LEADING

LIS Education And Data Science Integrated Network Group

LEADING Boot camp, Monday, June 14, 2021, 3:30-4:20 PM EST

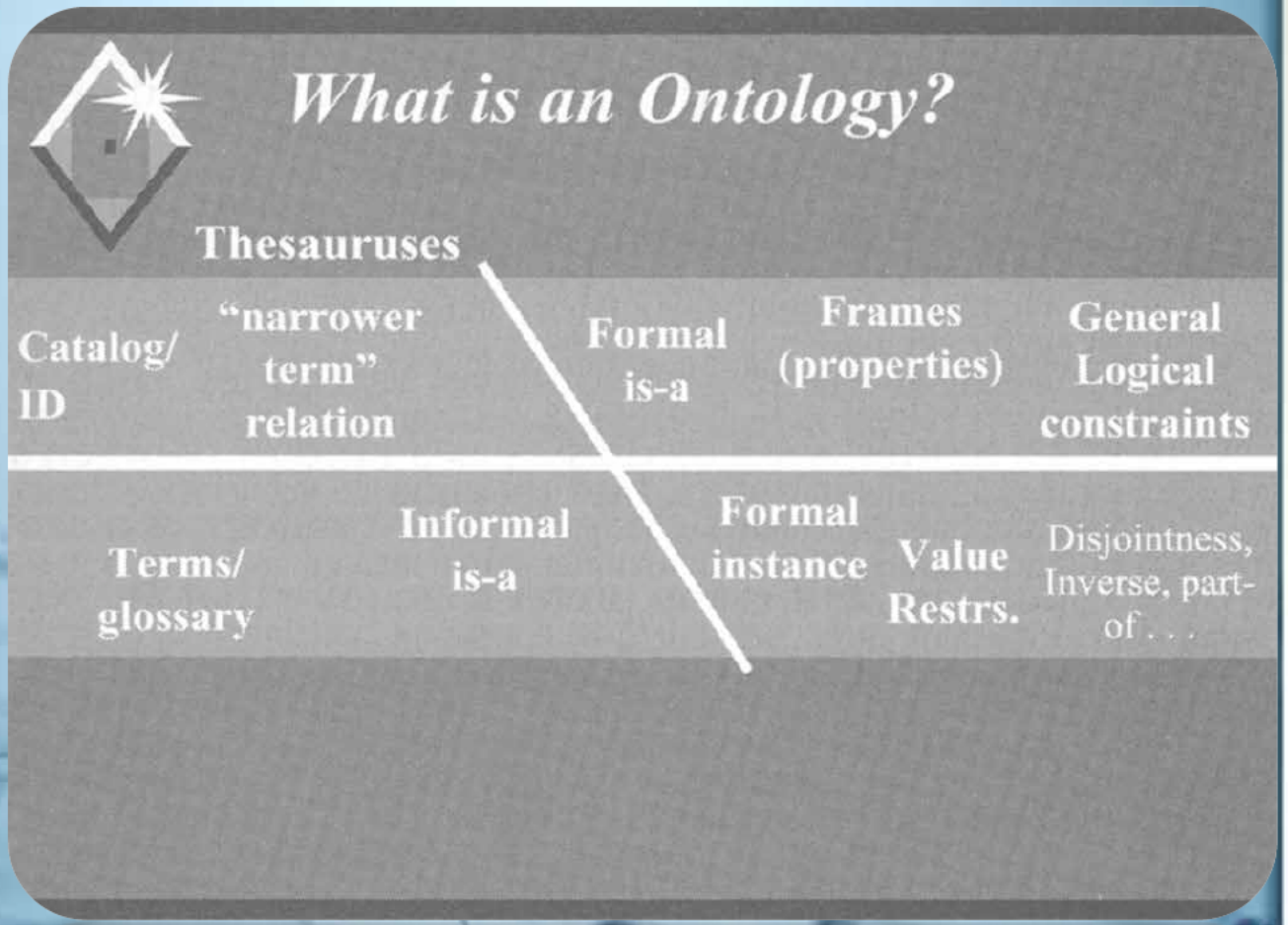
Agenda:

1. How's everyone doing
2. Ontology/Linked data
   ~ Protégé
3. Big metadata (brief comments)
4. Other/open discussion

- Metadata is a first-class data object
- There's lots of unstructured data, but…. a lot of data has metadata or needs better metadata to use it.

(McGuinness, D. L. (2003). Ontologies Come of Age.
In Fensel, et al, *Spinning the Semantic Web*.
(Cambridge, MIT Press)

Ontology



Ontologies and reality

Philosophy

Dates back to 5th Century B.C. when Empedocles divided the world into four elements – earth, fire, water and air.

Metaphysics: Defined by philosophers as the *nature of being or* ***existence***.

Aristotle · classification

Touches on "epistemology," which" is about knowledge and knowing

Ontology defined

Webster's dictionary

1. A science or study of being specifically, a branch of metaphysics relating to the nature and relations of being.
2. A theory concerning the kinds of entities and specifically the kinds of abstract entities that are to be admitted to a language system.

Thomas Gruber

"An ontology is a specification of a conceptualization"

