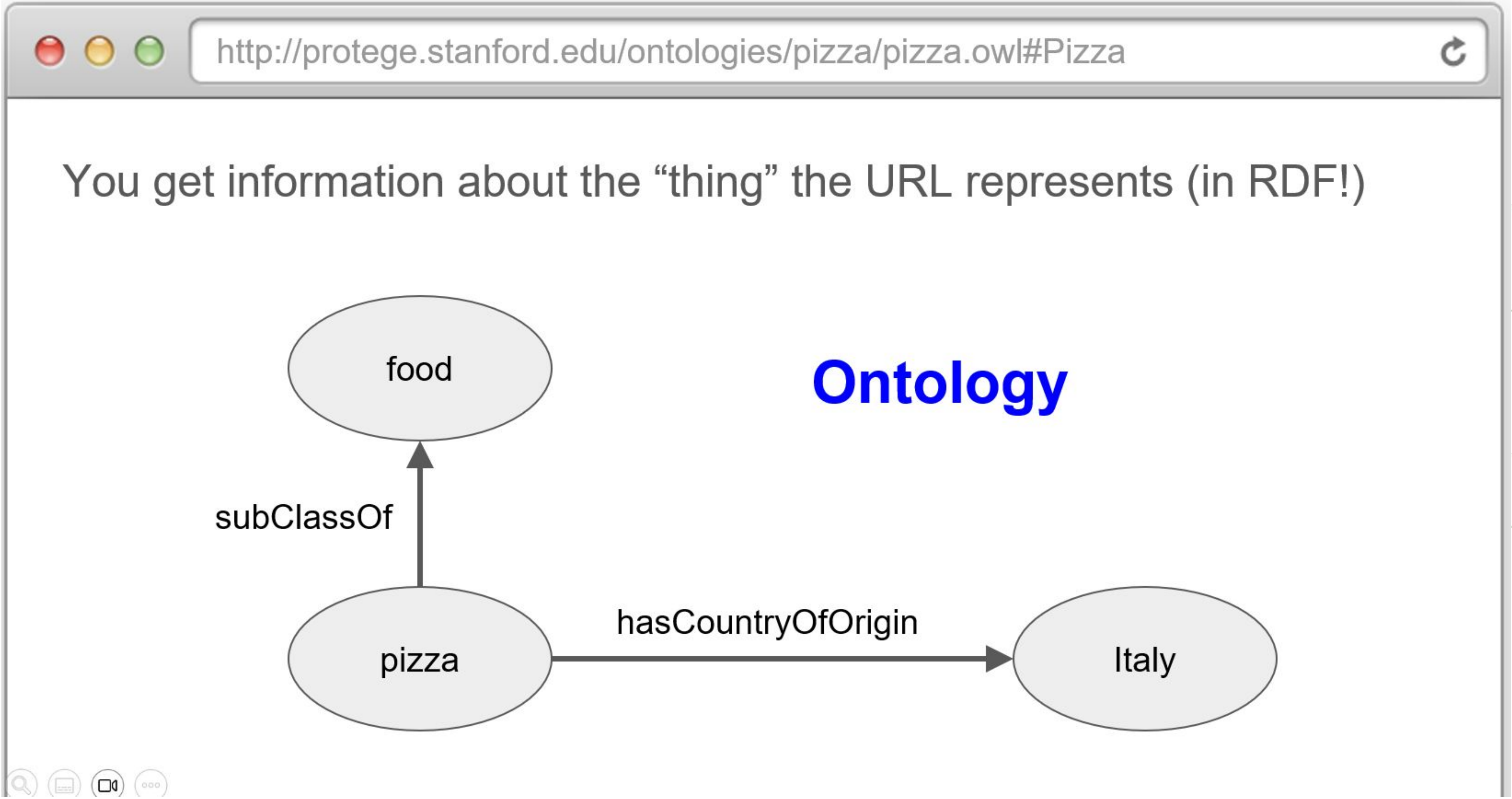
http://www.mysite.com/myOntology#like

https://orcid.org/0009-0001-9691-1989

http://protege.stanford.edu/ontologies/pizza/pizza.owl#Pizza



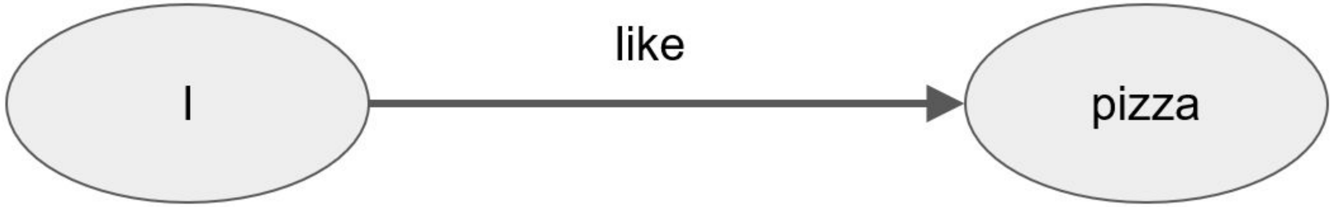
Provide useful information about a thing when it's looked up


http://protege.stanford.edu/ontologies/pizza/pizza.owl#Pizza

You get information about the “thing” the URL represents (in RDF!)

food
subClassOf
pizza
hasCountryOfOrigin
Italy
Ontology

How to represent RDF?

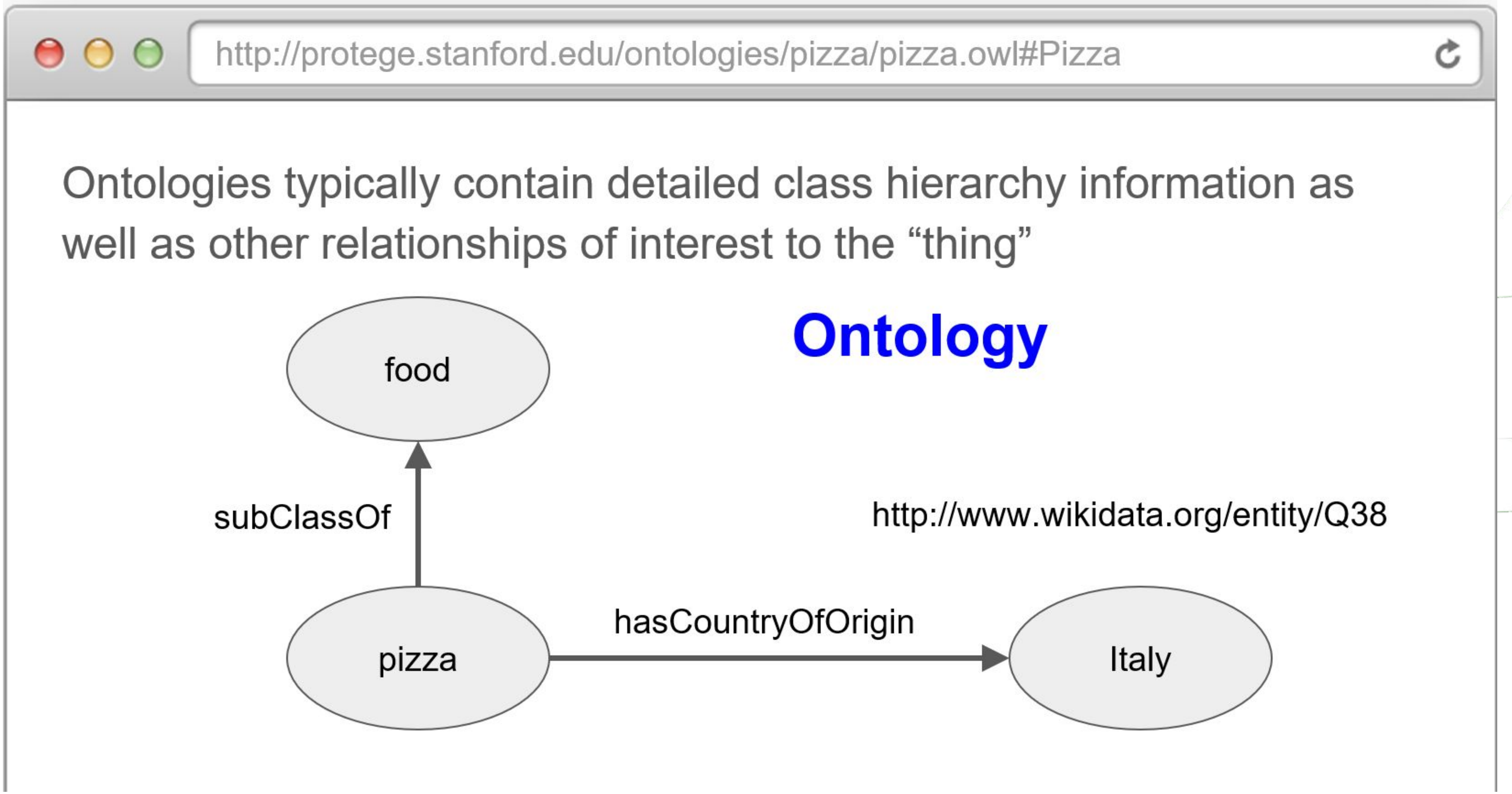


```
@prefix orcid: <http://orcid.org/> .
@prefix mo: <http://www.mysite.com/myOntology/> .
@prefix po: <http://protege.stanford.edu/ontologies/pizza/pizza.owl#> .

orcid:0000-0002-7633-1442  mo:like  po:Pizza  .
```