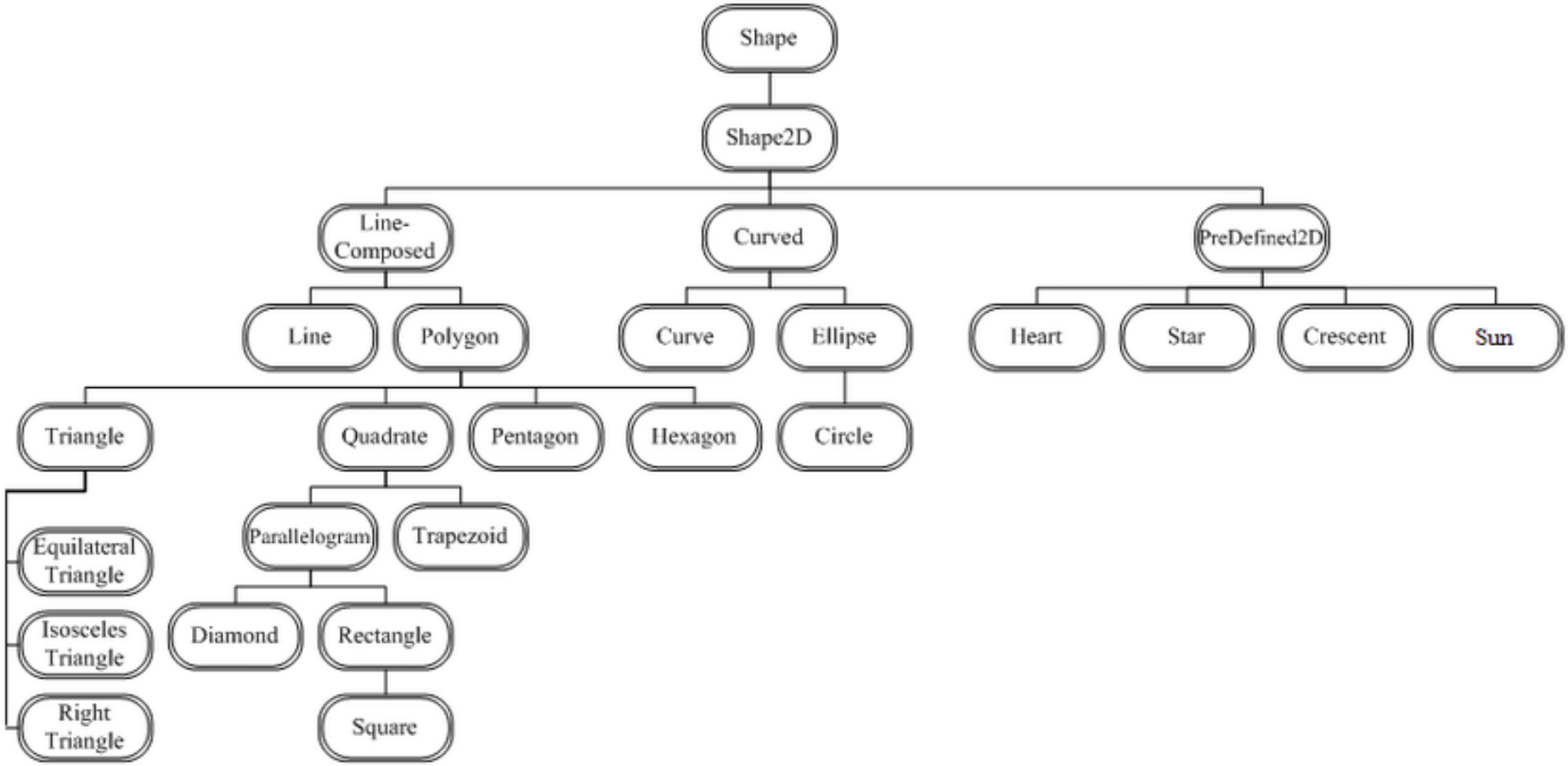
A must read!

Other important stuff

(comfy space)

- No exact definition
- *Generally speaking - ontology is* a WAY to convey a theory on how to represent a class of things and their relationships
- Knowledge representation, Knowledge organization (ILS)
- Blurry lines – w/database structure and knowledge graphs

AN ONTOLOGY OF SHAPES

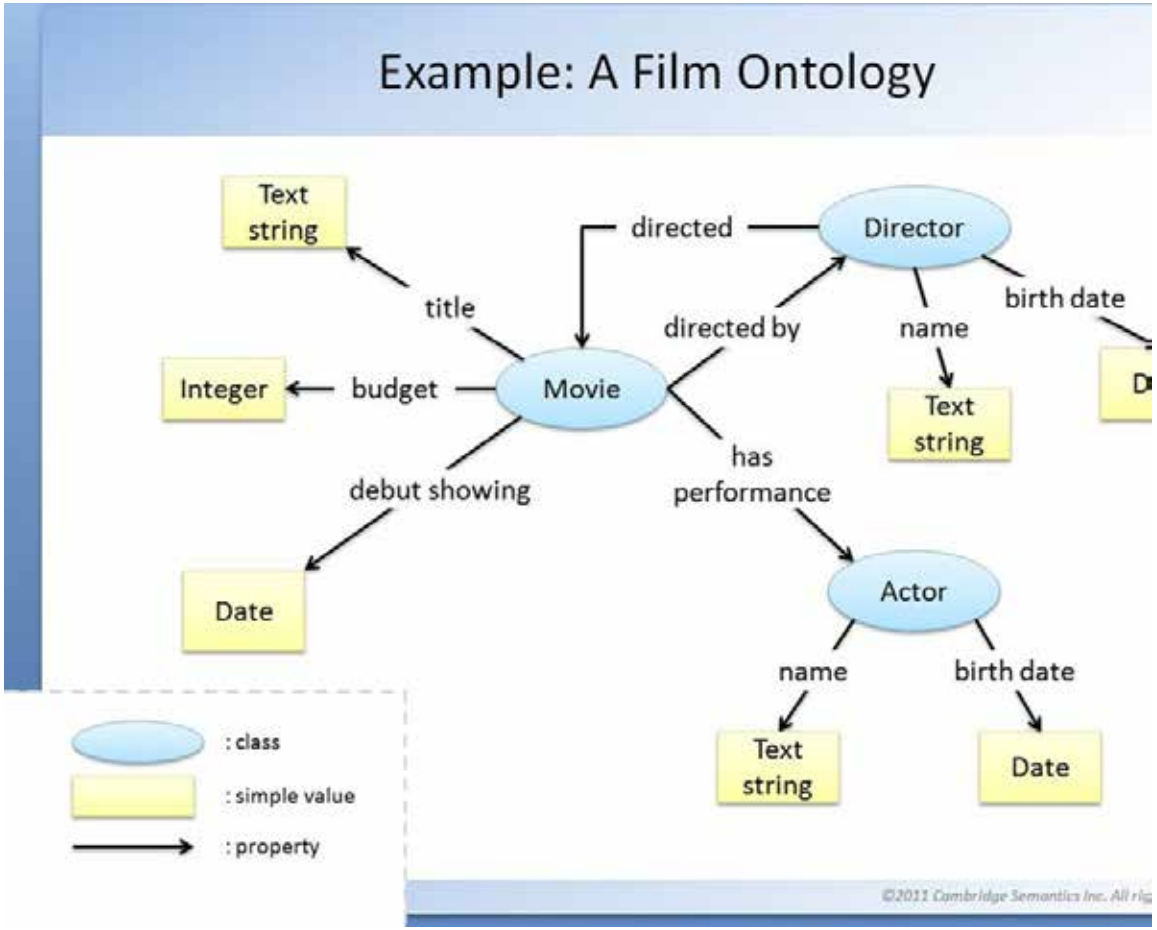


http://ceur-ws.org/Vol-812/paper9.pdf

Ontology

- National Center for Biological Ontologies: https://bioportal.bioontology.org/

- Example of a Film ontology à

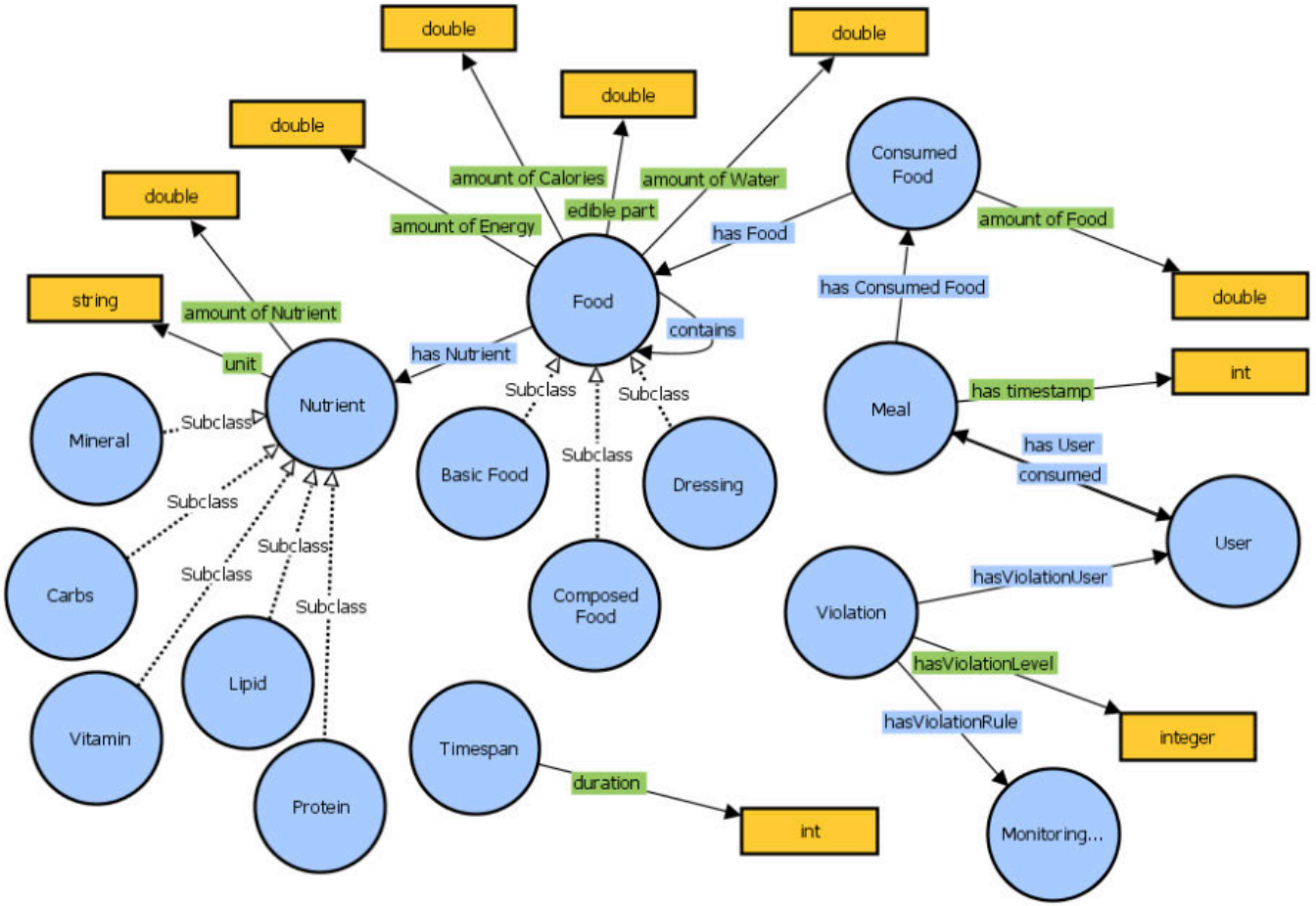


- Most ontologies structured as taxonomies, hierarchies
- People talk about ontologies in many, many, many… ways
- Basic ontology has two classes of elements: the entities and the relationships between them

- Class and sub-classes: Actor, Child actor
- Relationships
  - Is a type of
  - Movie "is directed by" director
  - Movie "has performance" actor

  This is more involved than a lexical-semantic thesaurus.
- More involved / RDF triples
  - Predicate (has a property of X)
  - E.g., Book has title
    - predicates are linguistic entities
  - Subject-predicate-object (think reverse ABC)
- Organized according to axioms or rules that control how the world will be defined.

AN ONTOLOGY OF FOOD



https://www.w3.org/community/owled/files/2016/11/OWLED-ORE-2016_paper_3.pdf

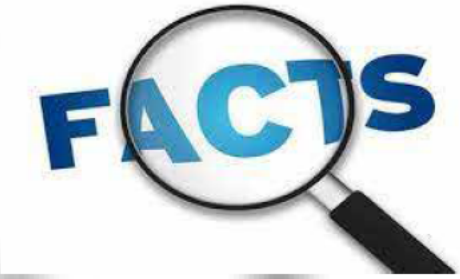
Let's go to Protégé

We will work with OWL: Web Ontology Language: https://www.w3.org/OWL/

- Modeling language
- Enabling technology

SKOS – Simple Knowledge Organization System: https://www.w3.org/2004/02/skos/

Important Facts



- What exists is only what is represented in the ontology
- Most ontologies focus on a specific area to conceptualize the world.
- Must be updated to keep up with dynamic world
- No set discipline or methodology!
- Joined together with instances, get closer toward knowledge graph… but fuzzy line

The Role of Ontology

Application areas

- Indexing
- Knowledge Sharing & Reuse
- Artificial Intelligence (AI)
- Enterprise Modeling
- Software Design
- Molecular Biology
- eCommerce
- Semantic Web….