Turtle

https://www.w3.org/TR/turtle/

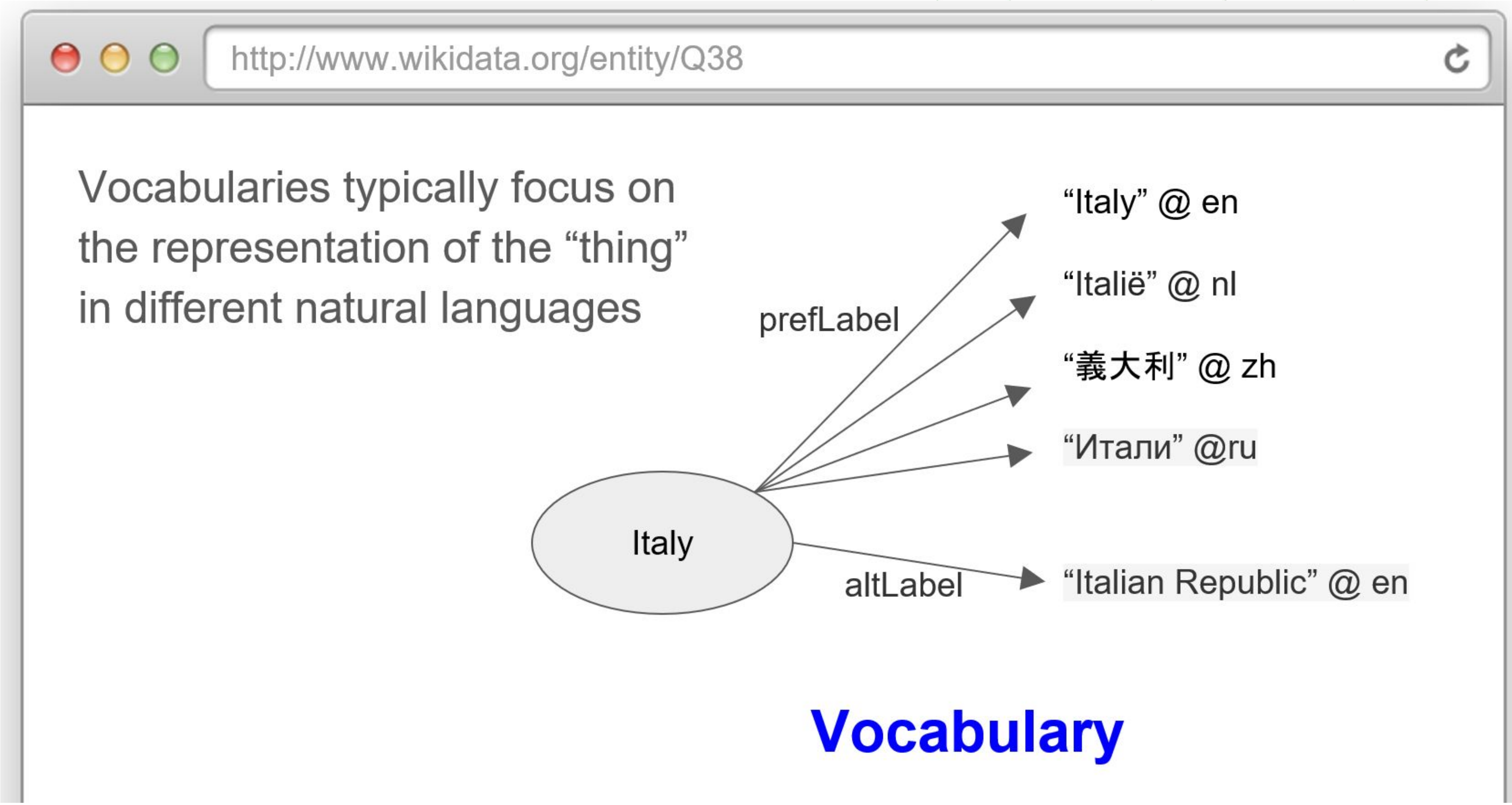
Provide useful information about a thing when it's looked up



http://protege.stanford.edu/ontologies/pizza/pizza.owl#Pizza

Ontologies typically contain detailed class hierarchy information as well as other relationships of interest to the “thing”

Provide useful information about a thing when it's looked up

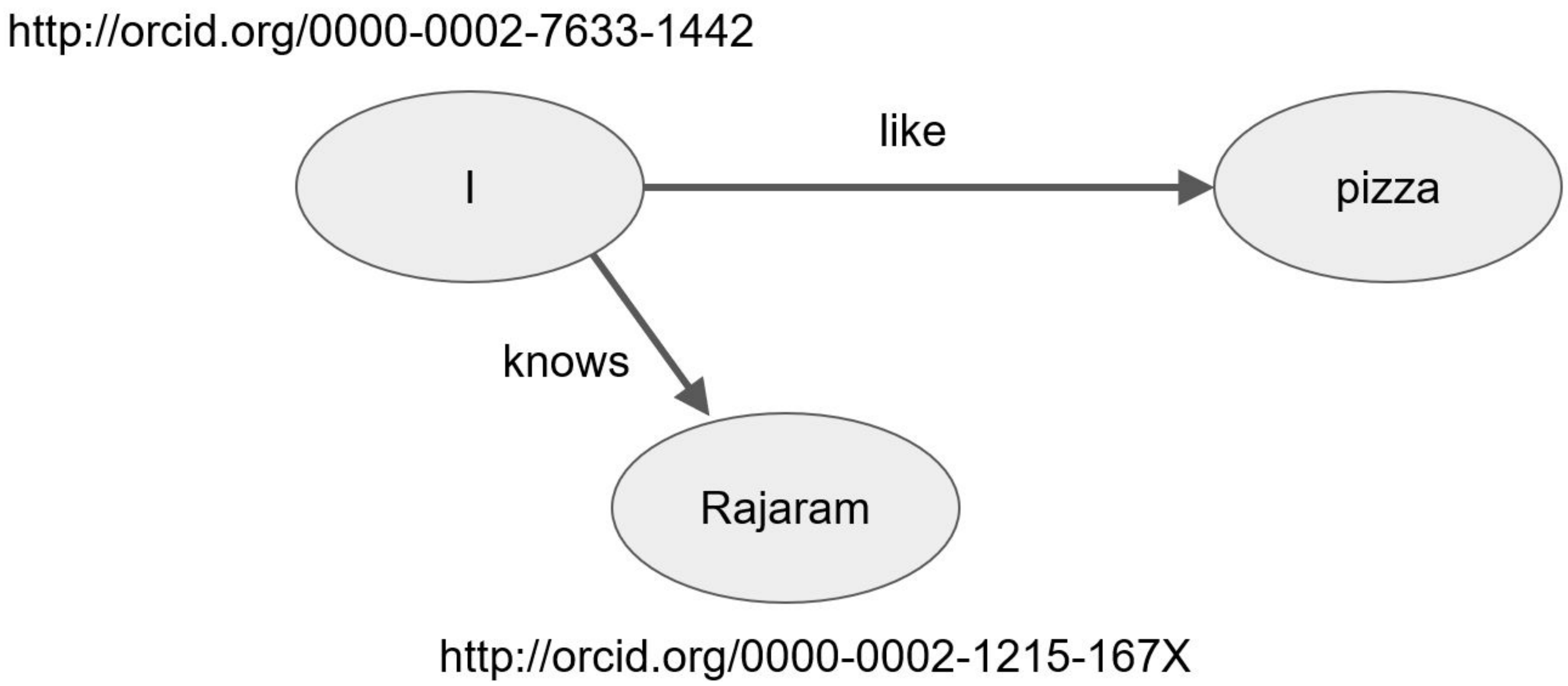


Vocabularies typically focus on the representation of the "thing" in different natural languages

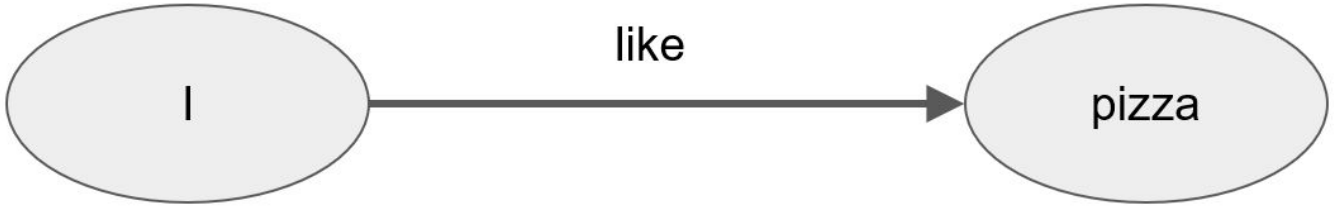
How do I publish my data as Linked Data?

1. Use URLs to name (identify) things
2. Provide useful information about a thing when it's looked up
3. Refer to other things (using their URL) when publishing data on the Web

For example: who do I know?


http://orcid.org/0000-0002-7633-1442
I
like
pizza
knows
Rajaram
http://orcid.org/0000-0002-1215-167X

How to represent RDF?