Indexing, browsing, finding, visualizing…

A language, reliable sharing of information

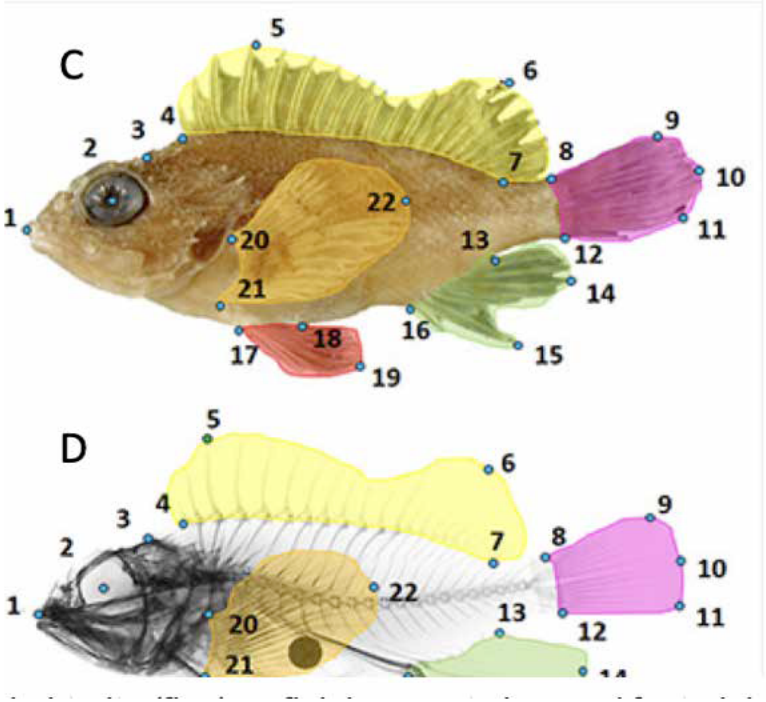
Enable reuse of domain knowledge

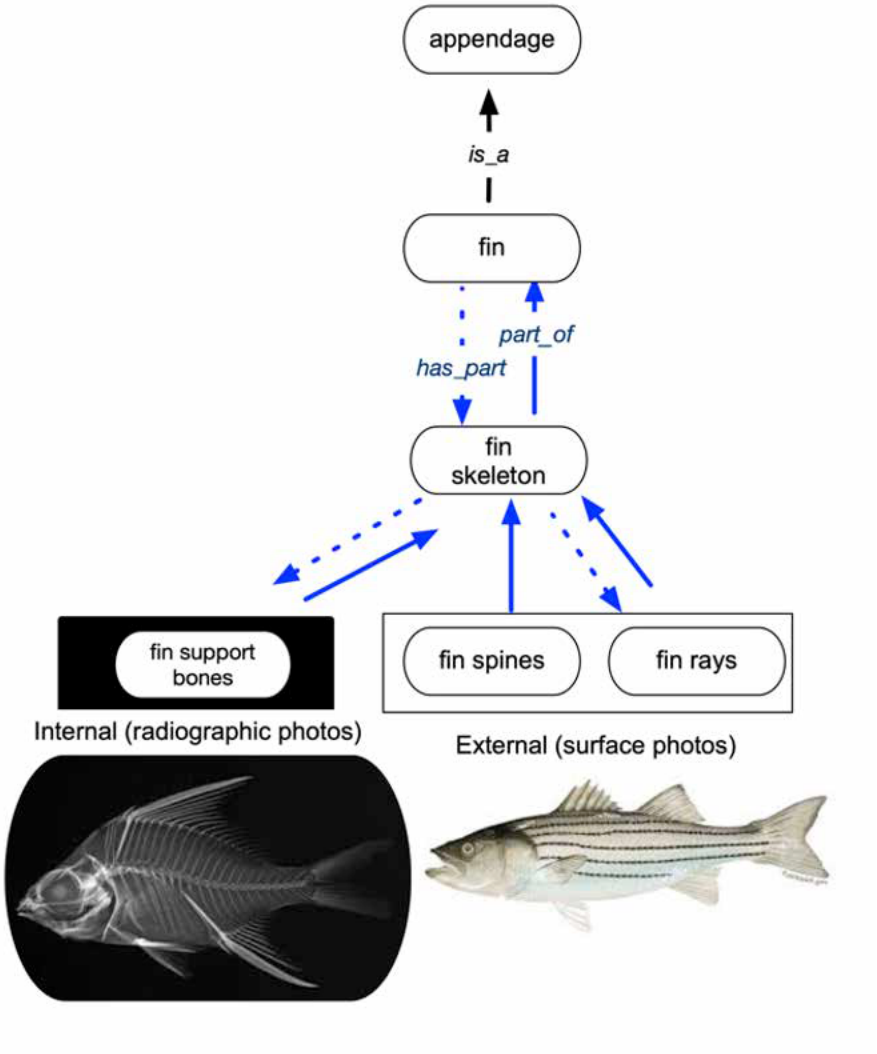
Makes assumptions more explicit

Analyze domain knowledge

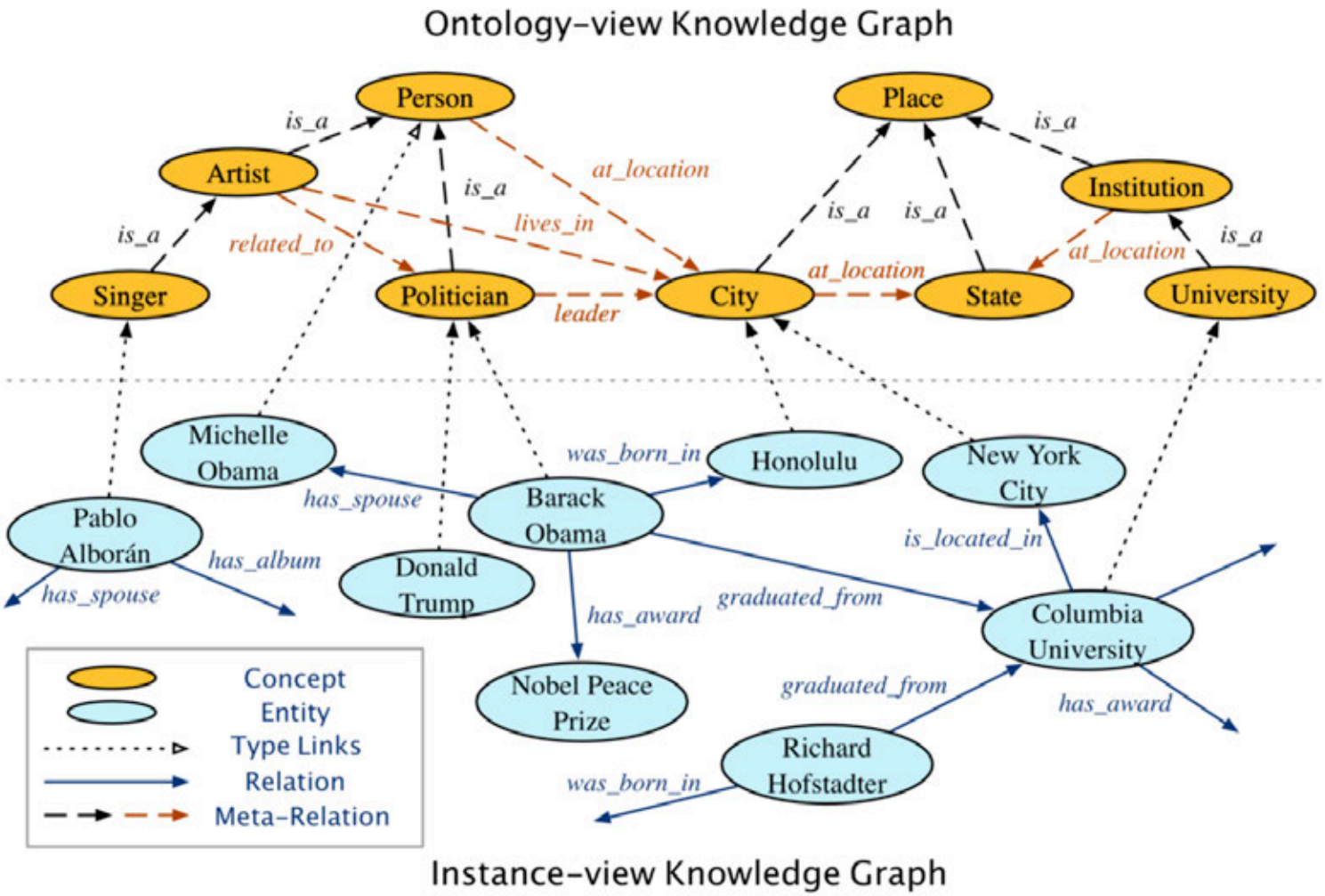
Reasoning

Reasoning to identify species variation





NSF-HDR Biology Guided Neural Networks (BGNN)



Publication: The 25th International ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2019)

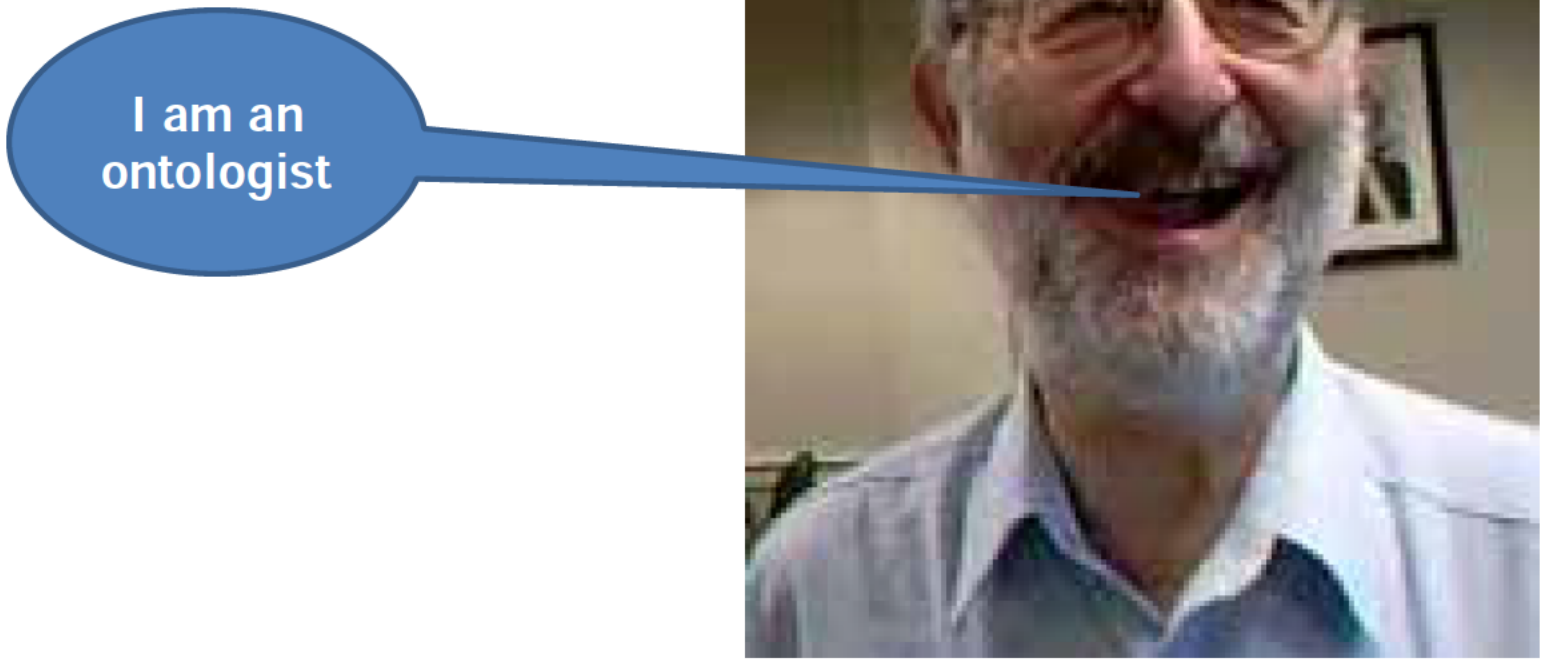
https://www.haojunheng.com/project/joie-kdd/

Ontologies as "structured terminologies"