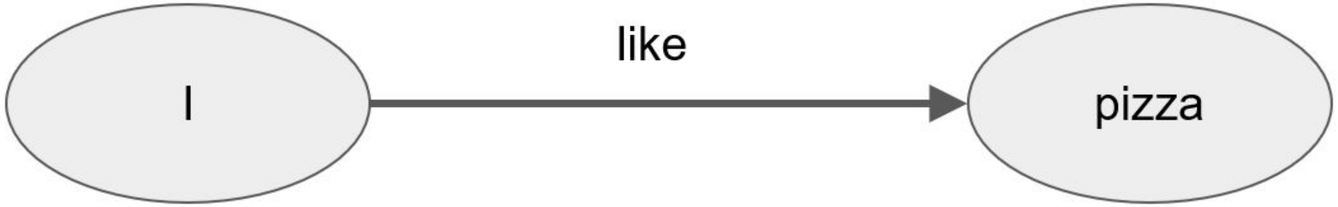


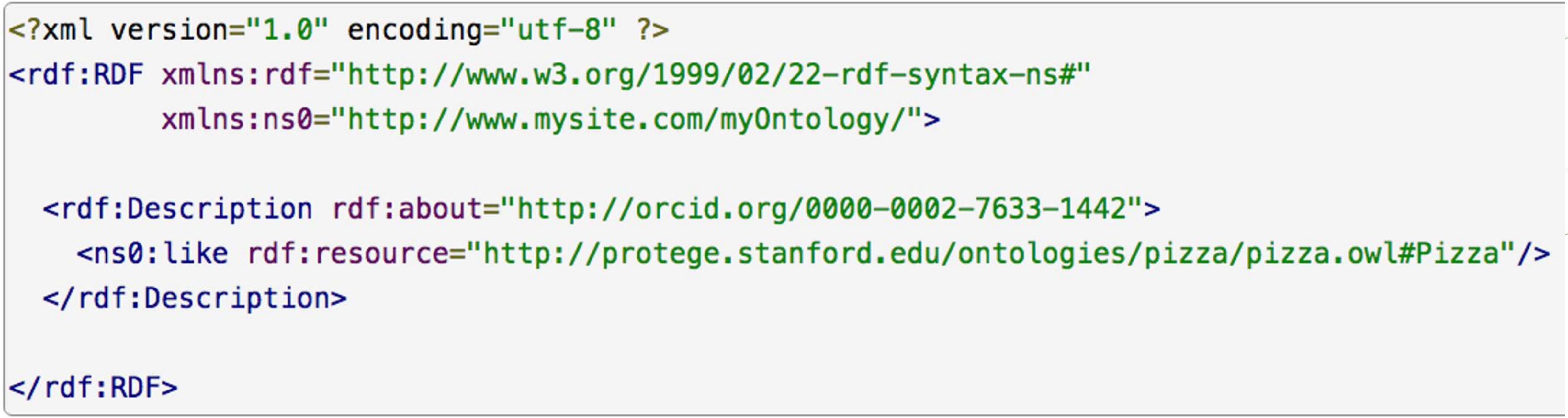
```
<http://orcid.org/0000-0002-7633-1442>
<http://www.mysite.com/myOntology#like>
<http://protege.stanford.edu/ontologies/pizza/pizza.owl#Pizza>.
```

n-triples

https://www.w3.org/TR/n-triples/

How to represent RDF?





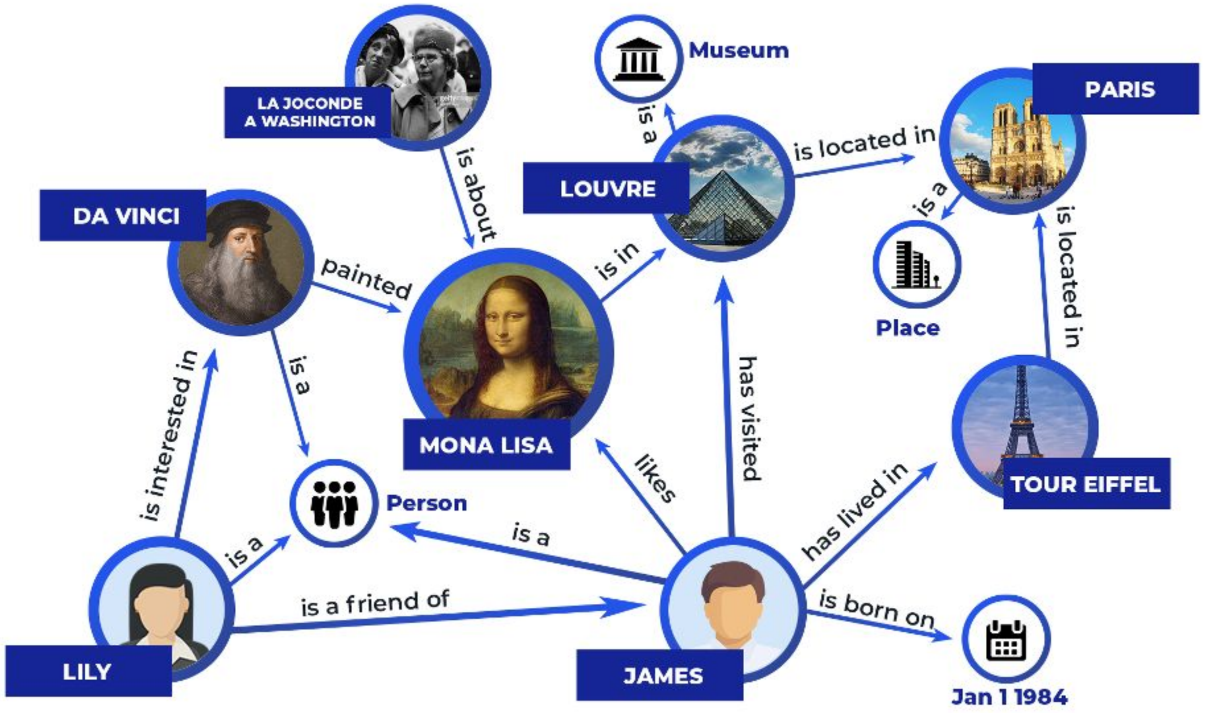
```
<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:ns0="http://www.mysite.com/myOntology/">

  <rdf:Description rdf:about="http://orcid.org/0000-0002-7633-1442">
    <ns0:like rdf:resource="http://protege.stanford.edu/ontologies/pizza/pizza.owl#Pizza"/>
  </rdf:Description>

</rdf:RDF>
```

RDF/XML

https://www.w3.org/TR/rdf-syntax-grammar/

Knowledge graph

A knowledge graph, also known as a semantic network, represents a network of real-world entities—such as objects, events, situations or concepts—and illustrates the relationship between them. This information is usually stored in a graph database and visualized as a graph structure, prompting the term knowledge "graph."



https://zilliz.com/learn/what-is-knowledge-graph

https://www.ibm.com/think/topics/knowledge-graph

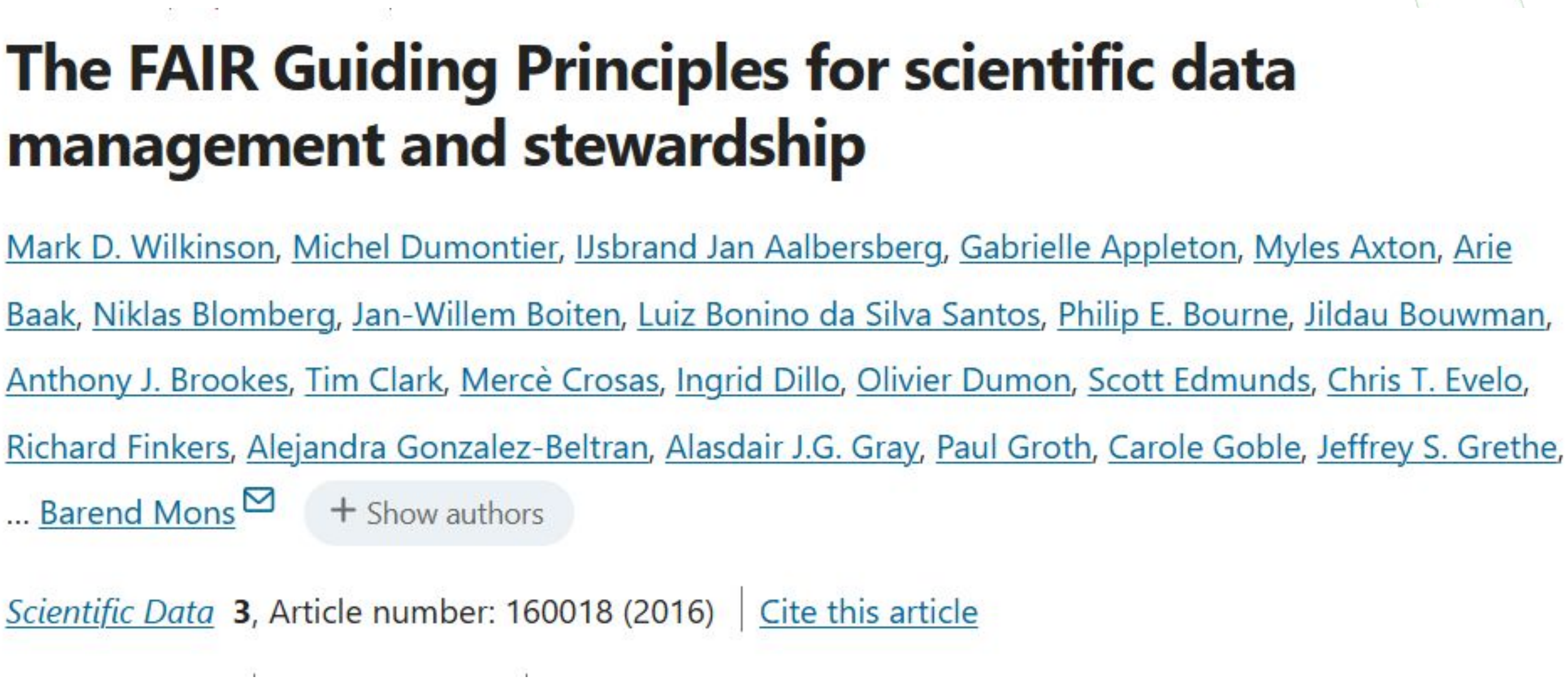
RDF triple stores

RDF graph objects can be persisted in specialized databases, RDF graph databases also known as RDF triple stores. Some examples are:

- Allegrograph
- Blazegraph
- GraphDB
- Stardog
- Virtuoso

FAIR Principles

ITINERIS



The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, ... Barend Mons ✉ + Show authors

Scientific Data **3**, Article number: 160018 (2016) | Cite this article

WorldFAIR

ENVRI FAIR

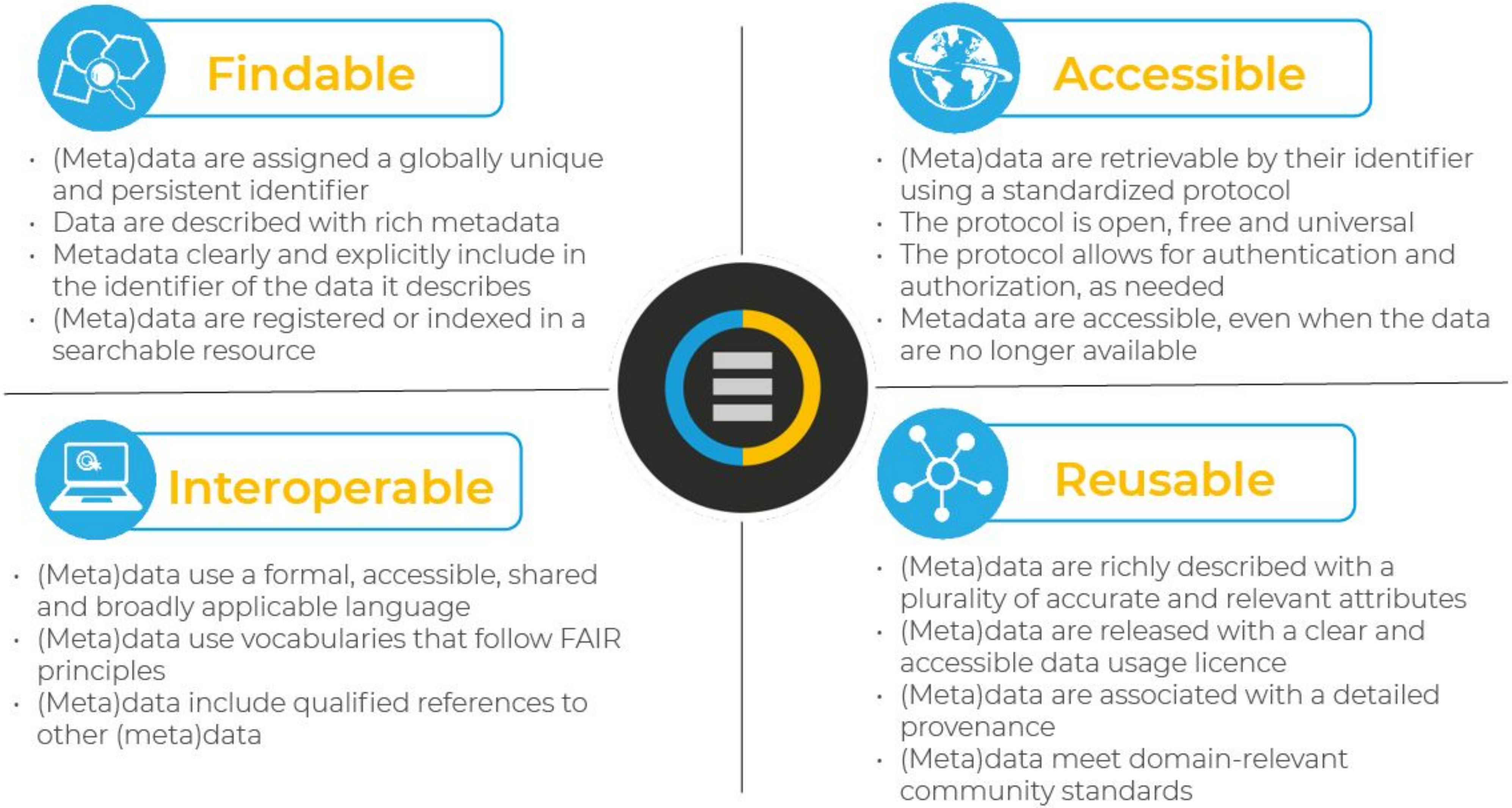
FAIR-IMPACT
Expanding FAIR Solutions across EOSC

FAIR-EASE

FAIRSFAIR
Fostering Fair Data Practices in Europe

FAIRCORE4EOSC

https://doi.org/10.1038/sdata.2016.18



Findable

- (Meta)data are assigned a globally unique and persistent identifier
- Data are described with rich metadata
- Metadata clearly and explicitly include in the identifier of the data it describes
- (Meta)data are registered or indexed in a searchable resource

Accessible

- (Meta)data are retrievable by their identifier using a standardized protocol
- The protocol is open, free and universal
- The protocol allows for authentication and authorization, as needed
- Metadata are accessible, even when the data are no longer available

Interoperable

- (Meta)data use a formal, accessible, shared and broadly applicable language
- (Meta)data use vocabularies that follow FAIR principles
- (Meta)data include qualified references to other (meta)data

Reusable

- (Meta)data are richly described with a plurality of accurate and relevant attributes
- (Meta)data are released with a clear and accessible data usage licence
- (Meta)data are associated with a detailed provenance
- (Meta)data meet domain-relevant community standards

https://www.ccdc.cam.ac.uk/solutions/about-the-csd/fair-data-principles/

Knowledge organisation systems (KOS)

- The term knowledge organization systems (KOS) is intended to encompass all types of schemes for organising information and promoting knowledge management.
- KOS are used to organise materials for the purpose of retrieval and to manage a collection. A KOS serves as a bridge between the user's information need and the material in the collection.
- Examples: authority files, gazetteers, taxonomies, thesauri, ontologies.

https://www.clir.org/pubs/reports/pub91/1knowledge/

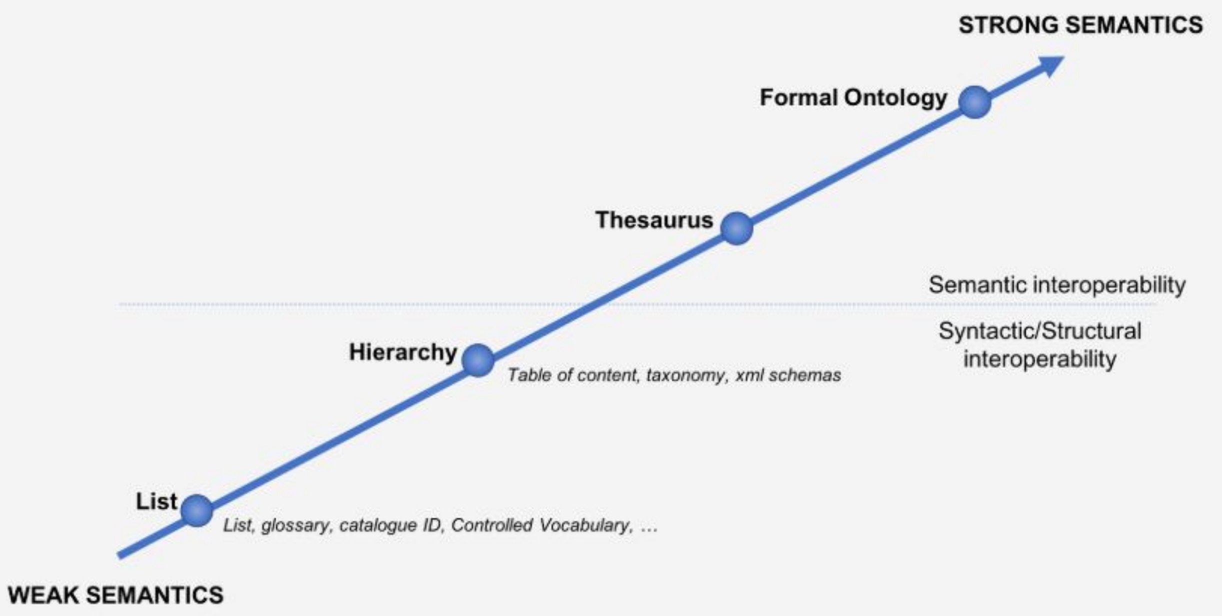
Some definitions

Thesaurus: “controlled and structured vocabulary in which concepts are represented by terms, organised so that relationships between concepts are made explicit, and preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms”.

Ontology: “A formal model that allows knowledge to be represented for a specific domain. An ontology describes the types of things that exist (classes), the relationships between them (properties) and the logical ways those classes and properties can be used together (axioms).”

From KOS to semantic artefacts

A semantic artefact as a machine-actionable formalisation (represented using appropriate formats and serialisations, including RDF and non-RDF standards) of a conceptualisation, enabling sharing and reuse by humans and machines.