- An island "is a type of" land area
- A nation "is a type of" geopolitical area

Relationships

- *Is at type of*
- *Causes*
- *Synonyms*

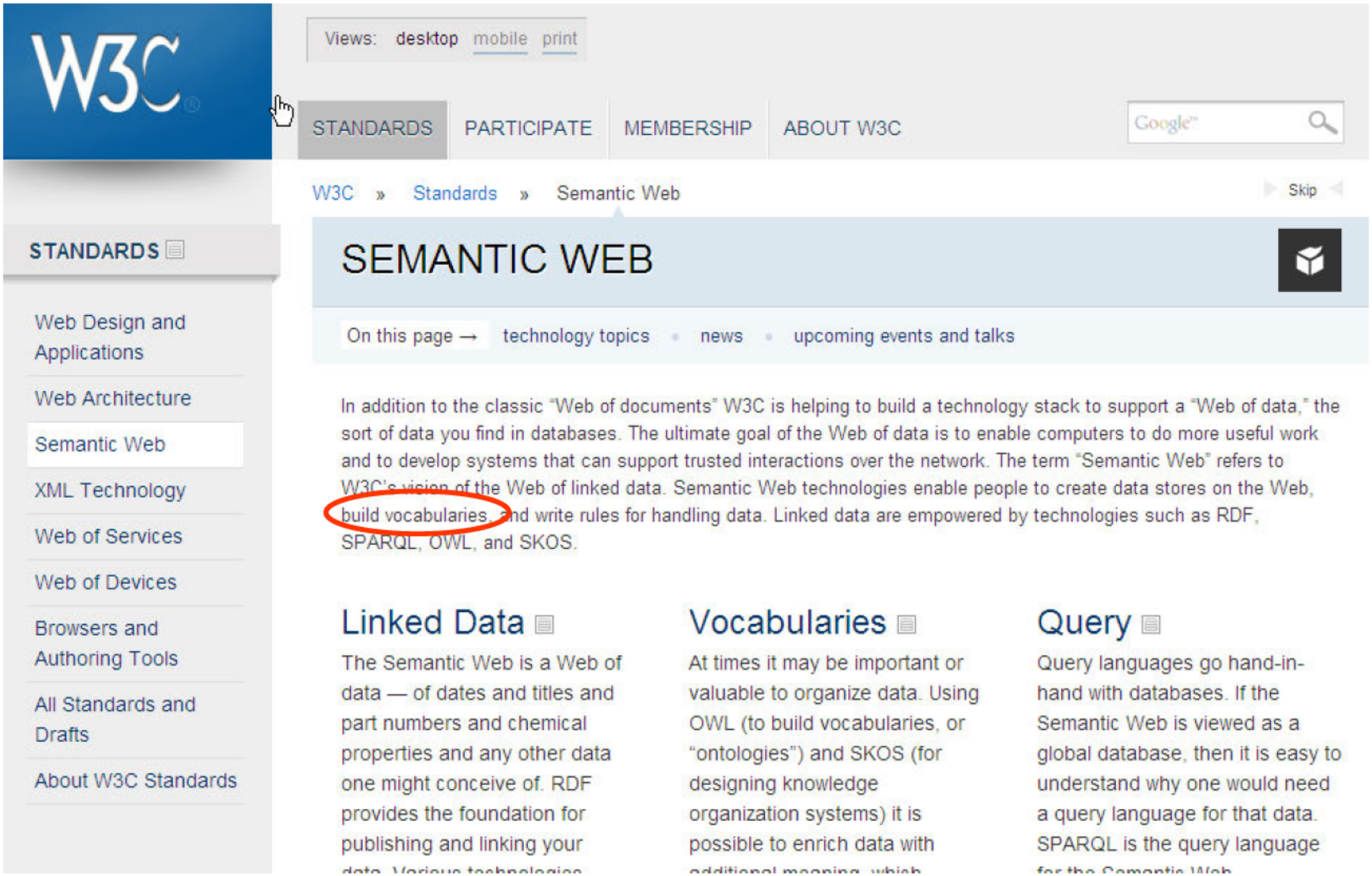


Dagobert Soergel, ACM/CL, 1997
http://nkos.slis.kent.edu/busch/summary.pdf

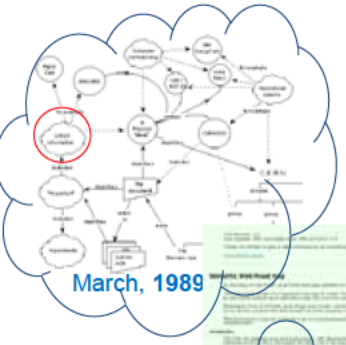
Linked data

- Strings and things
- Semantic Web vision
- Everything should have a unique ID, in fact a Persistent ID (PID)

Evolution of the Web http://www.w3.org



(E. Mendez, 2012)



Tim BL Roadmap on **Semantic Web** (Sept1998)

W3C Semantic Web Activity starts, 2001

FOAF (Early 2000)

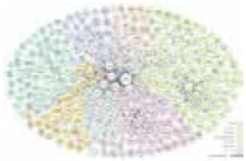
Berners Lee, Hendler & Lassila (May2001)

FOAF Voc. Spec (May2005)

1st **LOD Cloud** Cyganiak & Jentzsch (May2007)

Tim BL **LOD principles**

Tim BL TED Talks (Feb 2009 & Mar 2010)



Last **LOD Cloud** Cyganiak & Jentzsc (Nov2011)

OIL
DAML

DCMES (ISO-2003)

Tim BL **Linked Data** (July2006)

DCMI *Singapore Framework* (Aug2007)

1996 1998 2000 2002 2004 2006 2008 2010 2012

PICS W3C-REC–Oct96

RDFS (RDF Schema) 1st Draft –1998

RDF Model & Syntax 1st Draft–Oct97

RDF M&S W3C-REC, Oct99

RDFS (RDF Schema) W3C-CR, March1998

RDFS Vocabulary Description Language

W3C RDF

RDF M&S W3C-REC, Feb2004

SKOS 1st Draft (May2005)

W3C SKOS

W3C-REC Aug2009

W3C Incubator Activity

Library Linked Data XG (May 2010) Final Reports (Oct2011)

W3C SPARQL

W3C-REC, Jan2008

MCF Meta Content Framework–Jun97

W3C OWL

OWL specifications W3C-REC, Feb2004

W3C OWL2

OWL2 specifications W3C-REC, Oct2009

W3C Linked Data Platform LDP XG (May2012)

W3C RDFa

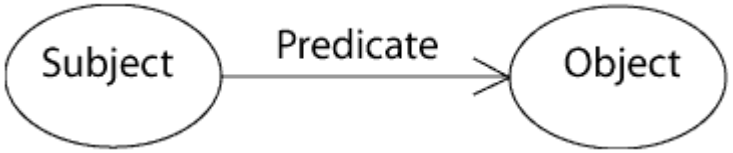
In XHTML W3C-REC, Oct2008

Linked Data based on RDF data model

Is similar to Entity-Relationship or Class diagrams, statements about resource in subject-predicate object expressions called "triples".

subject= resource

predicate=traits or aspects of the resource and expresses a relationship between the subject and the object.



http://www.w3.org/TR/rdf-concepts/

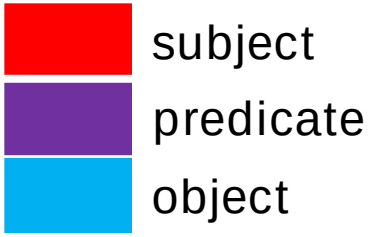
The sky has the color blue

RDF triple:

- A subject denoting "the sky"
- A predicate denoting "has the color"
- An object denoting "blue"