STRONG SEMANTICS
Formal Ontology
Thesaurus
Semantic interoperability
Syntactic/Structural interoperability
Hierarchy
Table of content, taxonomy, xml schemas
List
List, glossary, catalogue ID, Controlled Vocabulary, ...
WEAK SEMANTICS

Semantic artefact spectrum. Derived from Leo Obrst, 2010

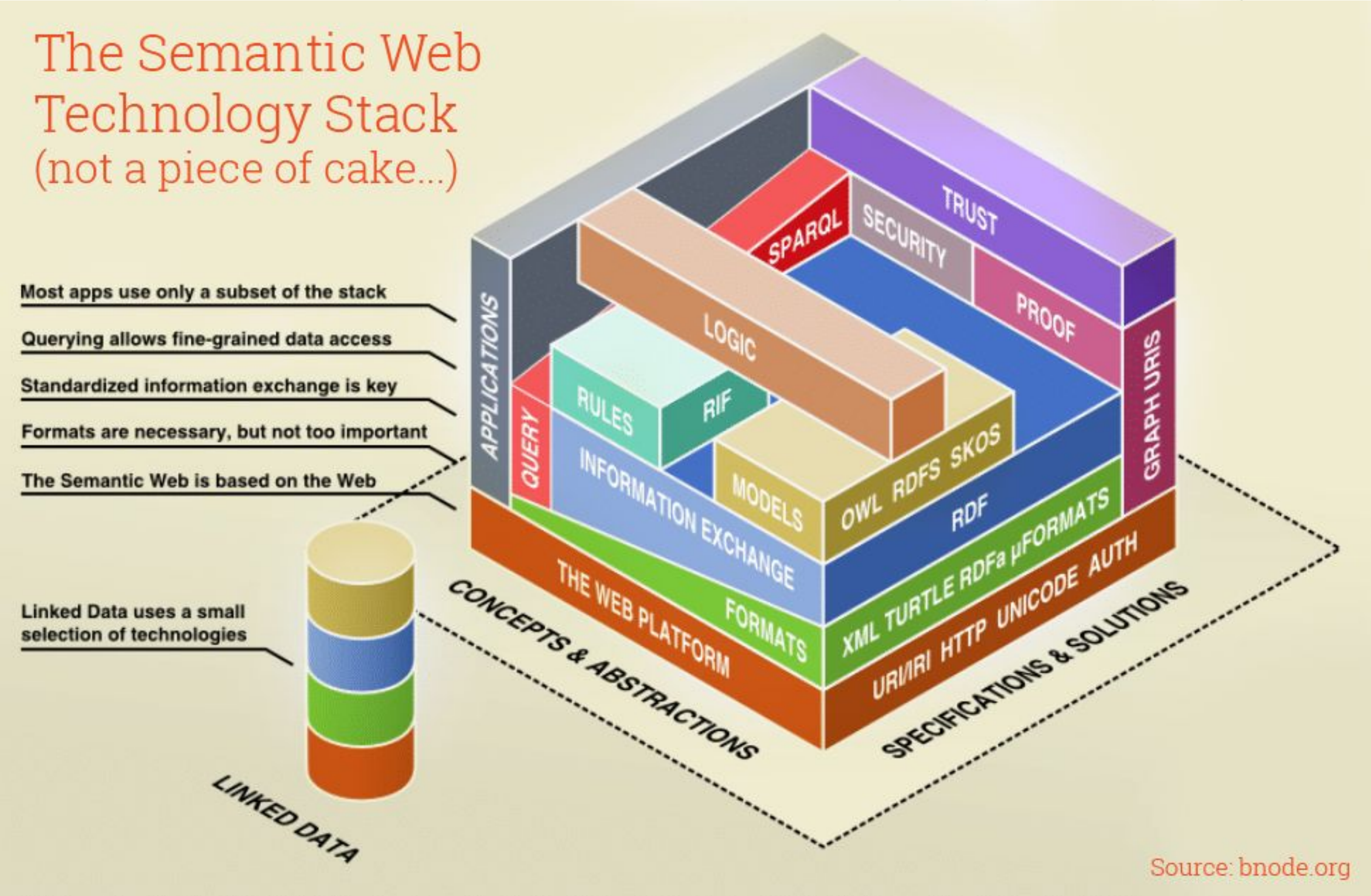
https://doi.org/10.1038/s41597-024-03185-4

Semantic Web technologies

vocabulary	scope	prefix	namespace URI
RDF	Basic RDF elements	`rdf:`	http://www.w3.org/1999/02/22-rdf-syntax-ns#
RDF Schema	RDF Schema elements	`rdfs:`	http://www.w3.org/2000/01/rdf-schema#
Web Ontology Language (OWL)	OWL elements	`owl:`	http://www.w3.org/2002/07/owl#
SKOS	SKOS elements	`skos:`	http://www.w3.org/2004/02/skos/core#
SHACL	SHACL elements	`sh:`	http://www.w3.org/ns/shacl#

https://csiro-enviro-informatics.github.io/info-engineering/standards.html

Semantic Web technologies



The Semantic Web Technology Stack
(not a piece of cake...)

Most apps use only a subset of the stack

Querying allows fine-grained data access

Standardized information exchange is key

Formats are necessary, but not too important

The Semantic Web is based on the Web

Linked Data uses a small selection of technologies

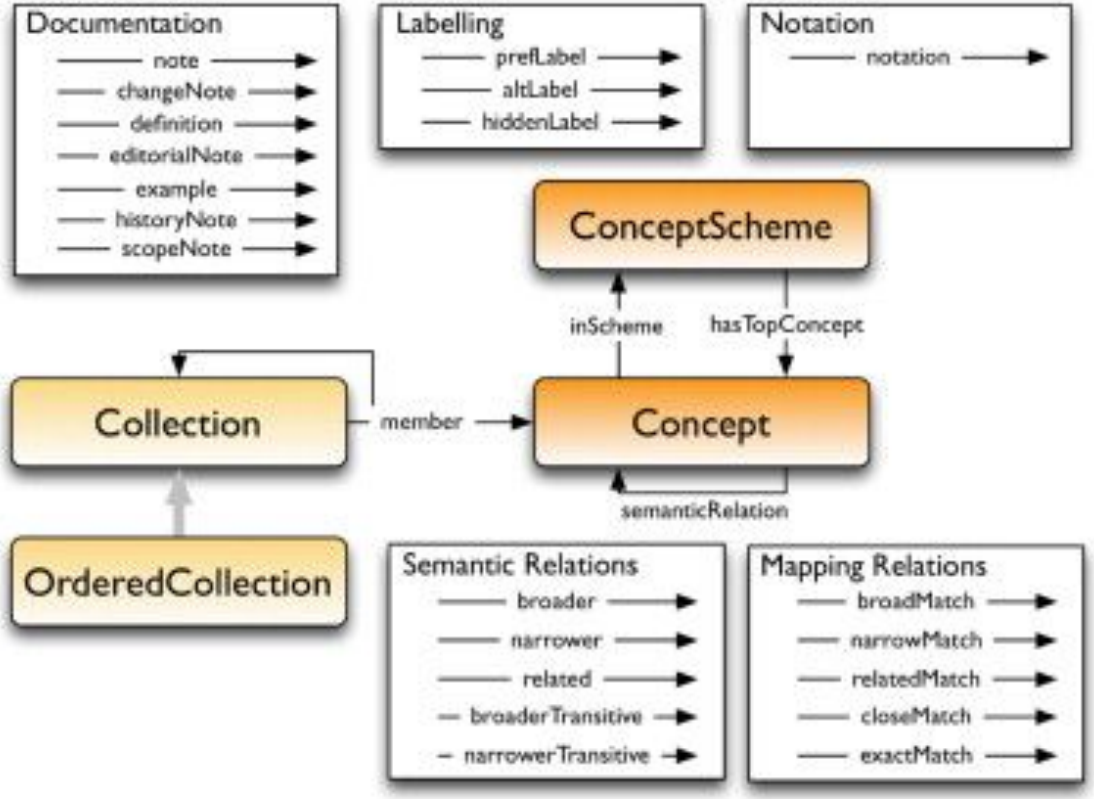
Source: bnode.org

https://csiro-enviro-informatics.github.io/info-engineering/standards.html

Simple Knowledge Organization System (SKOS)