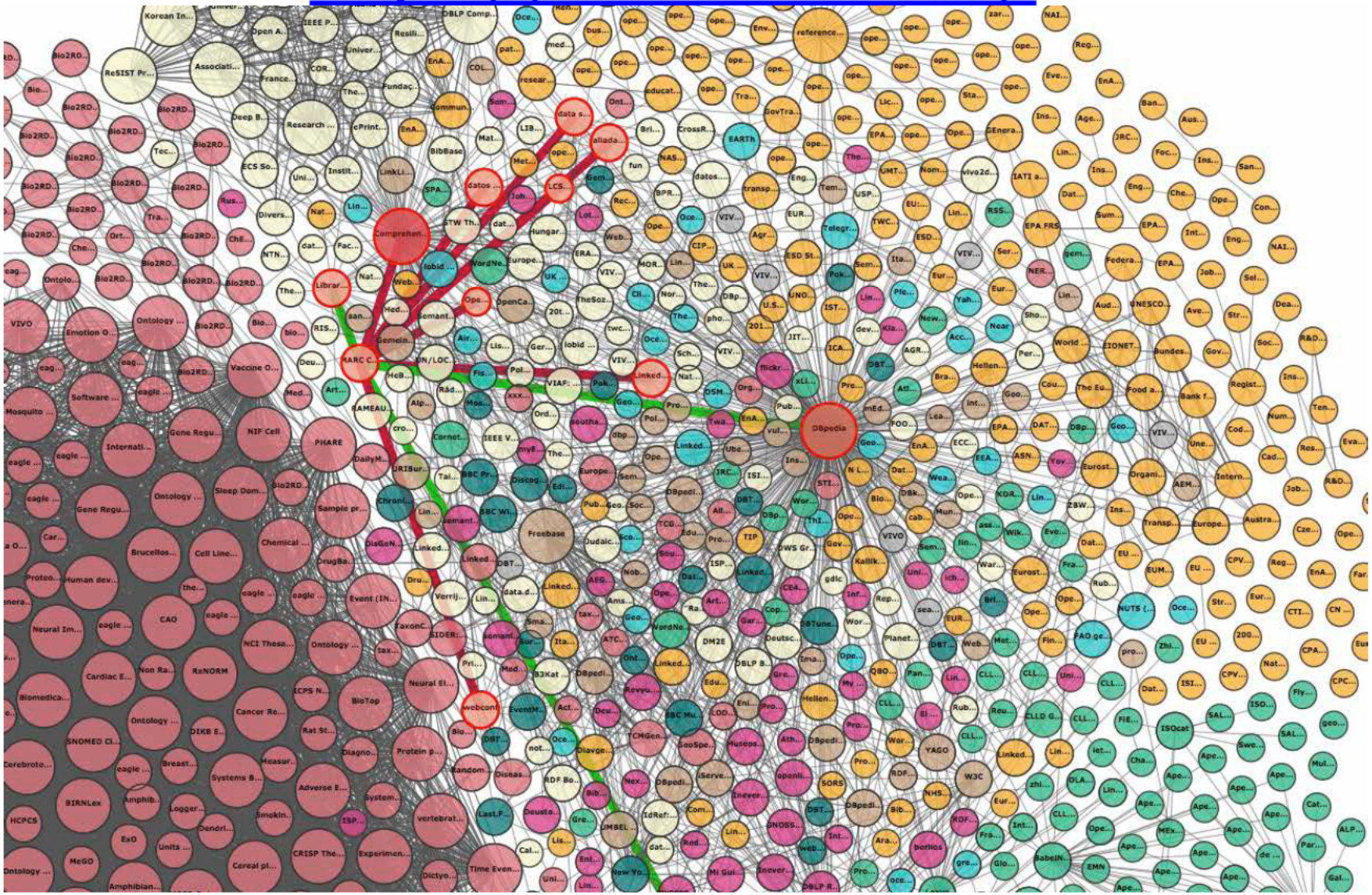
Example of RDF

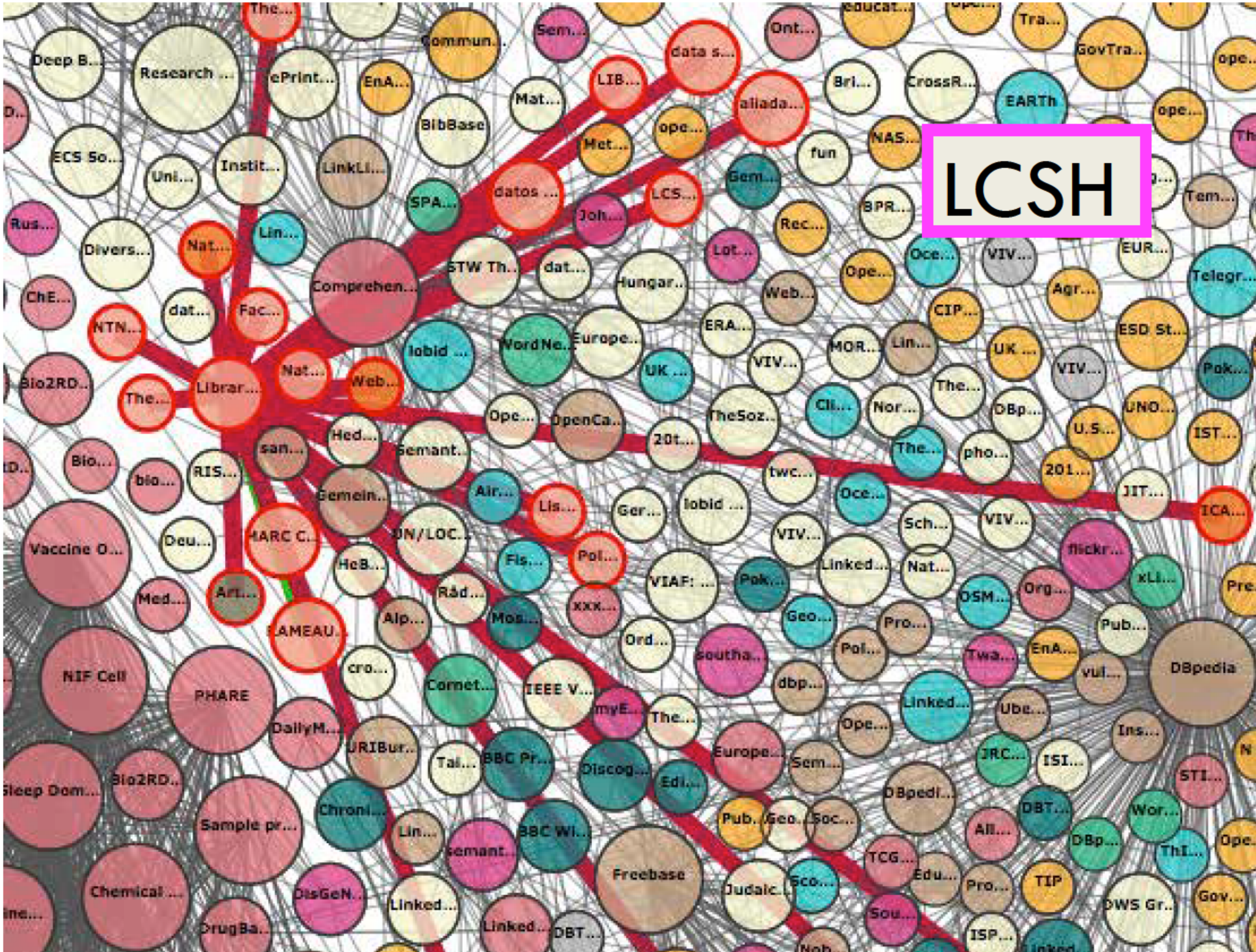
```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/">
<rdf:Description
rdf:about="https://www.sciencemag.org/news/2020/07/during-pandemic-students-do-field-and-lab-work-without-leaving-home">
<dc:title>During the pandemic, students do field...
</dc:title>
  </rdf:Description>
</rdf>
```



https://www.sciencemag.org/news/2020/07/during-pandemic-students-do-field-and-lab-work-without-leaving-home

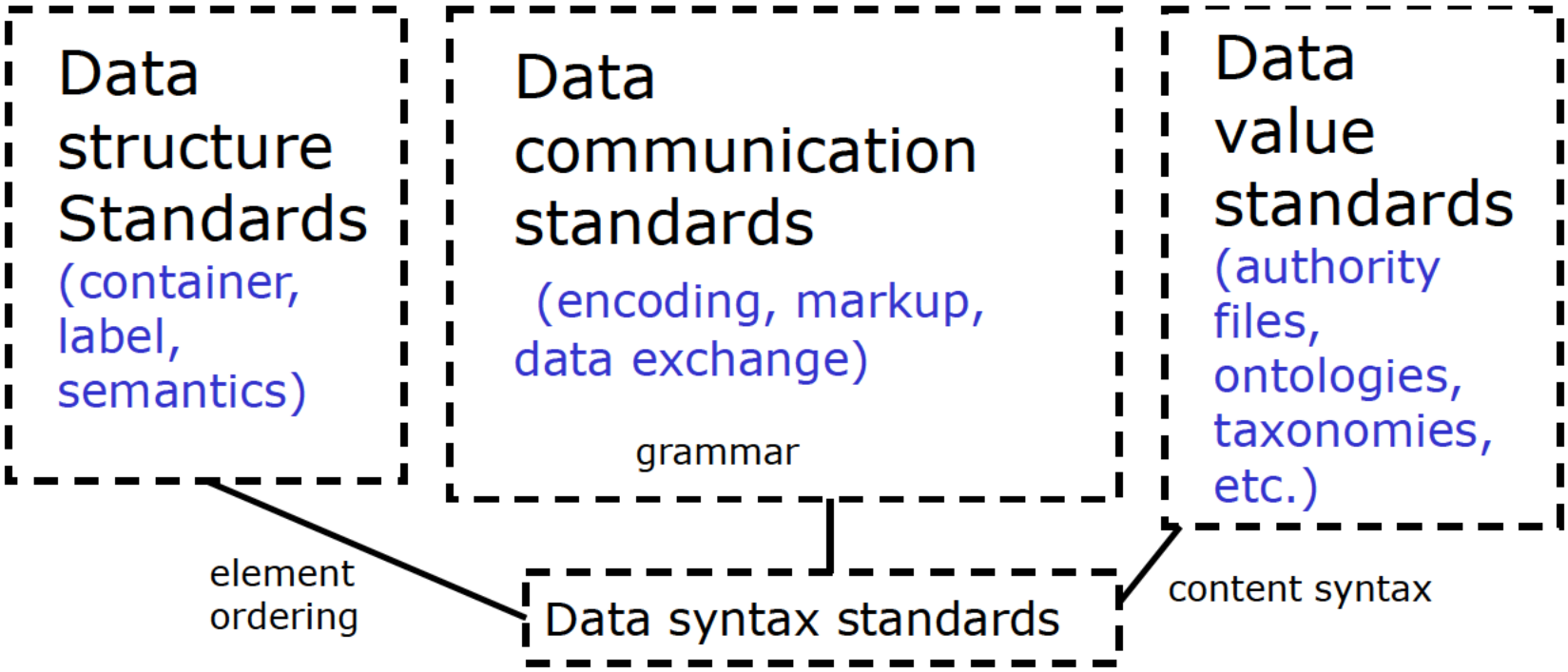
https://lod-cloud.net/




LCSH
BibBase
EARTh
fun
DBpedia
Freebase
PHARE
NIF Cell

Peter Fox, RDA,
Sweden, 2014

World of Metadata Standards



Big Metadata

Table 1. The five Vs of big metadata.

Five Vs	Definition
Volume	The quantity and usefulness of metadata generated daily confirms the existence of big metadata. At times metadata is less than or equal to the extent of the data it describes in size (bytes). During other times the metadata exceeds the data being described or tracked, due to the complexity of the data lifecycle activity. Linked data offers an example, with metadata renderings that can be larger than the volume of data object(s) being represented. Like big data, not all big metadata is useful, and a challenge is to identify the big metadata that is useful for data science and analytic endeavors.
Velocity	Metadata is generated via automatic processes at immense speed correlating with rate of digital transactions. For example, searching Google, answering an email, purchasing an item online, and day-to-day office activities such as word processing of all log data, as well as associated metadata.
Variety	Metadata reflects the wide variety of data formats, types, and genres along with the extensive range of data and metadata lifecycles. In addition, the different types of metadata (e.g. discovery, technical, preservation, etc.) as well as unique domain specific metadata requirements intensify the variety.
Variability	There is an unmistakable unevenness of metadata across the digital ecosystem. Lack of uniformity is extensive for data descriptions across different domains, systems, and processes. This unevenness can even be profound within domains, given economic factors supporting metadata generation, competing standards, or, simply, differing adoption policies. For example, two organizations may use the same metadata standard, but have different implementation practices. Even when standardization is imposed, an organization, process, and human activity can contribute to inconsistencies.
Value	*If data is the new black gold*—akin to petroleum requiring purification, but also a money maker, then metadata is the new platinum—a malleable substance that keeps its toughness, and can serve as a catalyst, sparking a reaction.* Metadata, as the *new platinum*, can be modified, while remaining a strong, independent data type. Metadata stands as a durable data object that triggers various functions—the catalyst, and achieves results—a reaction. Metadata is vital to accurate data interpretation and use by both humans and machines, and the value of metadata for data science endeavors cannot be overstated or diminished.

Note. *Singh (2013) identified data as the new black gold on Wired.com.

Greenberg, J. (2017): doi.org/10.1515/jdis-2017-0012