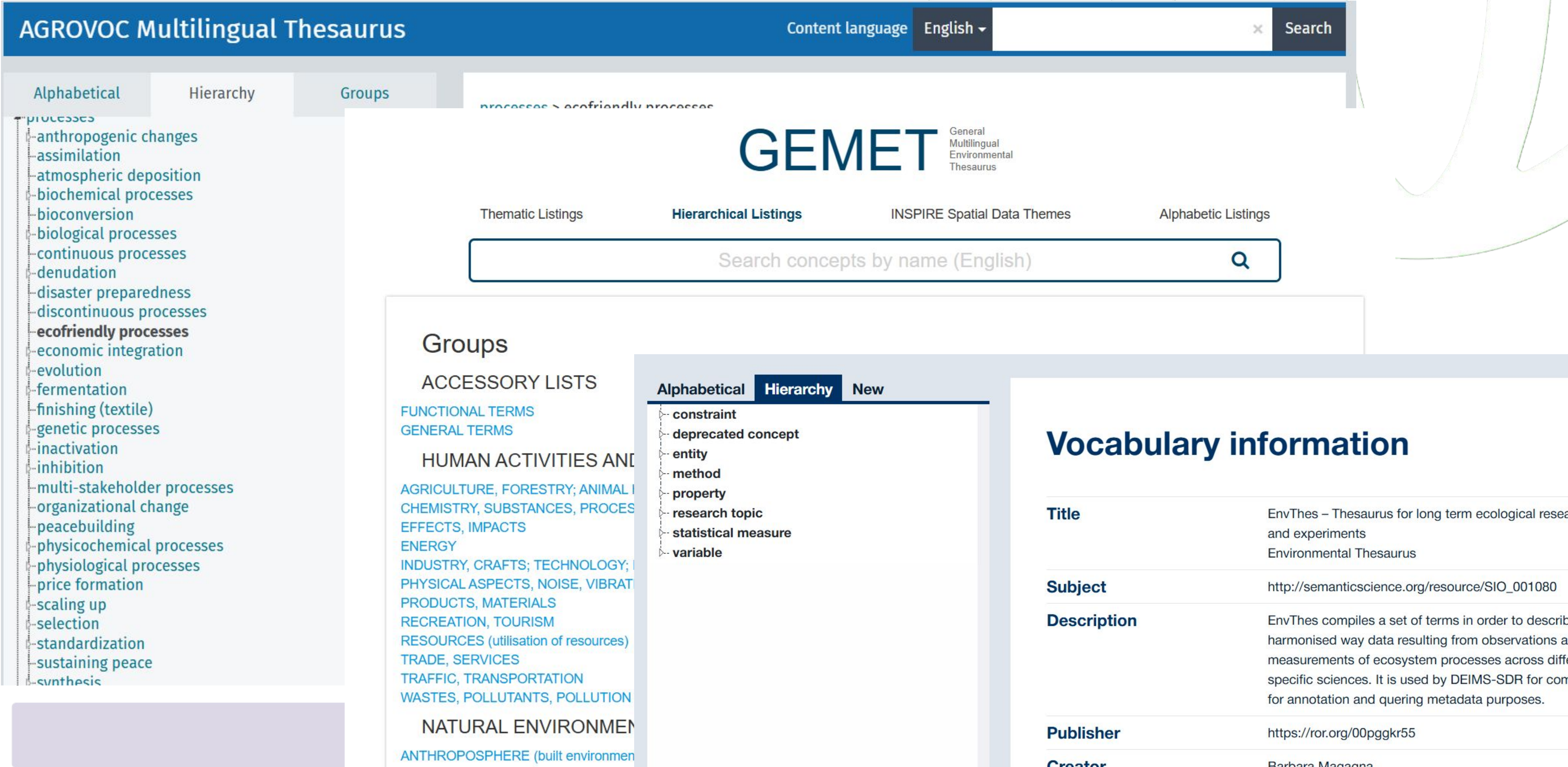
The Simple Knowledge Organization System is a common data model for KOS such as thesauri, classification schemes, subject heading systems and taxonomies.

Using SKOS, a knowledge organization system can be expressed as machine-readable data. It can then be exchanged between computer applications and published in a machine-readable format in the Web.



https://www.w3.org/TR/skos-reference/
https://doi.org/10.1016/j.websem.2013.05.001

SKOS thesauri in environmental sciences



processes > ecofriendly processes

processes
- anthropogenic changes
- assimilation
- atmospheric deposition
- biochemical processes
- bioconversion
- biological processes
- continuous processes
- denudation
- disaster preparedness
- discontinuous processes
- **ecofriendly processes**
- economic integration
- evolution
- fermentation
- finishing (textile)
- genetic processes
- inactivation
- inhibition
- multi-stakeholder processes
- organizational change
- peacebuilding
- physicochemical processes
- physiological processes
- price formation
- scaling up
- selection
- standardization
- sustaining peace
- synthesis

Groups

ACCESSORY LISTS

FUNCTIONAL TERMS
GENERAL TERMS

HUMAN ACTIVITIES AND

AGRICULTURE, FORESTRY; ANIMAL
CHEMISTRY, SUBSTANCES, PROCES
EFFECTS, IMPACTS
ENERGY
INDUSTRY, CRAFTS; TECHNOLOGY;
PHYSICAL ASPECTS, NOISE, VIBRAT
PRODUCTS, MATERIALS
RECREATION, TOURISM
RESOURCES (utilisation of resources)
TRADE, SERVICES
TRAFFIC, TRANSPORTATION
WASTES, POLLUTANTS, POLLUTION

NATURAL ENVIRONMEN

ANTHROPOSPHERE (built environmen

- constraint
- deprecated concept
- entity
- method
- property
- research topic
- statistical measure
- variable

Vocabulary information

Title	EnvThes – Thesaurus for long term ecological resea and experiments Environmental Thesaurus
Subject	http://semanticscience.org/resource/SIO_001080
Description	EnvThes compiles a set of terms in order to describ harmonised way data resulting from observations a measurements of ecosystem processes across diff specific sciences. It is used by DEIMS-SDR for com for annotation and quering metadata purposes.
Publisher	https://ror.org/00pggkr55
Creator	Barbara Magagna

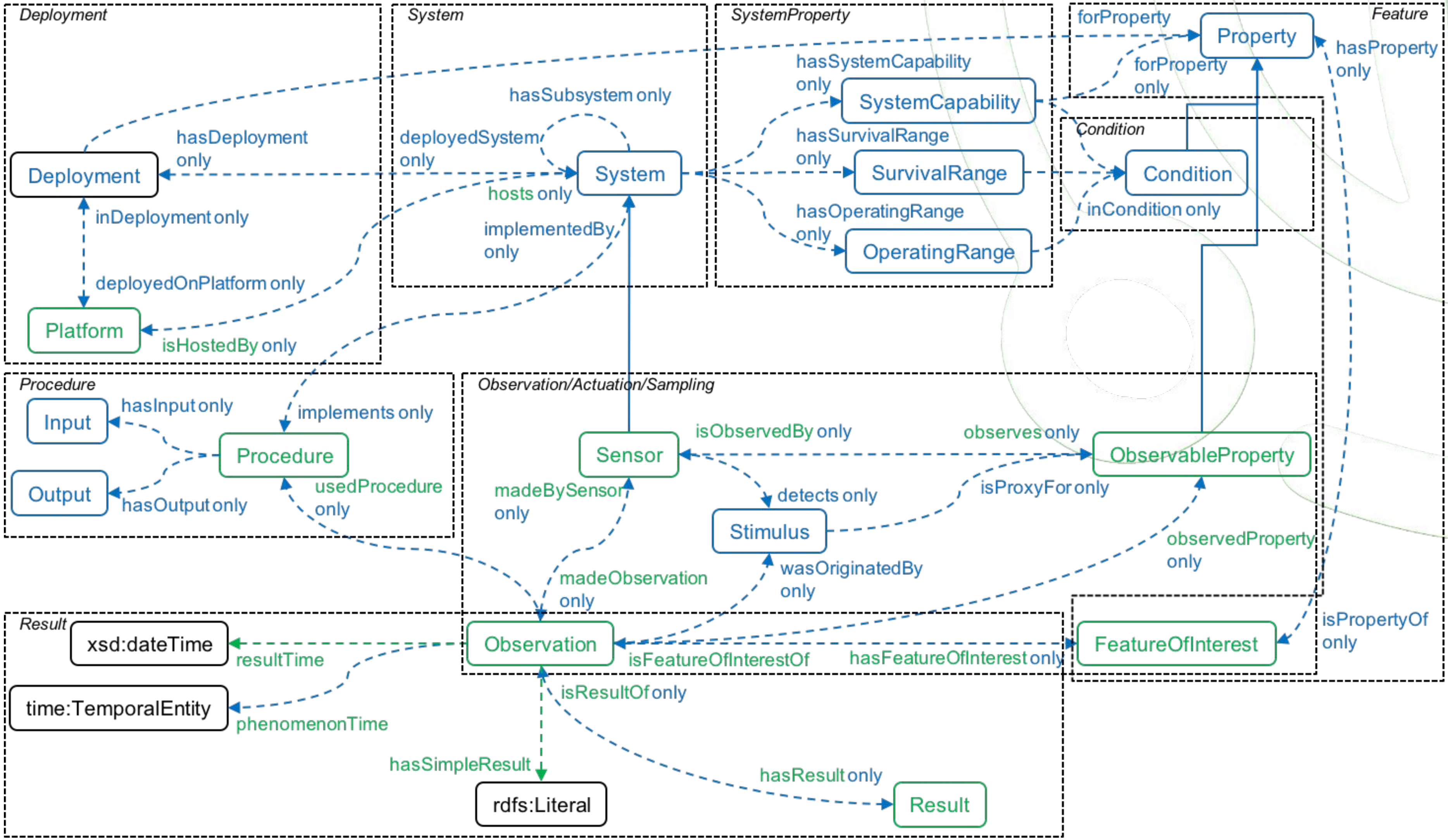
Web Ontology Language (OWL)

The W3C Web Ontology Language (OWL) is a Semantic Web language designed to represent rich and complex knowledge about things, groups of things, and relations between things.

Main elements are: classes, data properties, object properties, individuals

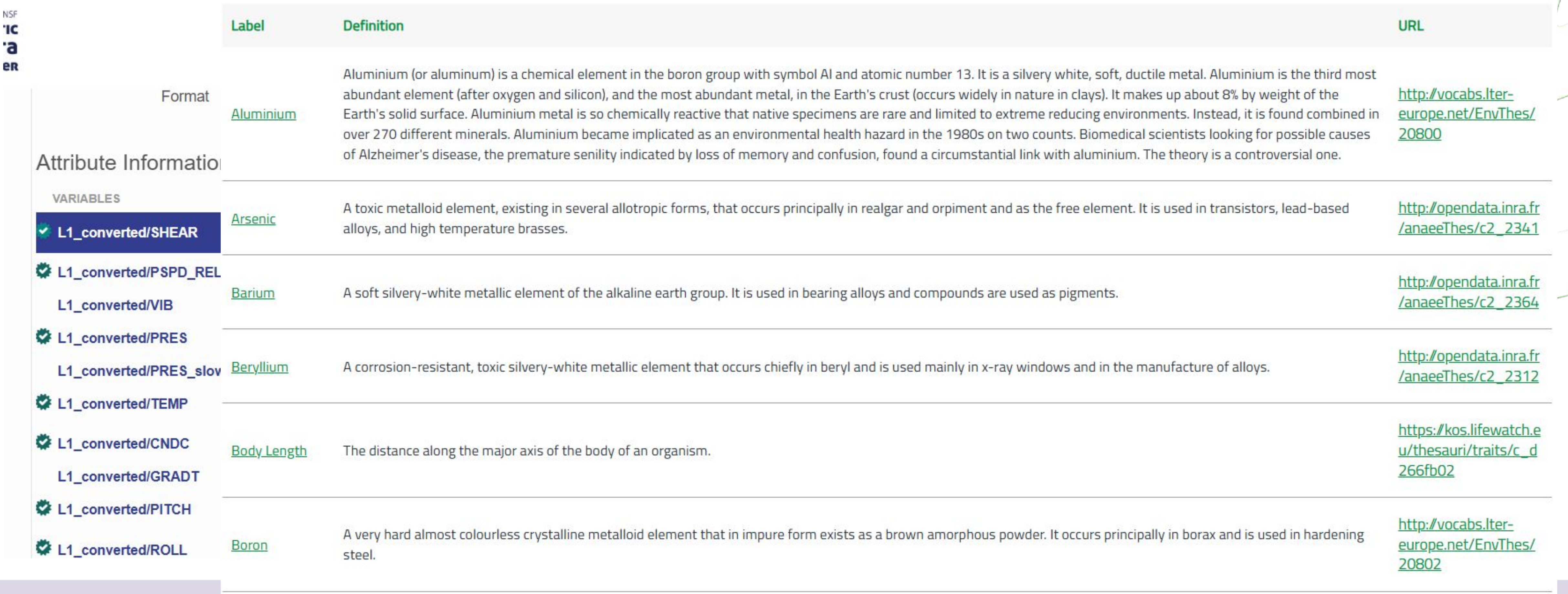
https://www.w3.org/TR/2012/REC-owl2-syntax-20121211/

OWL ontologies in environmental sciences



What are their applications?

Main applications of thesauri and ontologies are in the data management systems, but also in data integration systems.



NSF
ic
a
er

Format

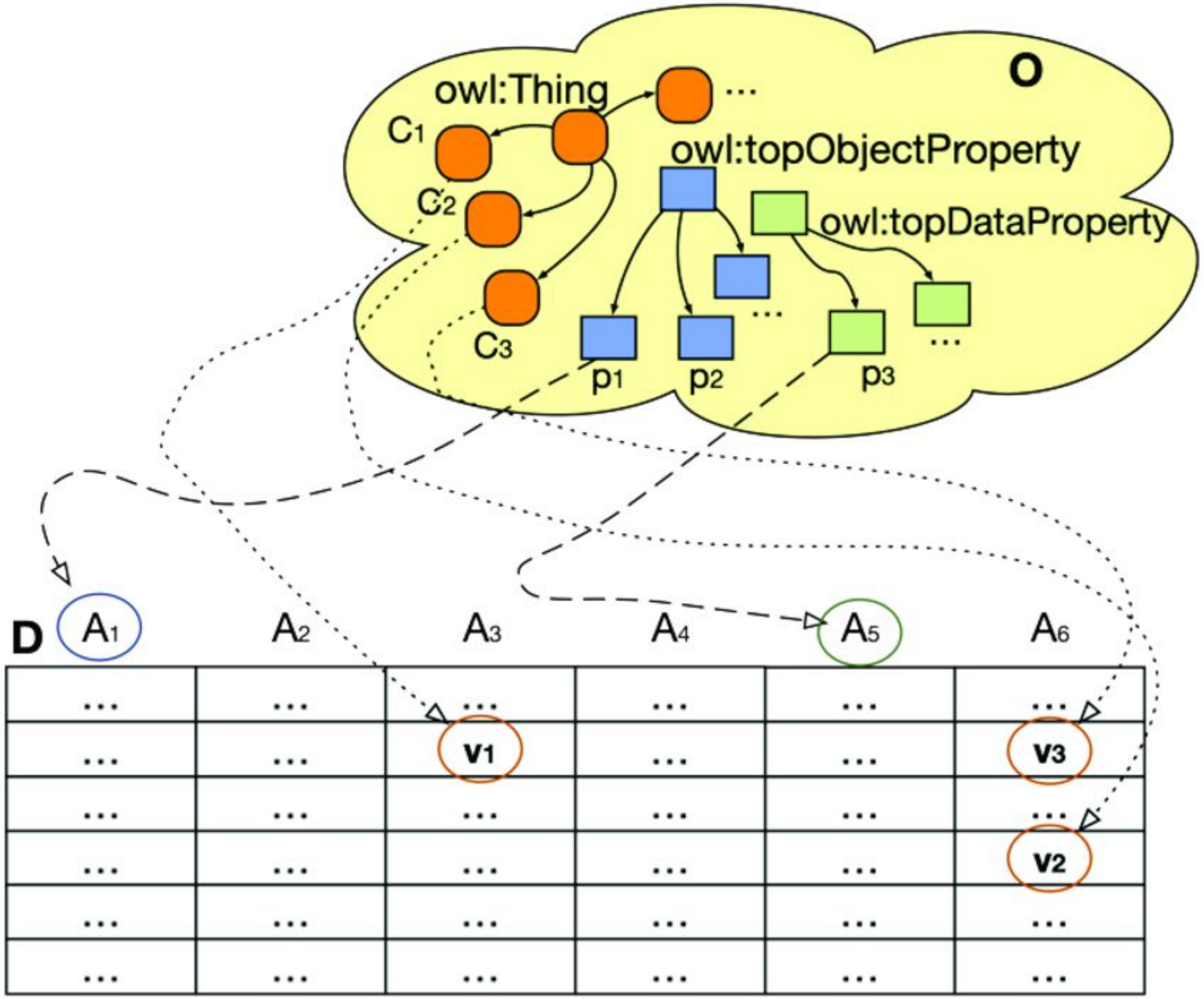
Attribute Informatio

VARIABLES

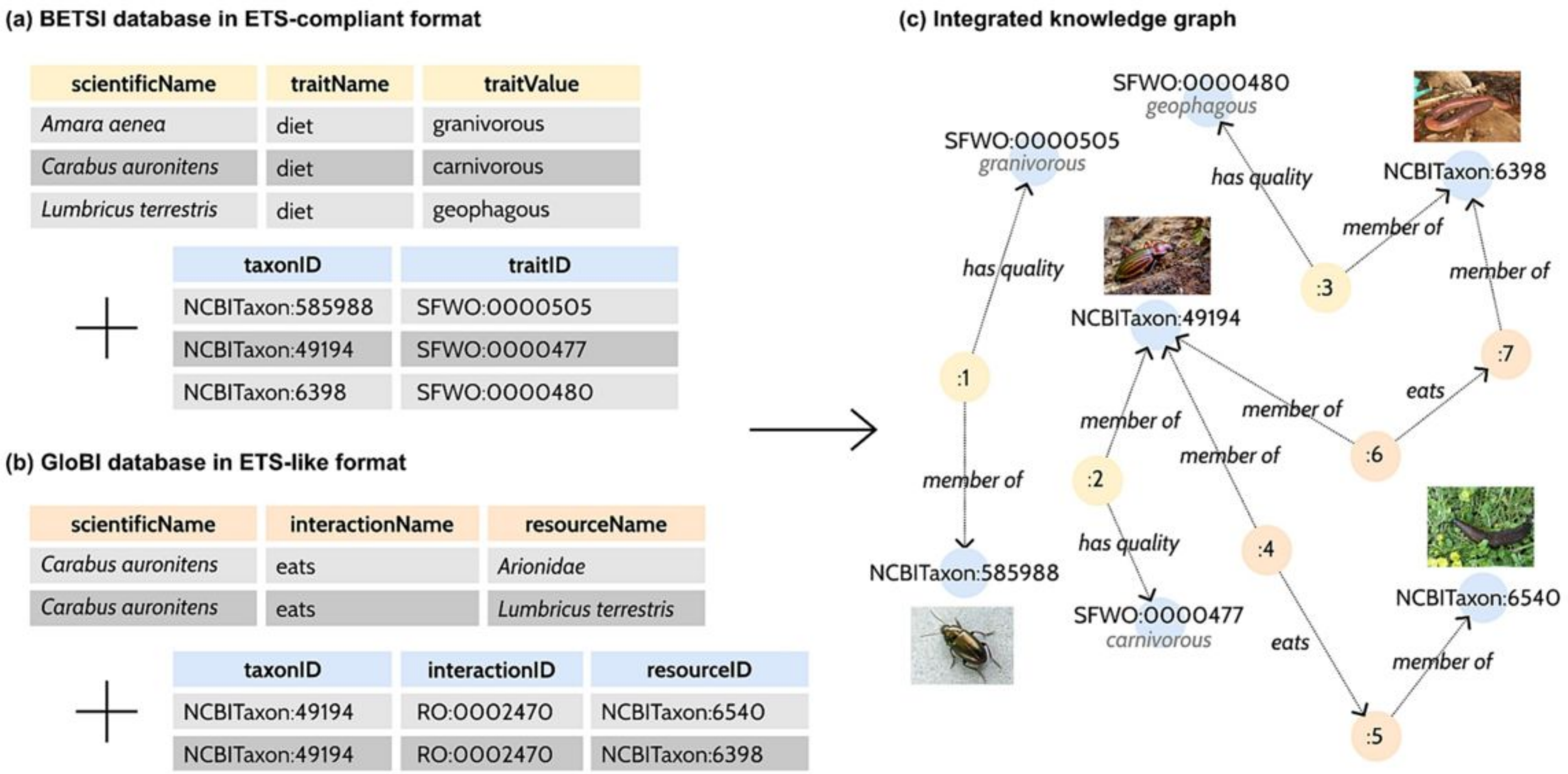
- ✓ L1_converted/SHEAR
- ✓ L1_converted/PSPD_REL
- L1_converted/VIB
- ✓ L1_converted/PRES
- L1_converted/PRES_slov
- ✓ L1_converted/TEMP
- ✓ L1_converted/CNDC
- L1_converted/GRADT
- ✓ L1_converted/PITCH
- ✓ L1_converted/ROLL

Label	Definition	URL
Aluminium	Aluminium (or aluminum) is a chemical element in the boron group with symbol Al and atomic number 13. It is a silvery white, soft, ductile metal. Aluminium is the third most abundant element (after oxygen and silicon), and the most abundant metal, in the Earth's crust (occurs widely in nature in clays). It makes up about 8% by weight of the Earth's solid surface. Aluminium metal is so chemically reactive that native specimens are rare and limited to extreme reducing environments. Instead, it is found combined in over 270 different minerals. Aluminium became implicated as an environmental health hazard in the 1980s on two counts. Biomedical scientists looking for possible causes of Alzheimer's disease, the premature senility indicated by loss of memory and confusion, found a circumstantial link with aluminium. The theory is a controversial one.	http://vocabs.lter-europe.net/EnvThes/20800
Arsenic	A toxic metalloid element, existing in several allotropic forms, that occurs principally in realgar and orpiment and as the free element. It is used in transistors, lead-based alloys, and high temperature brasses.	http://opendata.inra.fr/anaeeThes/c2_2341
Barium	A soft silvery-white metallic element of the alkaline earth group. It is used in bearing alloys and compounds are used as pigments.	http://opendata.inra.fr/anaeeThes/c2_2364
Beryllium	A corrosion-resistant, toxic silvery-white metallic element that occurs chiefly in beryl and is used mainly in x-ray windows and in the manufacture of alloys.	http://opendata.inra.fr/anaeeThes/c2_2312
Body Length	The distance along the major axis of the body of an organism.	https://kos.lifewatch.eu/thesauri/traits/c_d266fb02
Boron	A very hard almost colourless crystalline metalloid element that in impure form exists as a brown amorphous powder. It occurs principally in borax and is used in hardening steel.	http://vocabs.lter-europe.net/EnvThes/20802

Dataset mapping



A1	A2	A3	A4	A5	A6
...	...	...	...	...	...
...	...	v1	...	...	v3
...	...	...	...	...	...
...	...	...	...	...	v2
...	...	...	...	...	...
...	...	...	...	...	...



(a) BETSI database in ETS-compliant format

scientificName	traitName	traitValue
Amara aenea	diet	granivorous
Carabus auronitens	diet	carnivorous
Lumbricus terrestris	diet	geophagous

+

taxonID	traitID
NCBITaxon:585988	SFWO:0000505
NCBITaxon:49194	SFWO:0000477
NCBITaxon:6398	SFWO:0000480

(b) GloBI database in ETS-like format

scientificName	interactionName	resourceName
Carabus auronitens	eats	*Arionidae*
Carabus auronitens	eats	*Lumbricus terrestris*

+

taxonID	interactionID	resourceID
NCBITaxon:49194	RO:0002470	NCBITaxon:6540
NCBITaxon:49194	RO:0002470	NCBITaxon:6398

(c) Integrated knowledge graph

https://doi.org/10.1016/j.ecoinf.2023.102360

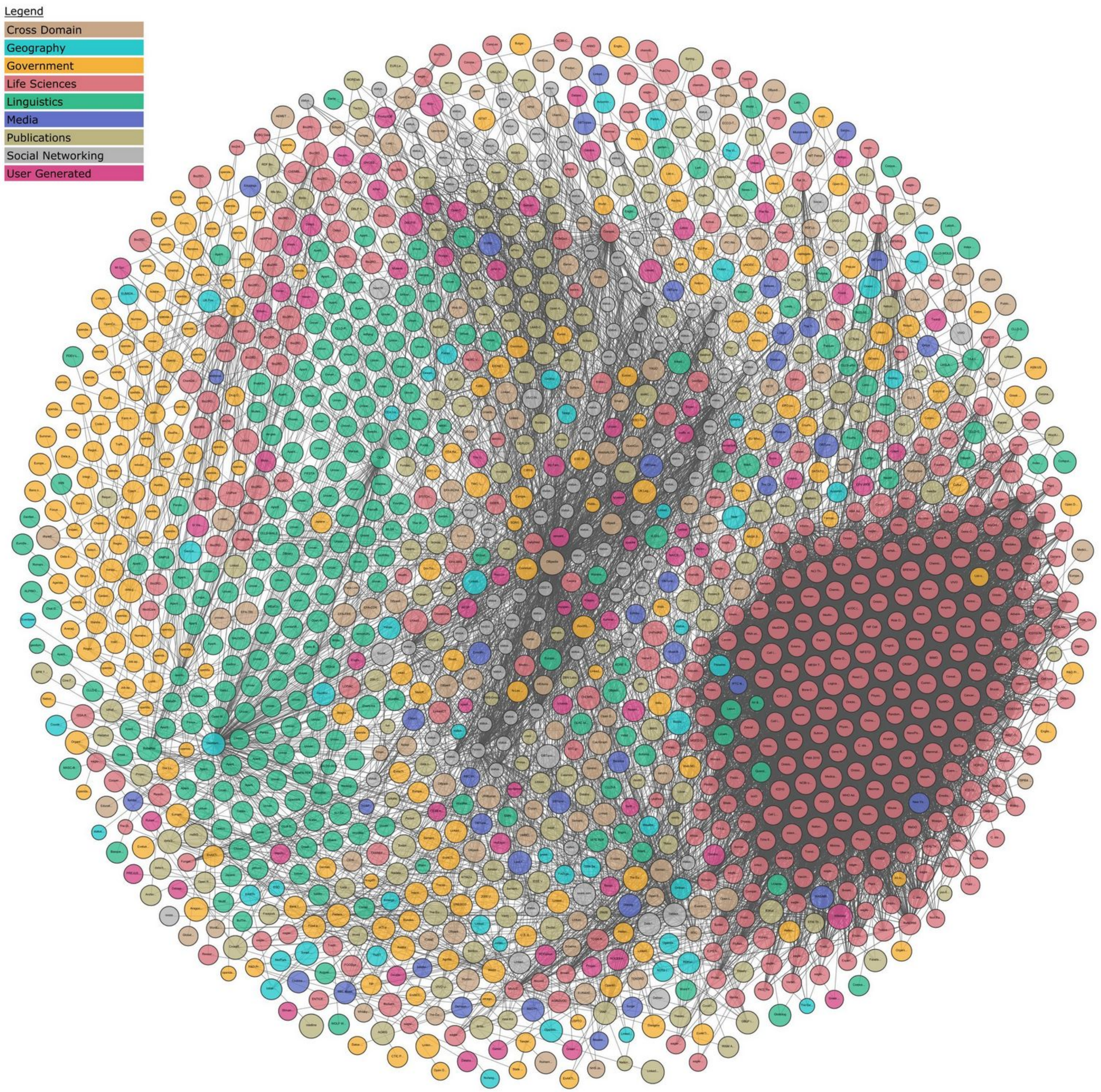
FAIR and knowledge graphs

- Metadata should be active → usable by software agents without the need of human intervention, thanks to the resolvable links found in the electronic documents and the associated semantics available to the agent.
- Providing metadata about a dataset in the form of Linked Data Graph is a significant path towards making data FAIR.

But FAIR data and knowledge graphs are not equivalent. Not all FAIR data is a knowledge graph and not all knowledge graphs are FAIR

https://faircookbook.elixir-europe.org/content/recipes/introduction/FAIR-and-knowledge-graphs.html

The Linked Open Data Cloud



https://lod-cloud.net/

Ten simple rules

- ente che custodisca il vocabolario
- avere una licenza
- controllare termini e definizioni siano univoche e comprensibili
- tracciare il processo di costruzione del vocabolario
- assegnare un URI univoco e persistente
- termini che siano machine-readable utilizzando modelli come SKOS E owl
- metadatare il vocabolario
- registrare il vocabolario in una repository
- rendere il vocabolario accessibile
- implementare un processo di revisione del vocabolario con l’aiuto degli esperti nel rispetto dei principi FAIR

Best practices

- URI che devono seguire i principi dei Linked Data
- sostenibilità a lungo termine
- versioning

ITINERIS

THANKS!

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System (D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 “Education and Research” - Component 2: “From research to business” - Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani