

Scaling up semantics

lessons learned from across the life sciences

Chris Mungall

Lawrence Berkeley National Laboratory

cjmungall@lbl.gov

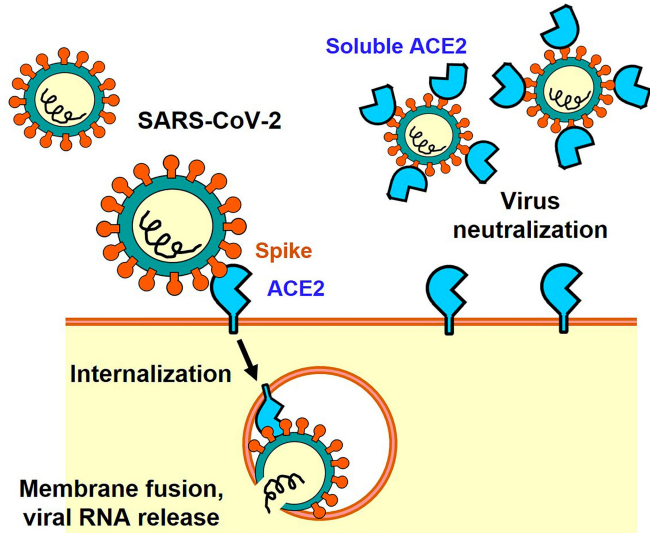
<https://bit.ly/mungall-swat-23>



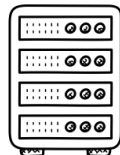
BERKELEY LAB



Providing Semantics for Systems Biology

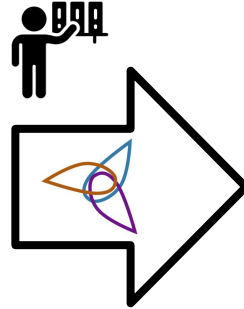
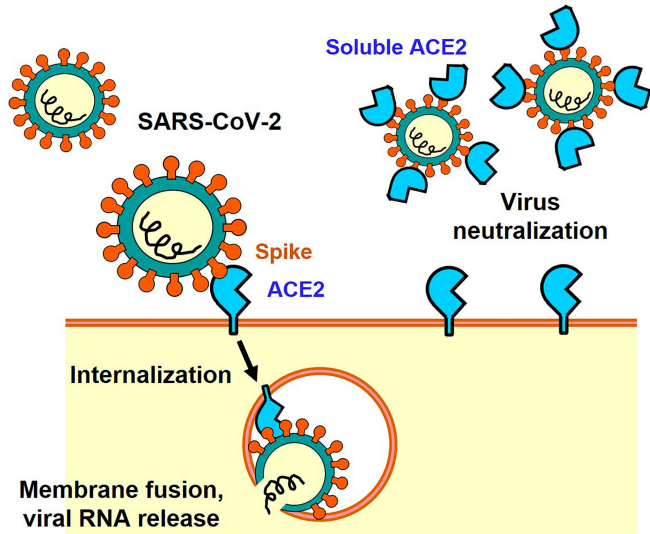


The angiotensin-converting enzyme 2 (ACE2) is a protein that has different roles such as catalytic, transporter of amino acids or viral receptor. It has an essential role in different systems, from cardiovascular to viral infection.



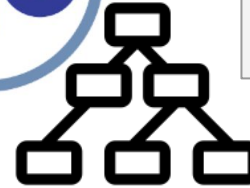
- The language of biological knowledge is text and images
- Digestion into **semantic models** necessary to be *properly* understood by machines

The Gene Ontology: Semantic Systems Biology



geneontology.org

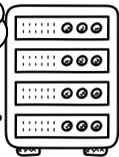
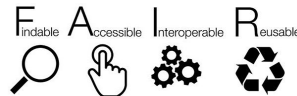
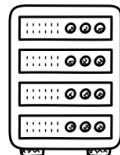
45k classes



750,000 experimentally supported annotations manually curated from 140,000 publications

1bn computed annotations

The angiotensin-converting enzyme 2 (ACE2) is a protein that has different roles such as catalytic, transporter of amino acids or viral receptor. It has an essential role in different systems, from cardiovascular to viral infection.



Semantic resources and life science applications



ontologies

semantic models



Semantic resources



Molecular biology



*Translational
research and health*



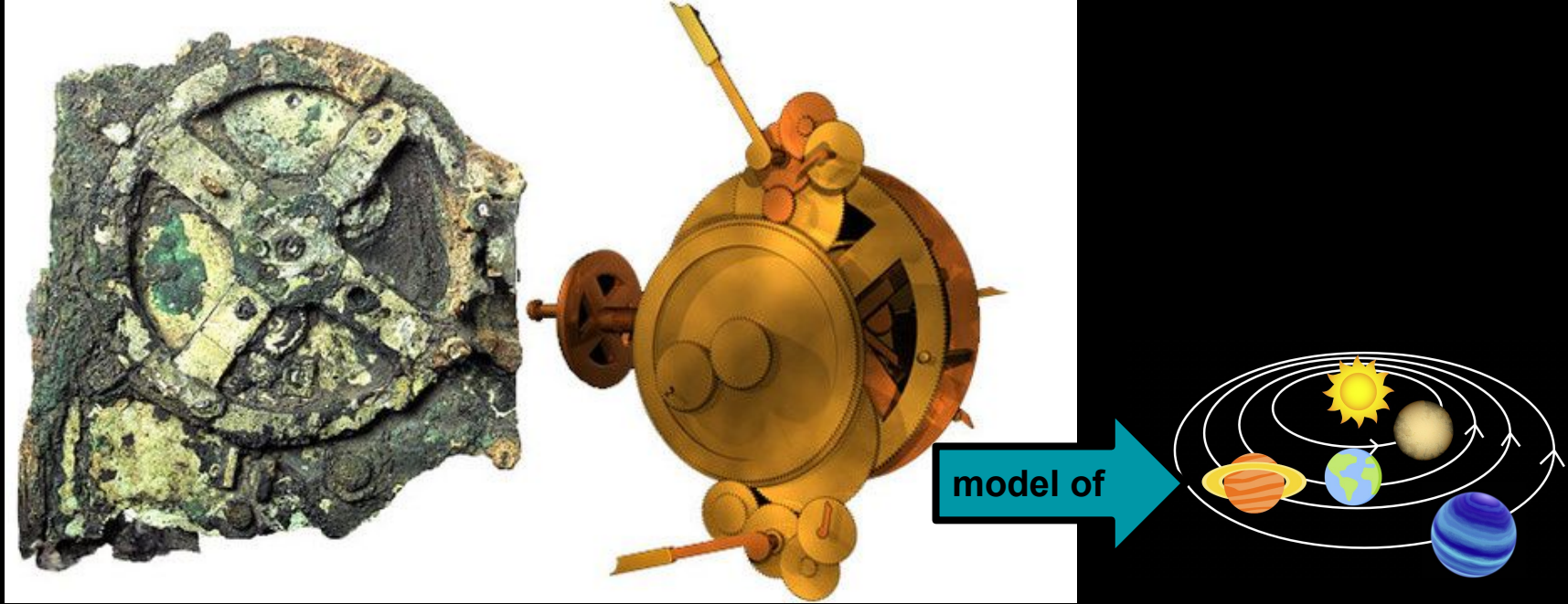
Microbiome science



multi-Omics integration

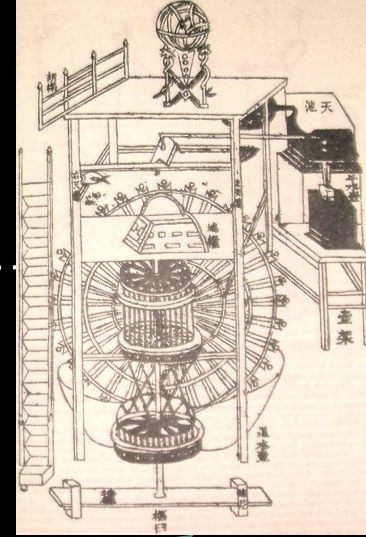
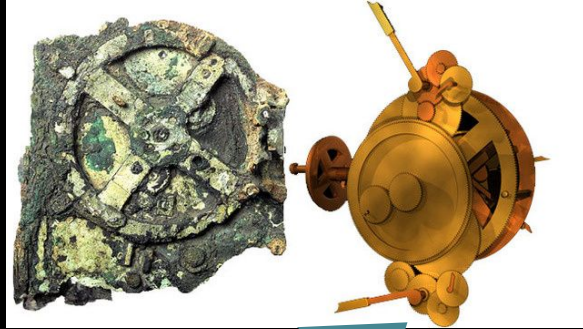
applications

We have always sought to model the world

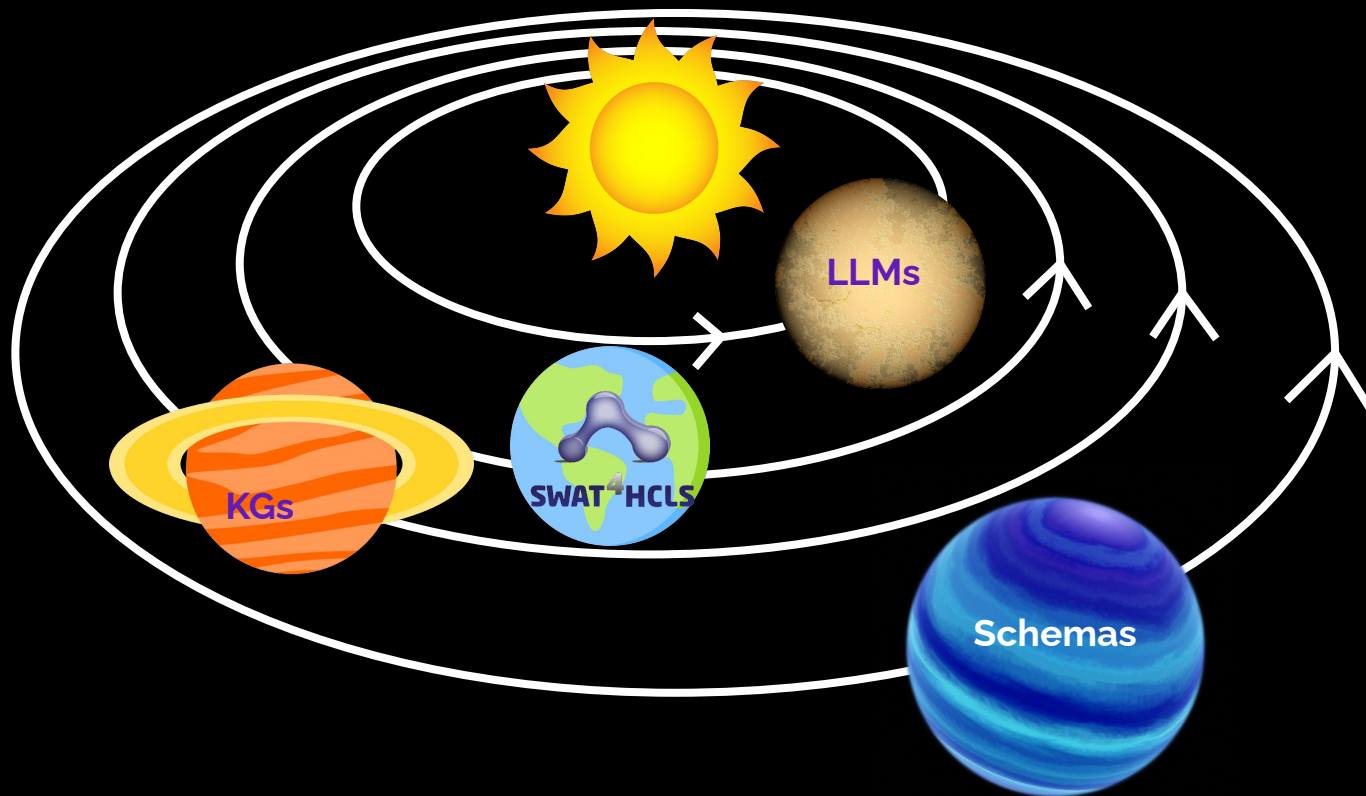


https://en.wikipedia.org/wiki/Antikythera_mechanism

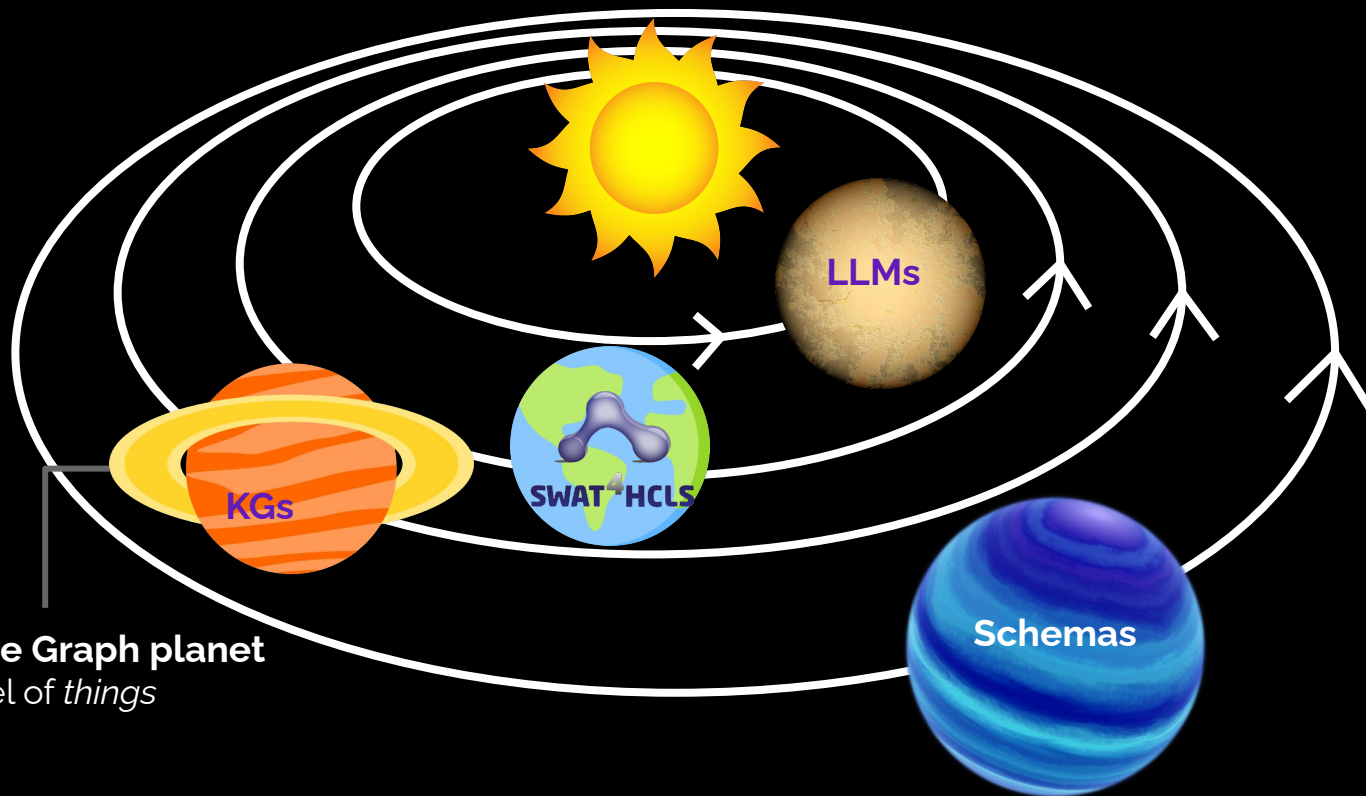
Preservation and rediscovery of knowledge



Modeling paradigms in the semantic solar system



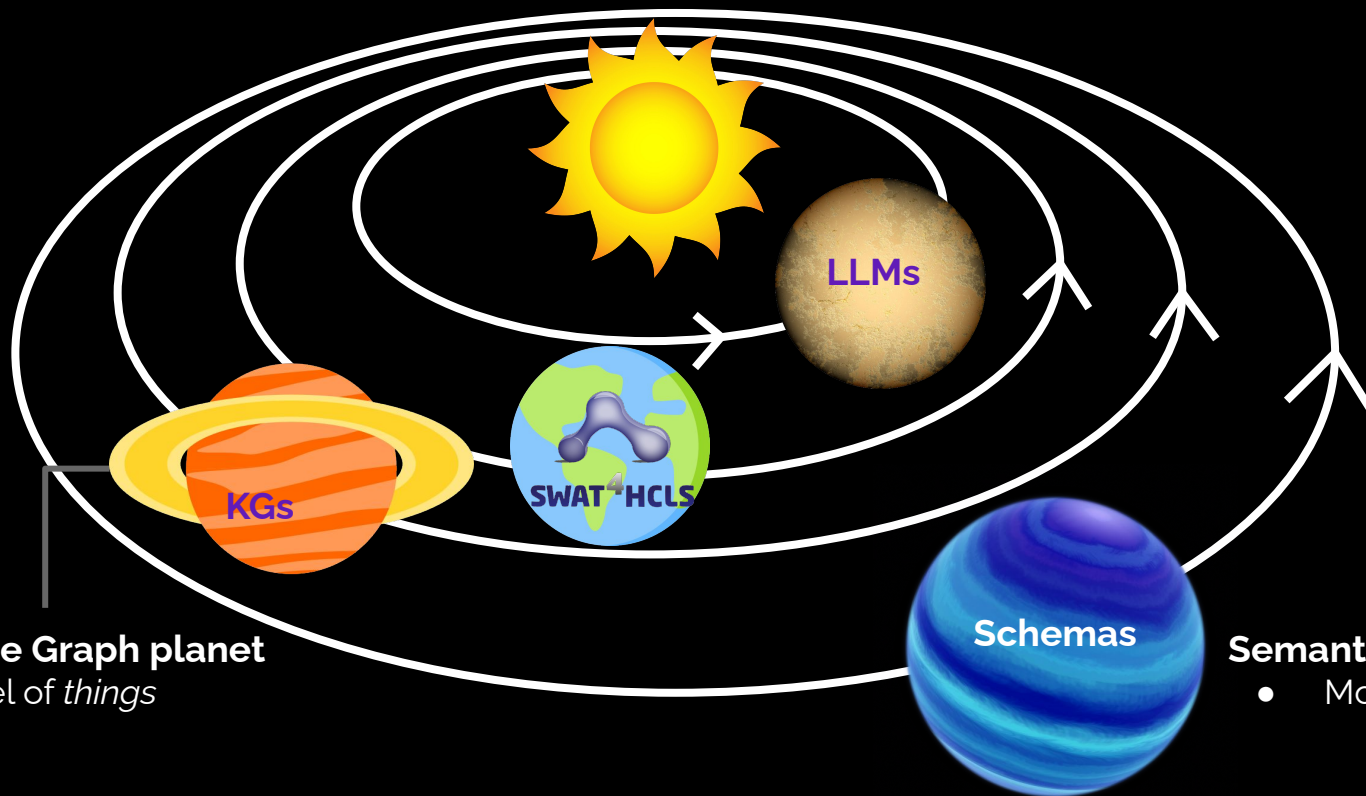
Modeling paradigms in the semantic solar system



Knowledge Graph planet

- Model of *things*

Modeling paradigms in the semantic solar system



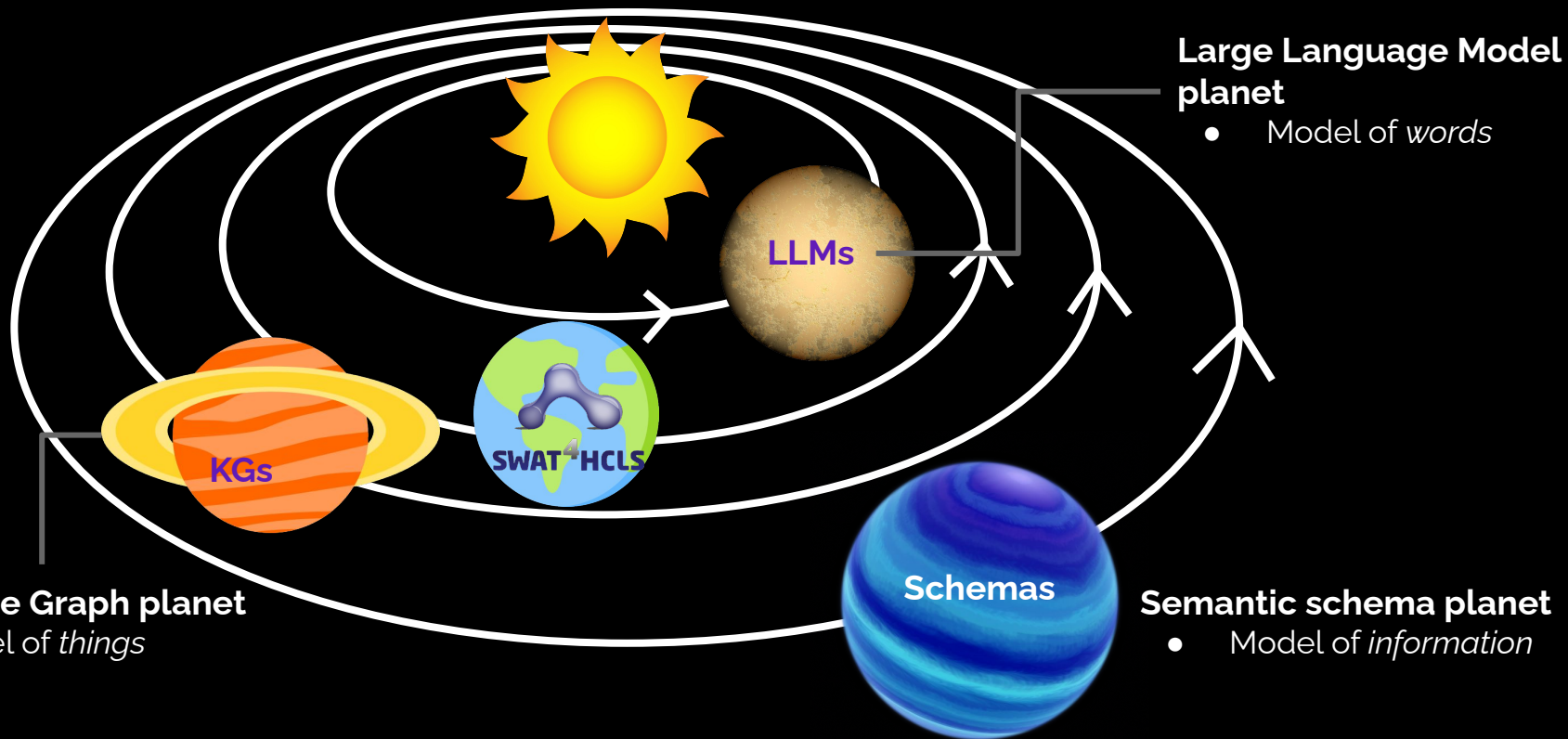
Knowledge Graph planet

- Model of *things*

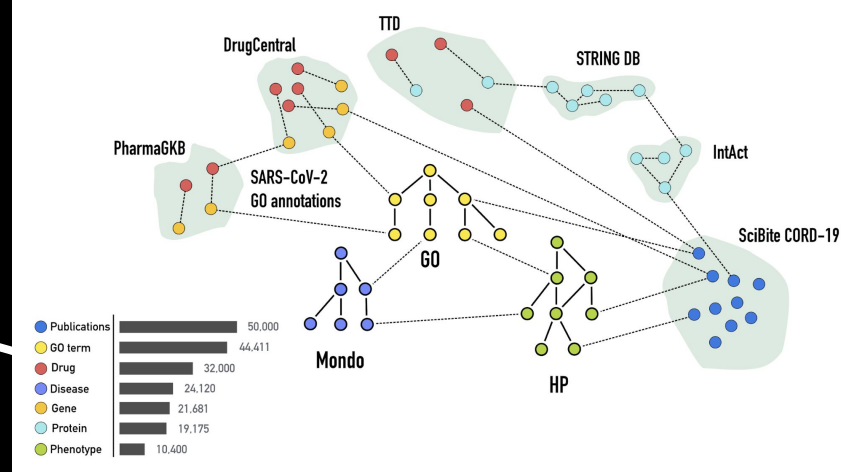
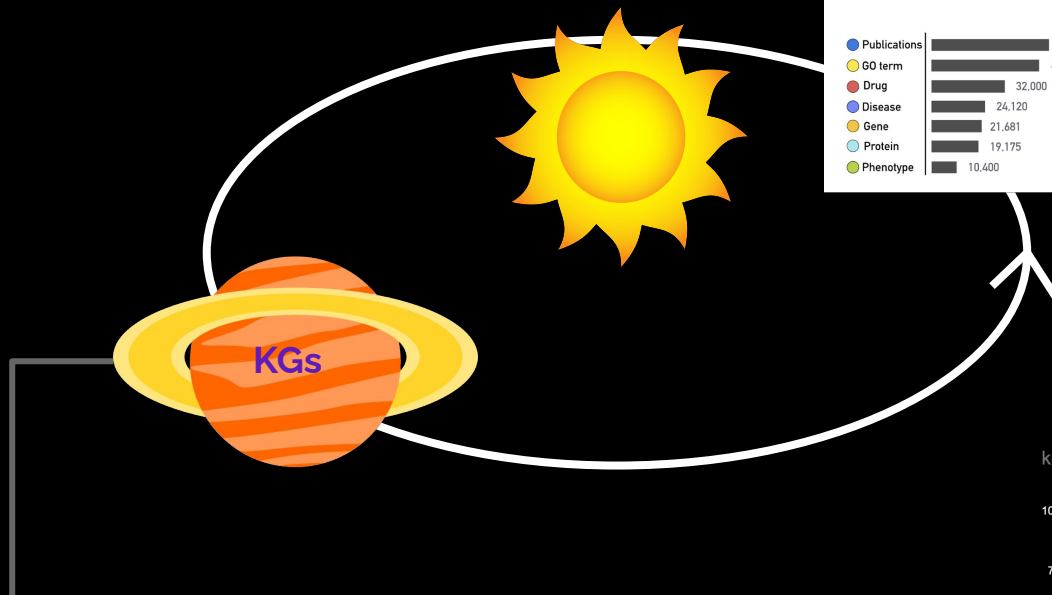
Semantic schema planet

- Model of *information*

Modeling paradigms in the semantic solar system



Planet KG: currently hot

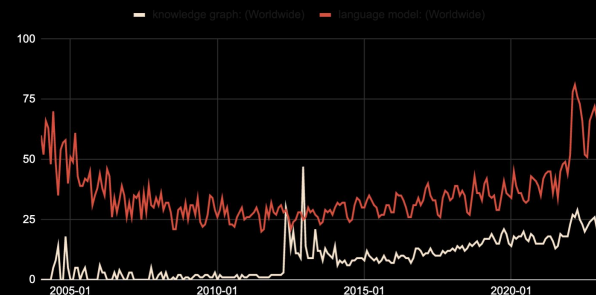


Reese J, **KG-COVID-19**
10.1101/2020.08.17.254839

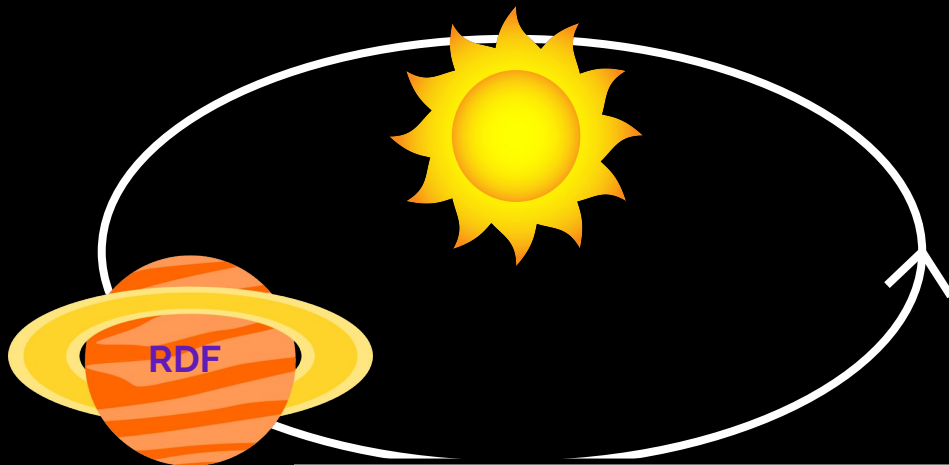
Knowledge Graph planet

- Model of *things*
- Temperature: **hot**

knowledge graph vs language model



Previous iterations of KG planet: Semantic Web 2001



Differences:

- Ad hoc IDs >> URIs
- Scale >> Correctness
- Property Graphs >> Triples
- Graphs, ML >> Reasoning
- Shiny databases >> Standards

SCIENTIFIC AMERICAN MAY 2001

The Semantic Web

A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities

By Tim Berners-Lee, James Hendler and Ora Lassila on May 1, 2001

a

DNA de

mitoc
genome

a

DNA metabolism

DNA degradation

DNA packaging

DNA replication

DNA repair

DNA recombination

mitochondrial genome maintenance

mitochondrial DNA replication

pre-replicative complex formation and maintenance

DNA unwinding

DNA priming

DNA initiation

DNA strand elongation

DNA ligation

lagging strand elongation

leading strand elongation

CDK9 mei-9 Lig1 Lig3

REV3 Rad1 Lig3 mei-9 Lig3 mus209 hay Rad51

DNA-lig I Lig1 DNA-lig II Lig3

Pcna Recc1

DNA pol- α 180

CDK2 DPB11 POL2 CDK9

mus209 CDK2 DNA pol- δ DPB11 hay POL2 Rad51

MCM2 Mcm2 Mcm22 MCM3 Mcm3 Mcm4 CDC54/MCM4 MCM4 MCM5 CDC46/MCM5 MCM6 Mcm6 Mcm66 CDC47/MCM7 Mcm7 Orc2 CDC54/MCM4 CDC46/MCM5 MCM6 CDC47/MCM7

MCM2 MCM3 CDC54/MCM4 CDC46/MCM5 MCM6 CDC47/MCM7

Mcm4 Mcm5

RNI35 Rnt1 Recc1 RNR1 Rur5 Rrm1 Rrm2

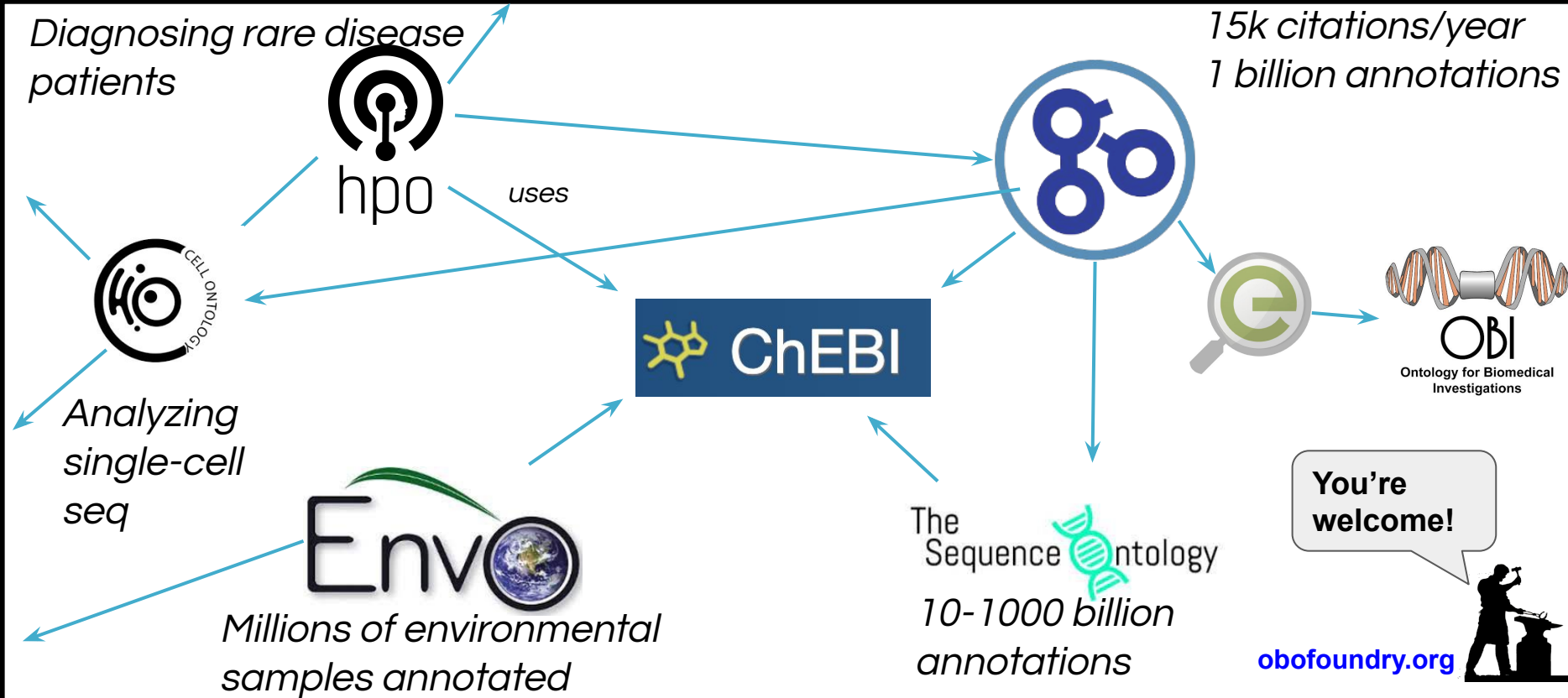
SACCHAROMYCES

DROSOPHILA

MUS

- All entities >> just terms
- Quantity >> Quality

The interconnected OBO KG 2000->present



OBO and Wikidata

glucan biosynthetic process (Q21758931)

The chemical reactions and pathways resulting in the formation of glucans, polysaccharides consisting only of glucose residues.

GO:0009250 | glucan synthesis | glucan formation | glucan anabolism | glucan biosynthesis

 [edit](#)



▾ [In more languages](#)

[Configure](#)

Language	Label	Description	Also known as
English	glucan biosynthetic process	The chemical reactions and pathways resulting in the formation of glucans, polysaccharides consisting only of glucose residues.	GO:0009250 glucan synthesis glucan formation glucan anabolism glucan biosynthesis
Spanish	No label defined	No description defined	
Traditional Chinese	No label defined	No description defined	
Chinese	No label defined	No description defined	

OBO and Wikidata

glucan biosynthetic process (Q21758931)

The chemical reactions and pathways resulting in the formation of glucans, polysaccharides consisting only of glucose residues.

GO:0009250 | glucan synthesis | glucan formation | glucan anabolism | glucan biosynthesis

 [edit](#)



▾ [In more languages](#)

[Configure](#)

has part(s)



glucan

 [edit](#)

object has role

product

▾ [1 reference](#)

stated in

Gene Ontology release 2019-11-16

[+ add reference](#)

[+ add value](#)



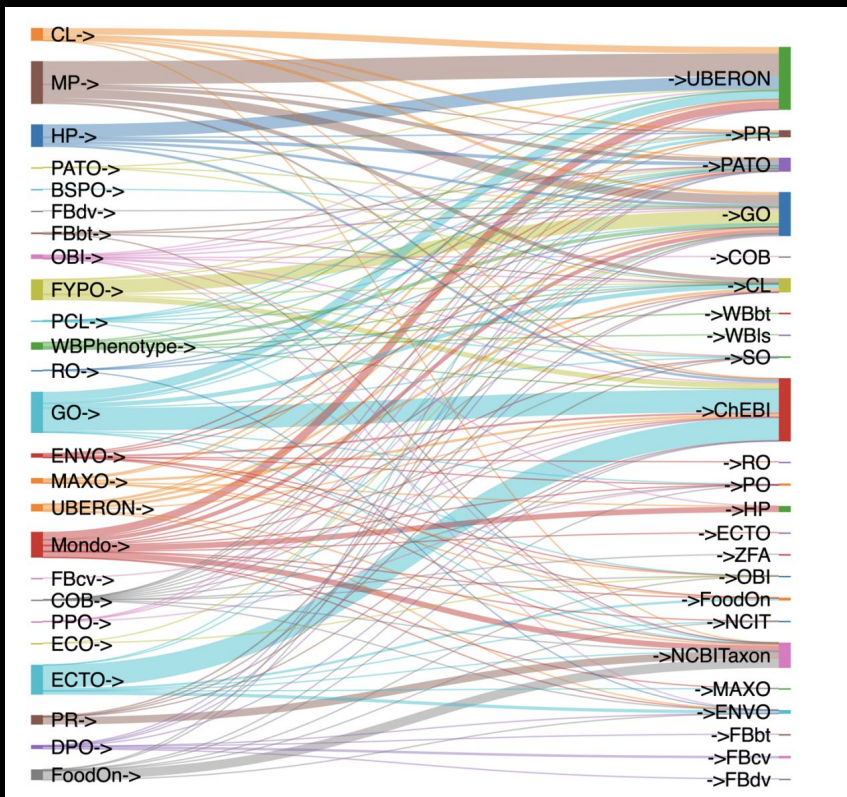
ChEBI

OBO and Wikidata

- Many more OBO ontologies to go!
 - Join us on #wikidata on OBO slack
 - <http://obofoundry.org>
- Next up:
 - Environment Ontology (ENVO)



OBO KG in Ubergraph



- Triplestore with interlinked subset of OBO
- Inference pre-entailed using Relation Graph
- Pre-categorized using Biolink
- Powerful federated queries with uniprot, wikidata, others

<https://ubergraph.apps.renci.org/sparql>

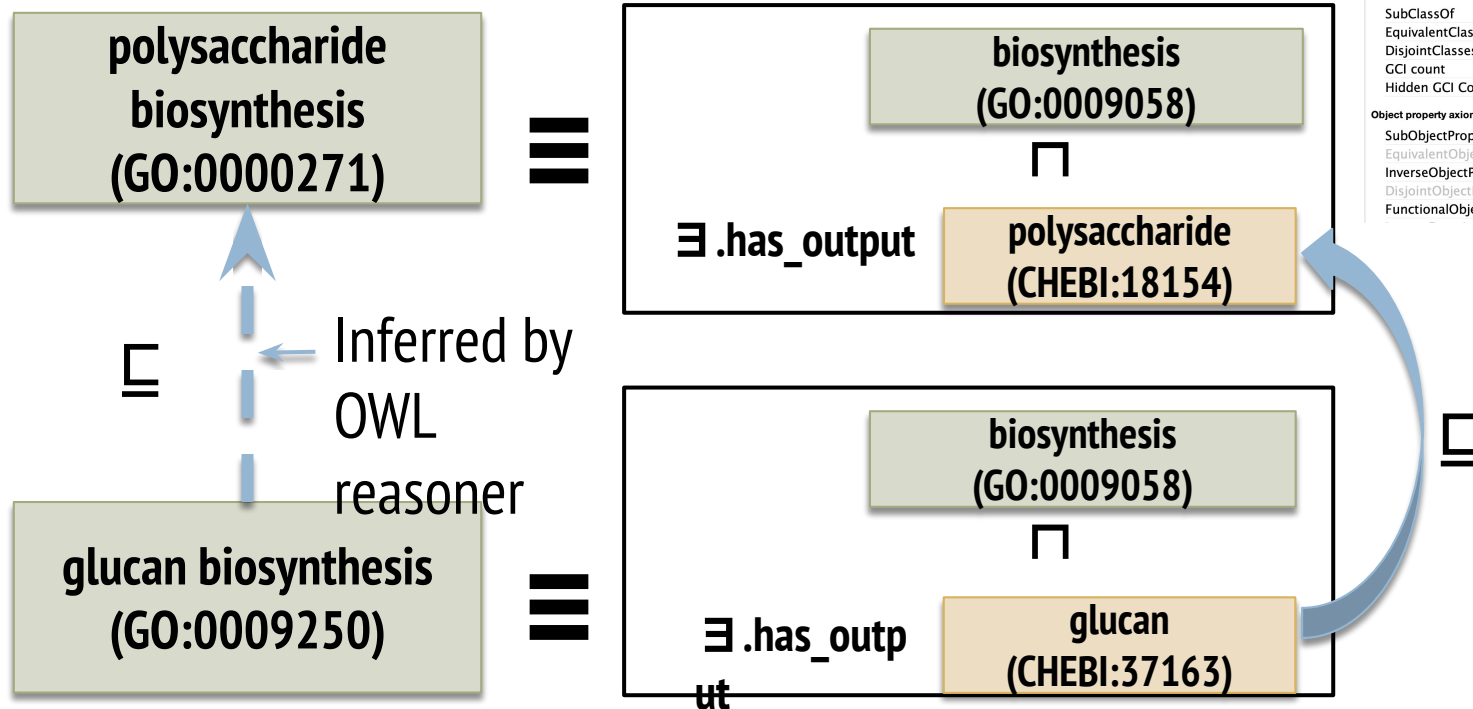
<https://github.com/INCATools/ubergraph>



*Inter-ontology connections
in Ubergraph, grouped by
OBO ontology*

https://icbo-conference.github.io/icbo2022/papers/ICBO-2022_paper_5005.pdf

Use of semantics in the connected KG



Ontology metrics:	
Metrics	
Axiom	753599
Logical axiom count	186317
Declaration axioms count	82489
Class count	82124
Object property count	314
Data property count	0
Individual count	0
Annotation Property count	58
Class axioms	
SubClassOf	114319
EquivalentClasses	69008
DisjointClasses	2475
GCI count	42269
Hidden GCI Count	13103
Object property axioms	
SubObjectPropertyOf	233
EquivalentObjectProperties	0
InverseObjectProperties	35
DisjointObjectProperties	0
FunctionalObjectProperty	2



Ontology Development Kit (ODK)

- ODK makes it easy to create a new GitHub repo following best practice + automation
- standardized, customizable and automatically executable workflows, and packages all required tooling in a single Docker image



Matentzoglou et al (2022) **Ontology Development Kit: a toolkit for building, maintaining and standardizing biomedical ontologies.**
Database 10.1093/database/baac087

<https://incatools.github.io/ontology-development-kit>

NEW: Ontology Access Kit (OAK). Python library for working with ontologies
<https://incatools.github.io/ontology-access-kit>

Gene Ontology Causal Activity Models: Systems KGs

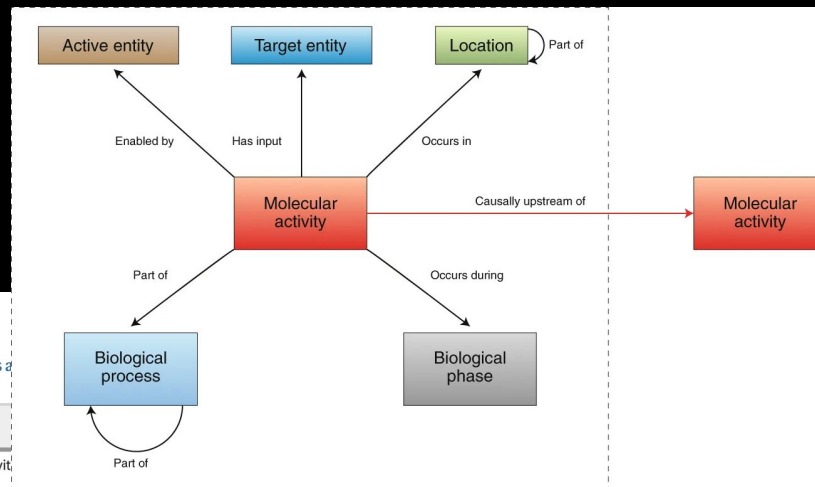
Table 1 GO-CAM elements and ontologies used

From: *Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and system.*

GO-CAM element (Fig. 2)	Ontology or identifier source(s)	Example
Molecular activity	GO molecular function	Ubiquitin-protein transferase activity (GO:0004842)
Biological process	GO biological process	Cellular response to UV (GO:0034644)
	GO cellular component	Nucleus (GO:0005634)
Location	Cell Ontology (CL) ²¹	Retinal cell (CL:0009004)
	Anatomy ontologies, for example, Uberon ²² , <i>C. elegans</i> gross anatomy ²³ or EMAPA ²⁴	Eye (UBERON:0000970)
Active entity	Gene, protein, RNA or complex identifier from a standard source, for example, HGNC for a human gene	NEDD4 (HGNC:7727)
Target entity	Same as active entity or chemical from Chemical Entities of Biological Interest (ChEBI) ²⁵	MAP2K1 (HGNC:6840)
Biological phase	GO biological phase (GO:0044848)	Mitotic G1 phase (GO:0000080)
	Developmental-phase ontology, for example, Mouse Developmental Stages	Theiler stage 02 (MmusDv:0000005)
Relations (arrows in Fig. 2)	Relations Ontology	'Occurs in' (BFO:0000066)

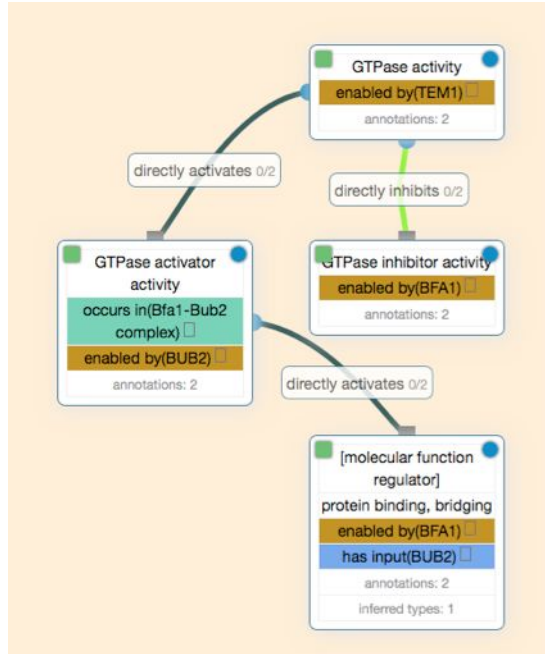
The formalism follows GO annotation practice: gene products (or complexes comprising multiple gene products) have molecular activities (GO molecular function), are active in specific locations (GO cellular component) and act as part of larger biological programs (GO biological process).

Other elements of GO-CAM provide further structured extensions of standard GO annotations. HGNC, HUGO Gene Nomenclature Committee.

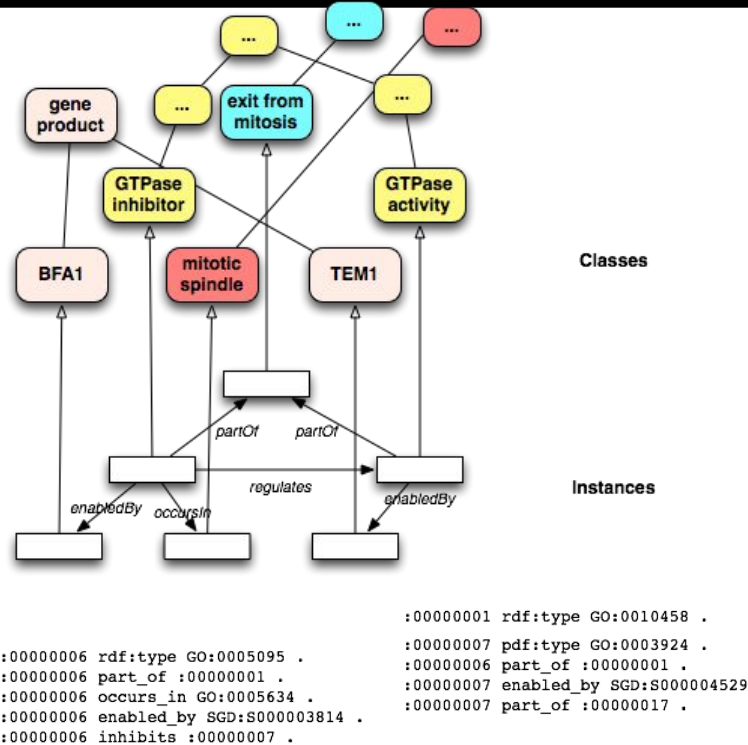


Thomas, PD (2019): *Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and system.* *Nat Gen* PMC7012280

Biologist View



RDF view



<http://rdf.geneontology.org/> - SPARQL endpoint

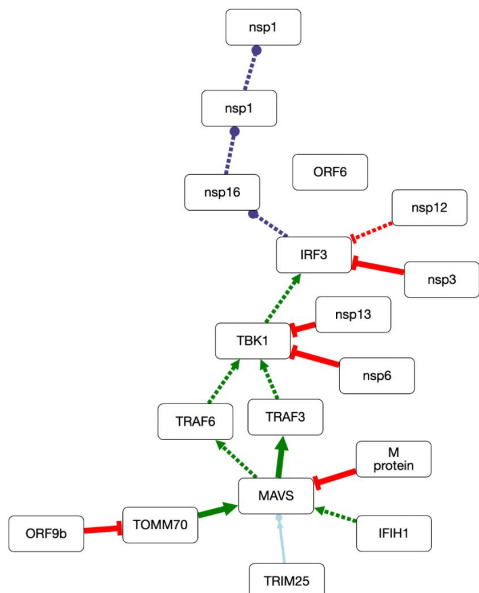
TBK1

Species	<i>Homo sapiens</i>
Symbol	TBK1
Name	TANK binding kinase 1

Pathways

Reactome Pathway (16)	Reactome Reactions (41)	GO-CAMs (12)
-----------------------	-------------------------	--------------

Available GO-CAMs: SARS-COV2/HOST




Processes and Activities

antiviral innate immune response


IFIH1 N/A

pattern recognition receptor
activity



occurs in cytoplasm 

RIGI D.melanogaster

pattern recognition receptor
activity

occurs in cytoplasm 

MAVS D.rerio

signaling adaptor activity 
occurs in mitochondrial
membrane 

- Pathway views on Alliance gene pages

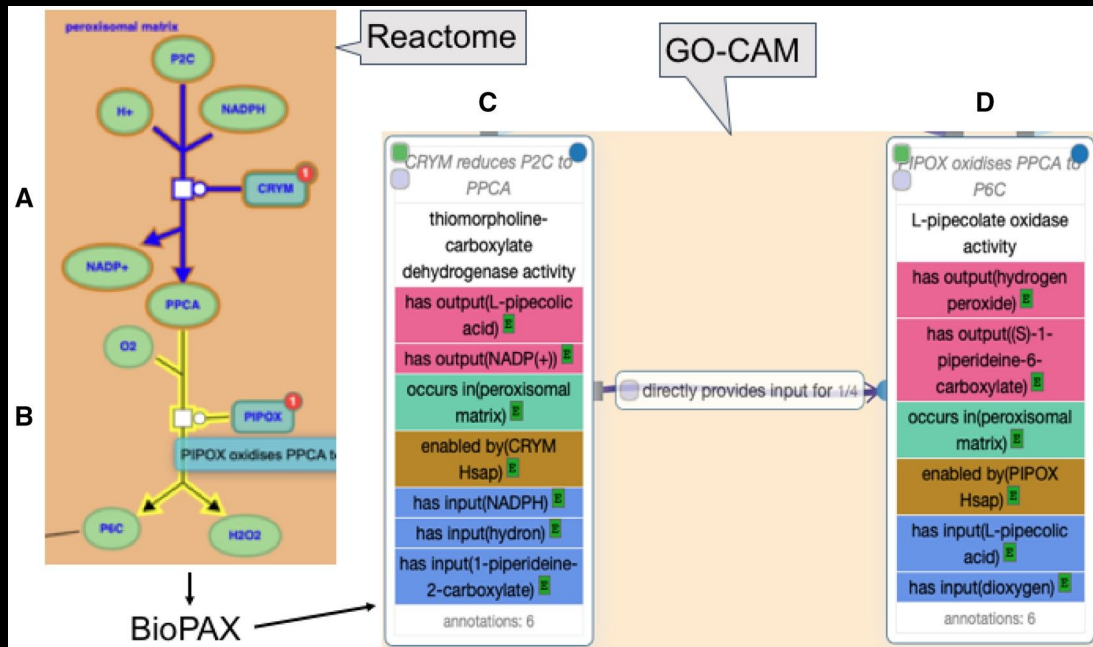
<https://www.alliancegenome.org/gene/HGNC:11584>

Reactome and the Gene Ontology: digital convergence of data resources

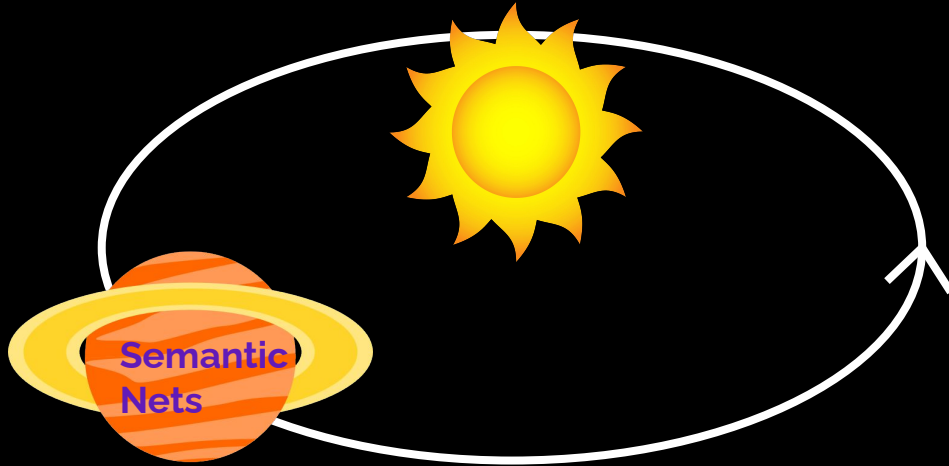
Benjamin M Good, Kimberly Van Auken, David P Hill, Huaiyu Mi, Seth Carbon,
James P Balhoff, Laurent-Philippe Albou, Paul D Thomas, Christopher J Mungall,
Judith A Blake, Peter D'Eustachio ✉

Bioinformatics, btab325, <https://doi.org/10.1093/bioinformatics/btab325>

Published: 08 May 2021 Article history ▼



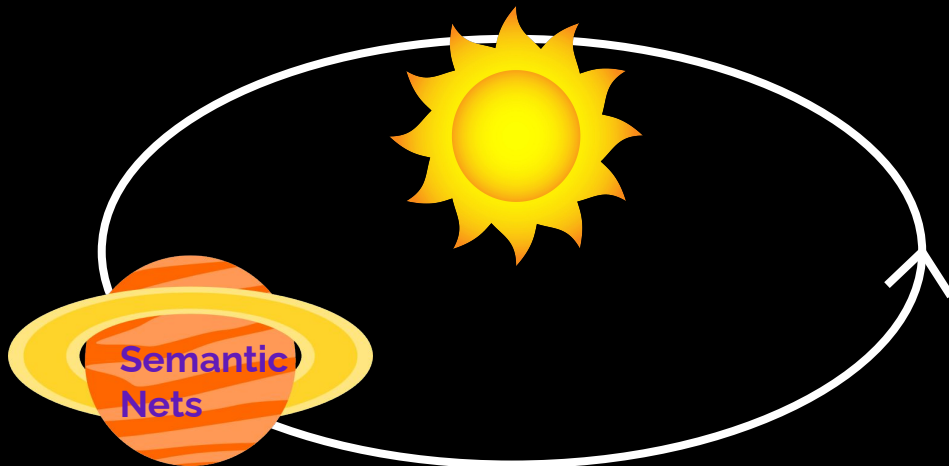
Previous iterations of planet KG



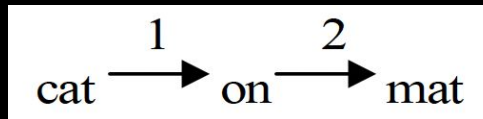
*Knowledge Representation with graphs
didn't start with the semantic web and
ontologies...*

Previous iterations: 1950s KGs

Richens RH, "Preprogramming for mechanical translation" *Mechanical Translation* 3 (1), July 1956, 20–25

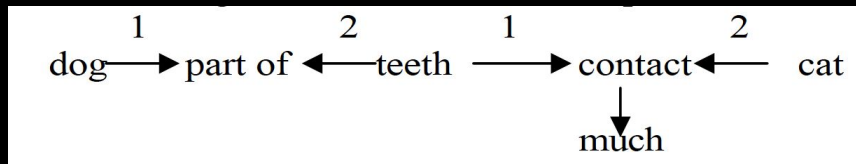


“The cat is on the mat”



Differences:

- Emphasis on language
- Applications: Machine Translation

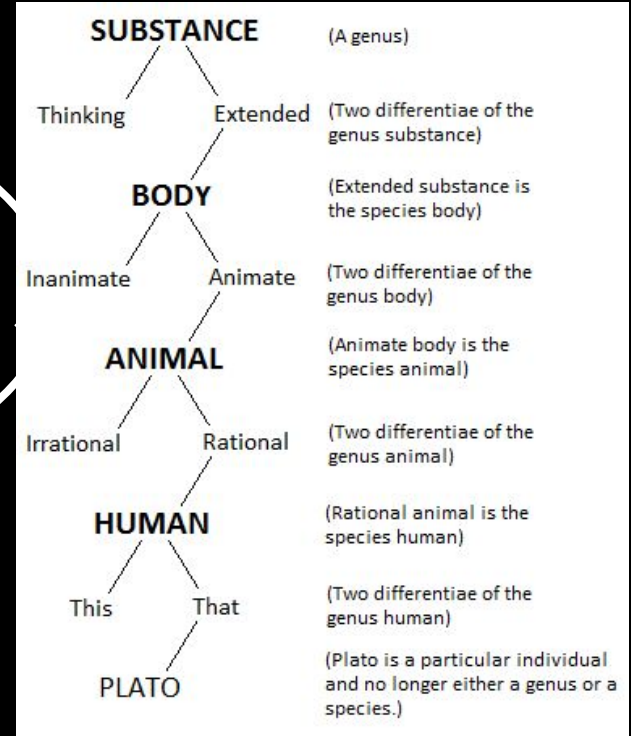
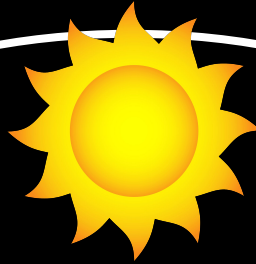


“The cat is bitten by the dog”

Previous iterations: KGs in the 3rd century CE



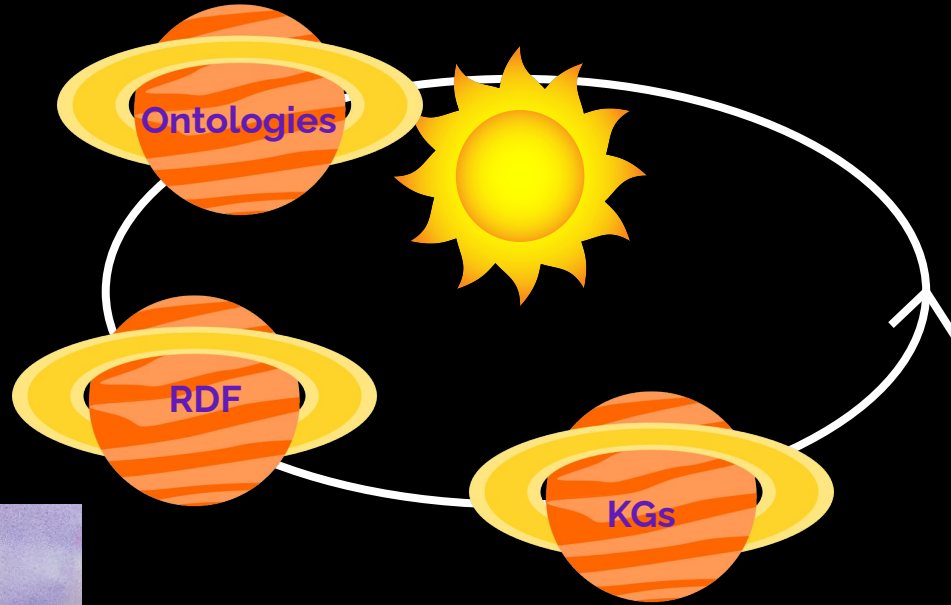
Porphirian
Trees



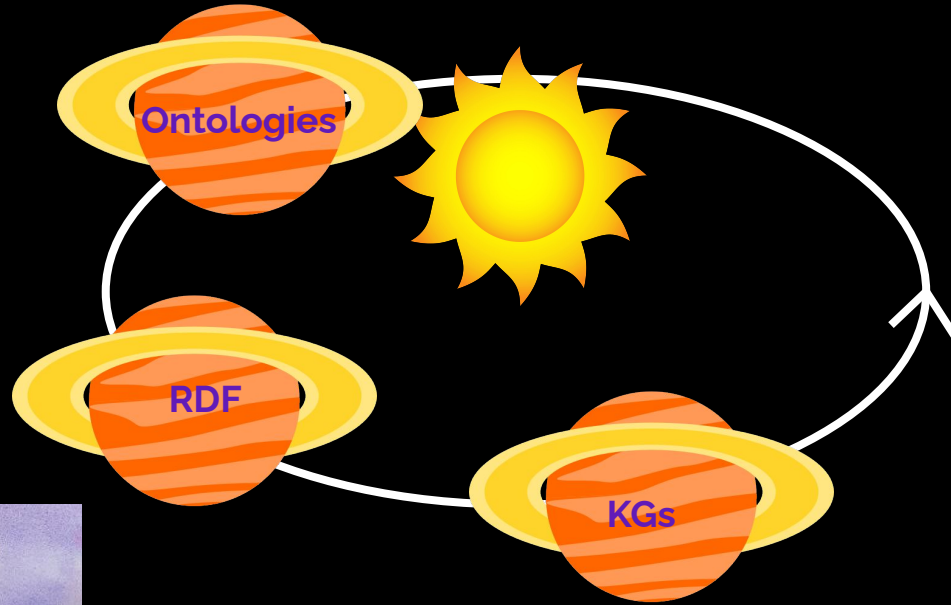
Differences:

- Philosophical inquiry >> Computation
- Theology >> science or empiricism

Curse of the eternal return



Curse of the eternal return



Embrace CURIEs

- URI are rarely used as identifiers outside RDF frameworks
 - URIs are awkward to work with outside RDF tooling
- Solution: CURIEs
 - Prefixed IDs have been standard in bio-databases before RDF
 - GO:0005634
 - UniProtKB:P12345
 - Key: *always make the prefixmap explicit*

<https://bioregistry.io/>
<https://github.com/linkml/prefixmaps/>

OLS / Gene Ontology GO / GO:0005634 Copy

nucleus

 http://purl.obolibrary.org/obo/GO_0005634  Copy

Hoyt, C. T., et al. (2022) [The Unifying the identification of biomedical entities with the Bioregistry](#). *Nature Scientific Data*, s41597-022-01807-3

Can we have semantics in our KGs?

The story:

RDF and OWL provide semantics for the web

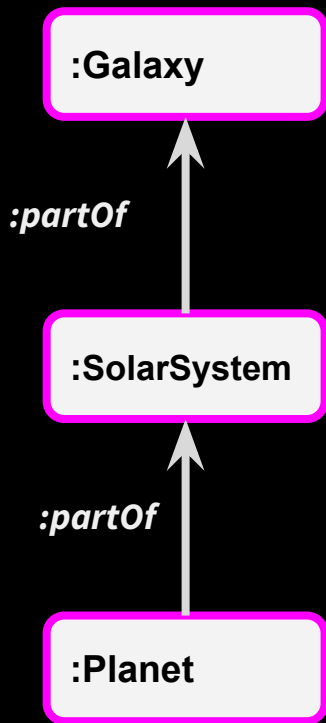
The reality:

OWL is great for terminology maintenance

OWL layering in RDF is problematic and ignored in non-RDF KGs

OWL is a poor fit for *contingent knowledge* common in the life sciences

Example: semantics of parthood



- Doesn't capture quantification

Example: semantics of parthood in OWL

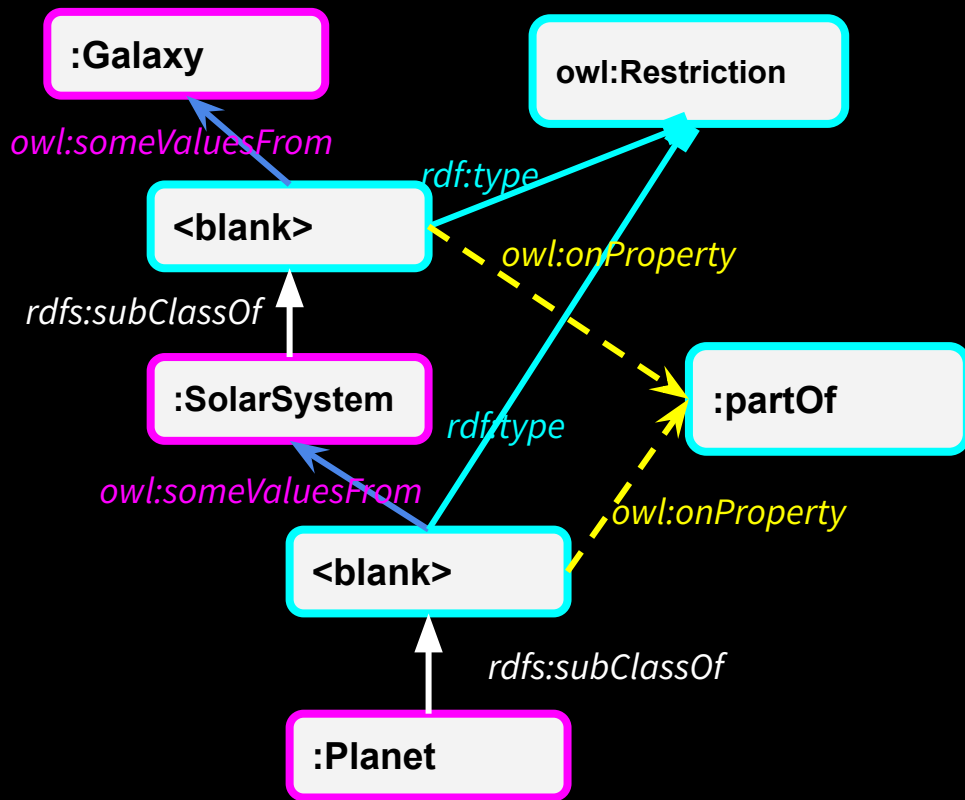
Class: Planet

SubClassOf: *partOf* some SolarSystem

Class: SolarSystem

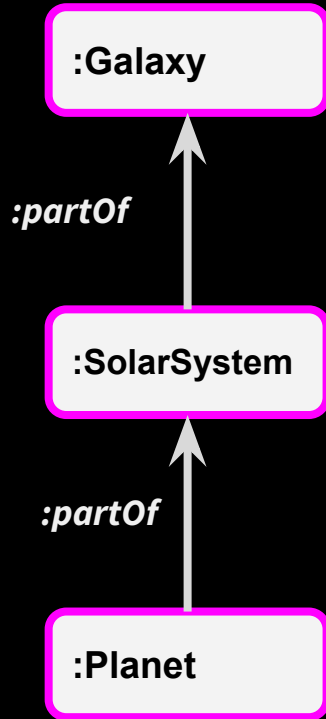
SubClassOf: *partOf* some Galaxy

RDF/OWL Ontology Rendering



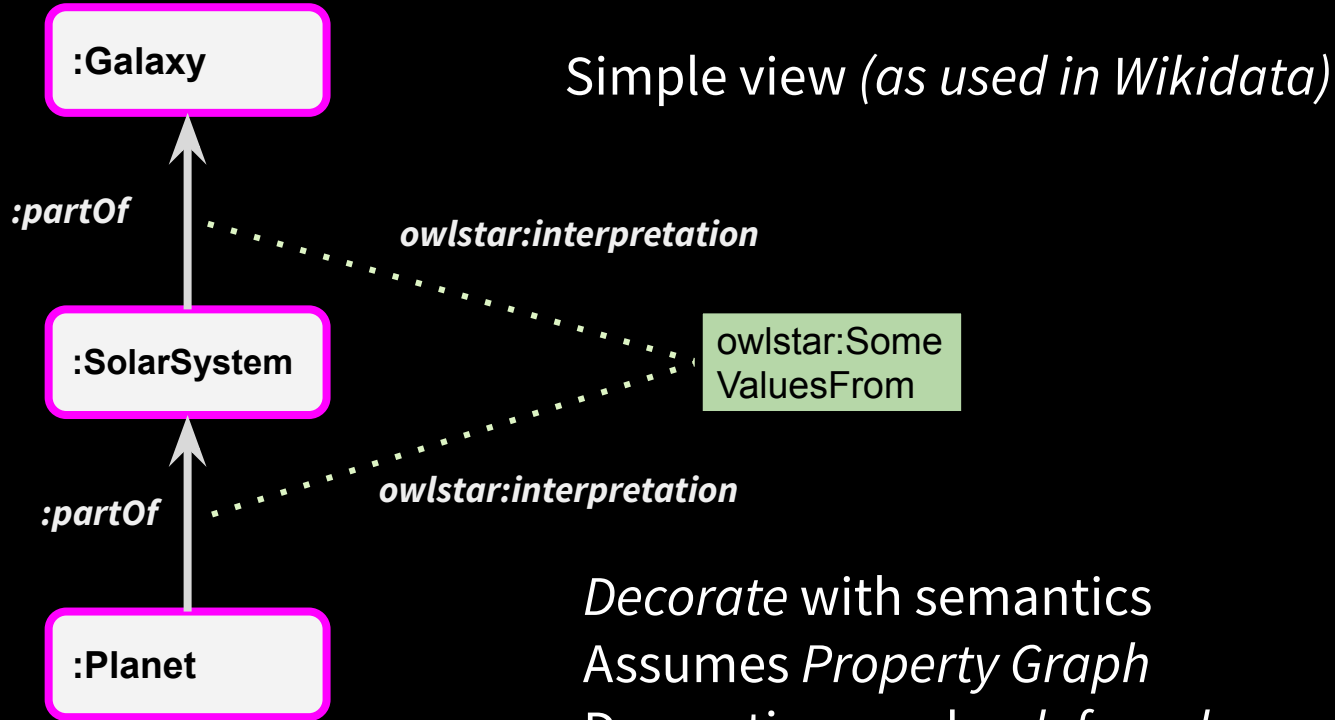
- Complex monstrosity
- Misaligned with mental model
- Confuses developers
- Not amenable to transitive/graph querying
 - E.g. *what kinds of things are found in galaxy?*

OWLStar Property Graph conceptual model

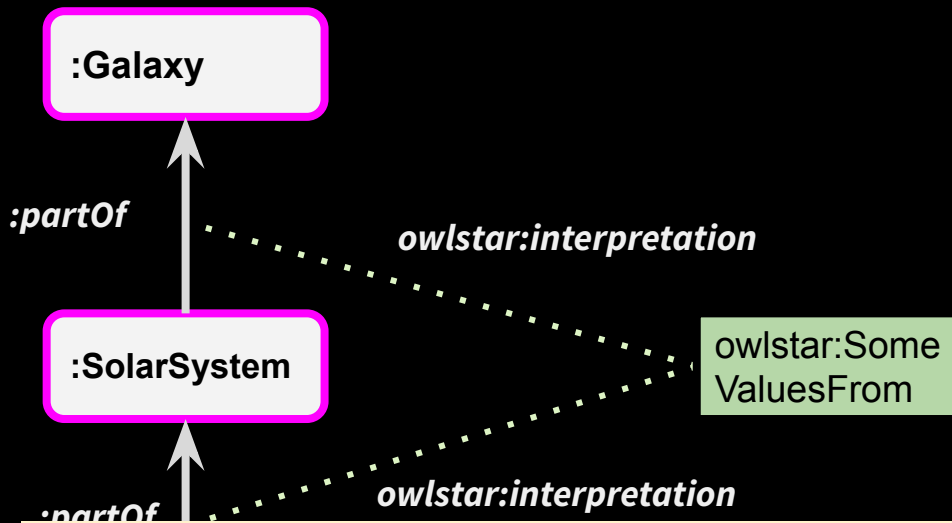


Simple view (*as used in Wikidata*)

OWLStar Property Graph conceptual model

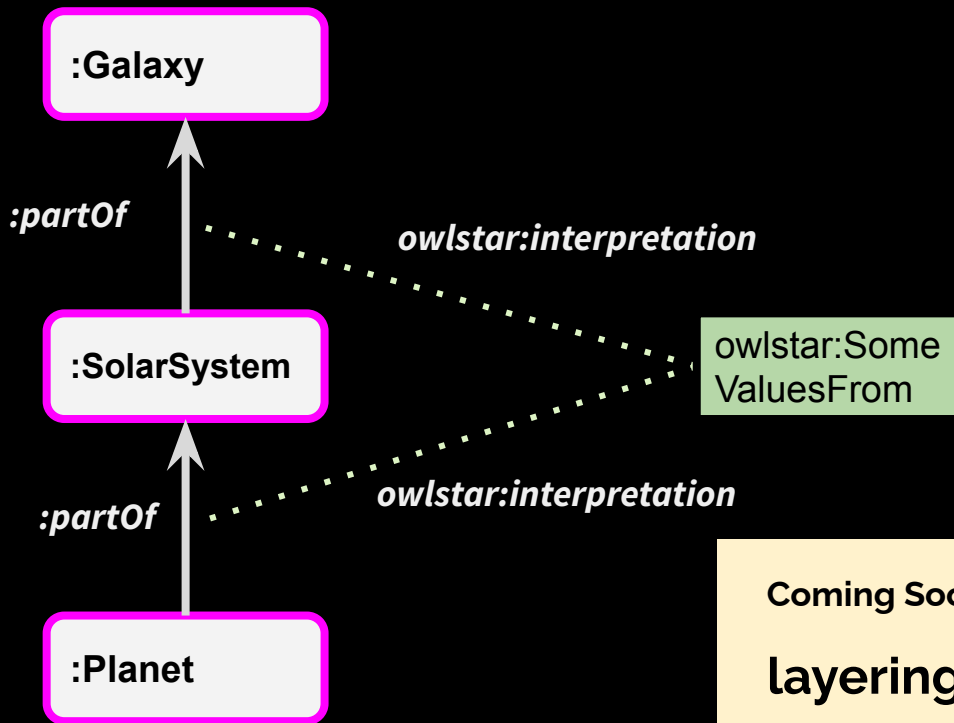


OWLStar Property Graph in RDFStar



```
<<:Planet :part-of :SolarSystem>> owlstar:interpretation owlstar:AllSomeInterpretation .  
<<:SolarSystem :part-of :Galaxy>> owlstar:interpretation owlstar:AllSomeInterpretation .
```

OWLStar Property Graph in OneGraph



Coming Soon: **OneGraph**
layering

Semantics of contingent knowledge in KGs

FOL/model theory is the basis of RDF and OWL

... yet is a poor fit for much biological knowledge

See:

Schulz S: **Strengths and limitations of formal ontologies in the biomedical domain** PMC2904529

Rector A: **On beyond Gruber: "Ontologies" in today's biomedical information systems and the limits of OWL** PMID34384571

Schulz S: **"lmo-2 interacts with Elf-2" On the Meaning of Common Statements in Biomedical Literature** KRMED 2006



:lmo-2

:elf-2

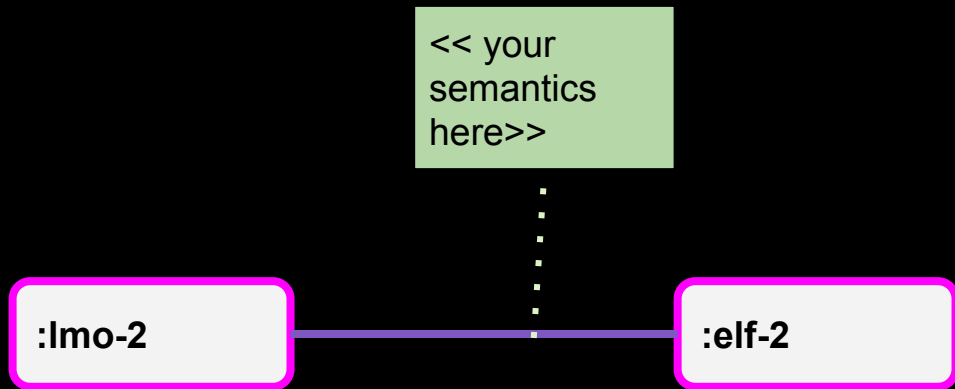
A new old logic layered on Property Graphs

Multimodal semantics:

Use whatever the use case demands

Defeasible reasoning, bayesian probabilities, ...

Or simply *defer*



BioLink Model: hierarchical categories for KGs

Biolink Model

Last uploaded: November 26, 2022

Summary

Classes

Properties

Notes

Mappings

Widgets

- entity
 - association
 - named thing
 - activity
 - administrative entity
 - attribute
 - biological entity
 - chemical entity
 - chemical mixture
 - complex molecular mixture
 - food
 - molecular mixture
 - processed material
 - environmental food contaminant
 - food additive
 - molecular entity**
 - nucleic acid entity
 - coding sequence
 - exon
 - transcript
 - small molecule
 - clinical entity
 - clinical intervention
 - clinical trial
 - device
 - event
 - information content entity
 - common data element
 - confidence level
 - dataset
 - dataset distribution
 - dataset summary
 - dataset version
 - evidence type
 - information resource
 - publication
 - study
 - study result
 - study variable
 - organism taxon
 - phenomenon
 - physical entity
 - planetary entity
 - procedure
 - treatment

Details

Visualization

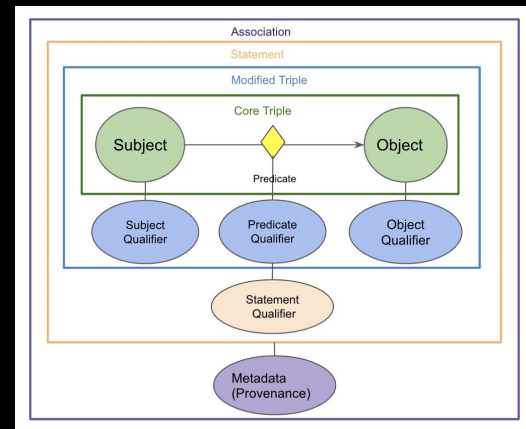
Notes (0)

Class Mappings (127)



Preferred Name	molecular entity
Synonyms	
Definitions	A molecular entity is a chemical entity composed of individual or covalently bonded atoms.
ID	https://w3id.org/biolink/vocab/MolecularEntity
definition	A molecular entity is a chemical entity composed of individual or covalently bonded atoms.
name	molecular entity
narrowMatch	http://purl.bioontology.org/ontology/STY/T088 https://bioschemas.org/MolecularEntity http://purl.obolibrary.org/obo/CHEBI_23367 http://purl.bioontology.org/ontology/STY/T085
prefixIRI	biolink:MolecularEntity
prefLabel	molecular entity
subClassOf	https://w3id.org/biolink/vocab/ChemicalEntity

Unni D et al (2022). Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science. *Clin. Tran. Sci*

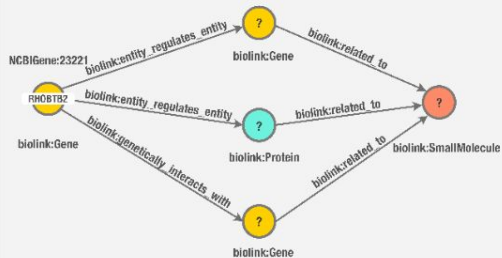


<https://w3id.org/biolink>

Standards in NCATS Biomedical Translator

Query

What chemicals or drugs might be used to treat neurological disorders such as epilepsy that are associated with genomic variants of RHOBTB2?

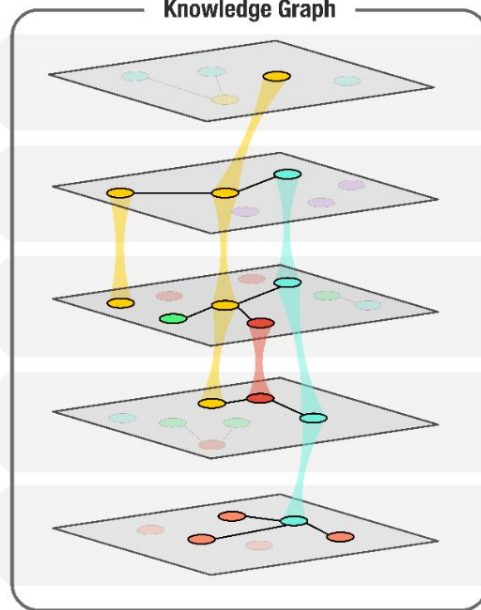


Result

CHEMBL.COMPOUND:CHEMBL3989516 Fostamatinib disodium

CHEMBL.COMPOUND:CHEMBL1789941 Ruxolitinib

Knowledge Graph



Biolink Model

Knowledge Provider



Knowledge Source

Gene Ontology
UniProt

Reactome

HP
MONDO
GO
UBERON

Monarch


ChEBI
ChEMBL

Standards in NCATS Biomedical Translator

What drugs may treat ▾

Q Addison Disease

Search

[View this disease on Monarch Initiative](#) 

[Need Help?](#)

Filters

Entity Search

Evidence

☒ Minimum Number of Evidence

1

10

99

ATC Classification 

Filter on organ or system where drug's therapeutic effect occurs.

☐ Alimentary Tract And Metabol... (20)

☐ Blood And Blood Forming Or... (1)

☐ Cardiovascular System (17)

☐ Dermatologicals (10)

Results

Showing 1-10 of 38 (502) Results

Minimum Evidence: 10 ×

NAME	EVIDENCE	SCORE ▲
Metyrapone 1 Path that may treat Congenital Adrenal Hyperplasia Due To 11-beta-hydroxylase Deficiency	View All Evidence (30)	93.4 ▲

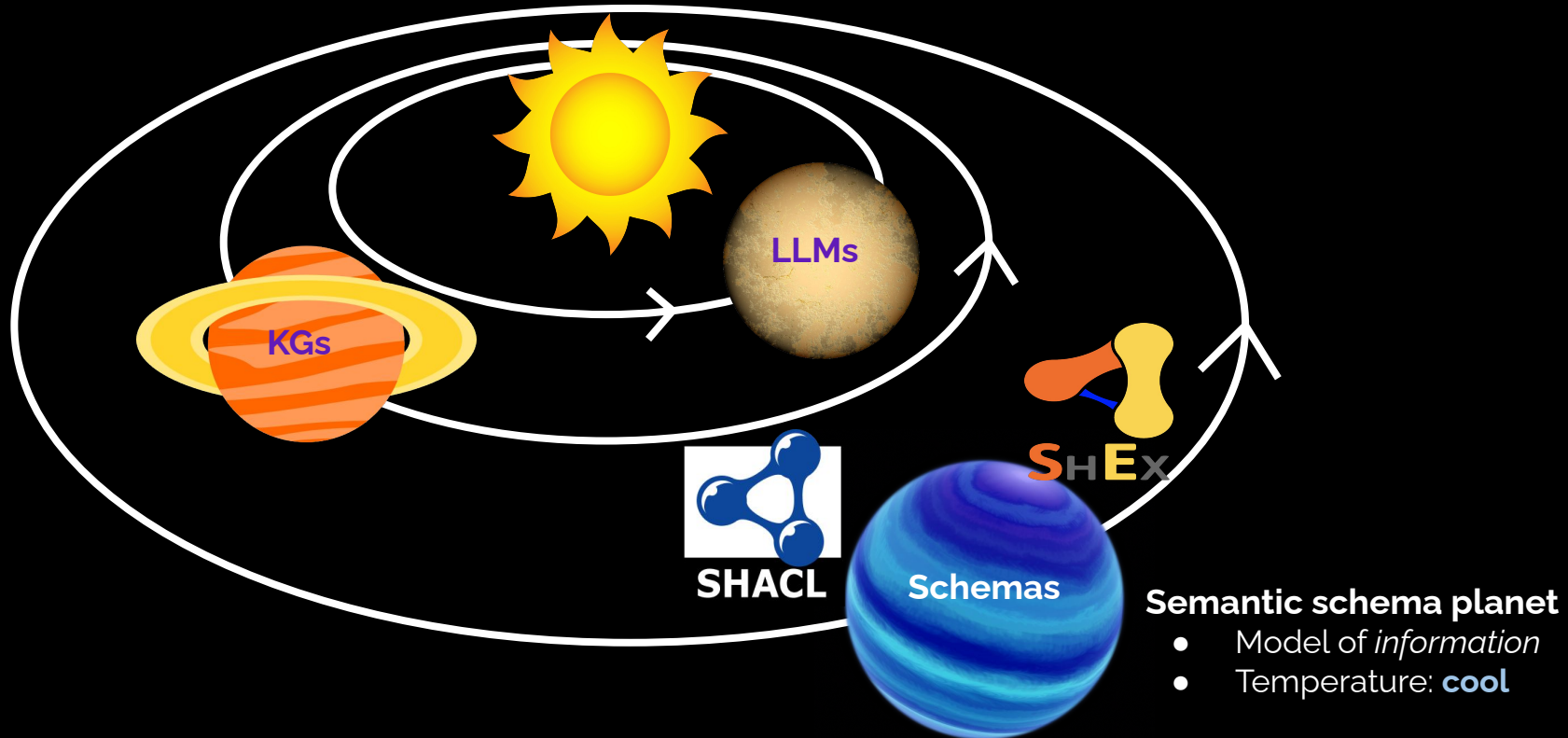
Paths

Click on any entity to view a definition (if available), or click on any relationship to view evidence that supports it.

ENTITY	RELATIONSHIP	ENTITY	RELATIONSHIP	ENTITY	RELATIONSHIP	TARGET
 Metyrapone	causes decre...	 Cyp11b1	causes	 Congenital Adr...		

[Was this helpful?](#) [Send Feedback](#)

Planet Semantic Schema



Previous iteration: OWL misused as schema language

<https://twitter.com/aaranged/status/1485670670134497281>

"I think it is fair to say that it is much easier to live without OWL than without SHACL" > Syntax, semantics, and the great OWL hoax / Jan Voskuil bit.ly/3uOG5qx

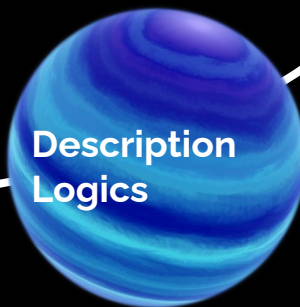
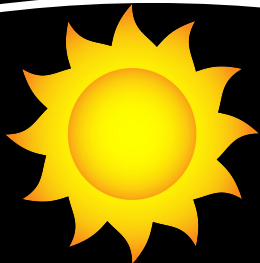
Why I Don't Use OWL Anymore

Ontology Modeling, Blog

The idea for this blog came from a question asked by a customer. They said:

"When do you use OWL and when do you use SHACL? Isn't OWL about semantics and SHACL about data validation? And what about RDFS?"

My answer was that I no longer used RDFS/OWL. I now use SHACL for everything. The only parts of RDFS I use are subclasses. I stopped using RDFS/OWL back in 2017 when [SHACL](#) became the official W3C standard.



**Description
Logics**

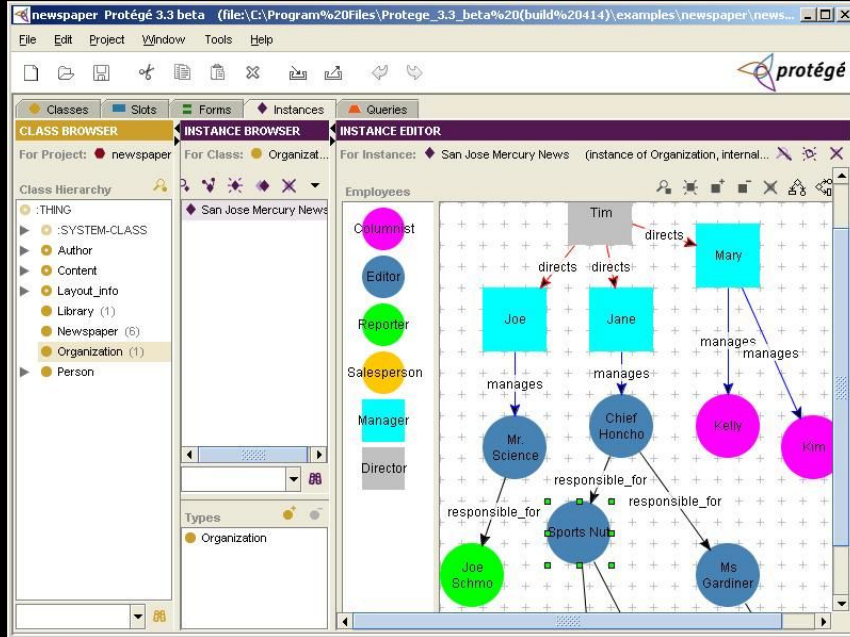
Semantic schema planet

- Model of *information*
- Temperature: **cool**

Lessons:

- OWL is for modeling terminological knowledge
- OWL is not a language to describe data

Previous iteration: Frames 1990s-2000s




- OWL succeeded DAML+OIL
- DAML normally frame-like

Frames

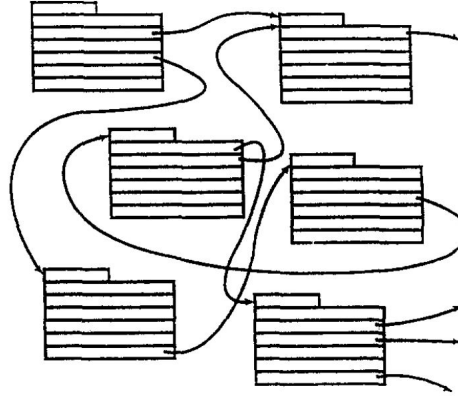
Semantic schema planet

- Model of *information*
- Temperature: **cool**

Previous revolution: Frames 1970s-1990s



(FIDO)	
Slot	Values
INSTANCE-OF:	value: (DOG, PET)
Name:	value: "Fido"
Color:	value: (BROWN)
Father:	value: (BOWSER)
Mother:	value: (WEENIE)
Owner:	value: (MR.-FITZCUBBINS)
Cost:	value: \$12.95
Has-as-ears:	value: (LEFT-EAR, RIGHT-EAR)
Number-of-ears:	value: 2



A Framework for Representing Knowledge: Minsky, M. *MIT-AI Laboratory Memo 306*, June, 1974

Lehmann, F. Semantic Networks. *Computers Math. Applic.* 23(2-5), 1-50, (1992)
[https://doi.org/10.1016/0898-1221\(92\)90135-5](https://doi.org/10.1016/0898-1221(92)90135-5)

Frames

Semantic schema planet

- Model of *information*
- Temperature: **cool**

We need semantic standards!

>1800 Databases

>1500 Standards

>1k Ontologies

~13.5m terms

Most data standards are underspecified

Most standards in fairsharing don't have an associated schema or validator

If you're lucky:

XML Schema (older)

JSON Schema (newer)

Bindings to ontology terms are frequently unclear

Most data standards are underspecified

Most standards in Fairsharing don't have an associated schema or validator

If you're lucky:

- XML Schema (older)

- JSON Schema (newer)

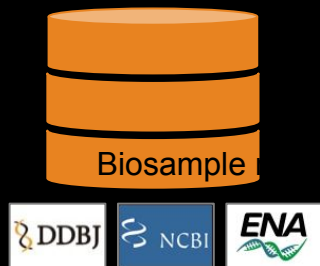
Bindings to ontology terms are frequently unclear

What about ShEx/SHACL standards?

- Most data in the biosciences not shared as RDF

- Some (e.g BioPAX) are really XML standards

Lack of adherence to schemas limits meta-analyses



Metagenome standards analysis findings:

- Standards are underspecified
- **Not *machine actionable***
- Public data has poor conformance

Data too noisy for environmental meta-analyses

Depth

MIxS specifies this should be {number} {unit}

Some example values that do not conform:

- N40.1164_W88.2543
- 25 santimeters
- 0 – 20 cm
- 3.149
- 30-60cm replicate6
- 1800, 1800
- 30ft
- 5m, 32m, 70m, 110m, 200m, 320m, 1000m
- Surface soil from deep water
- 0 m water depth
- Metamorph4 (19dpf) biological replicate 3

Actual data!

LinkML: Linked Data Modeling Language

*Create data models in simple YAML files,
optionally annotated using ontologies*



<https://linkml.io>

<https://github.com/linkml/linkml>

```
id: https://example.org/linkml/hello-world
title: Really basic LinkML model
name: hello-world
version: 0.0.1

prefixes:
  linkml: https://w3id.org/linkml/
  sdo: https://schema.org/
  ex: https://example.org/linkml/hello-world/

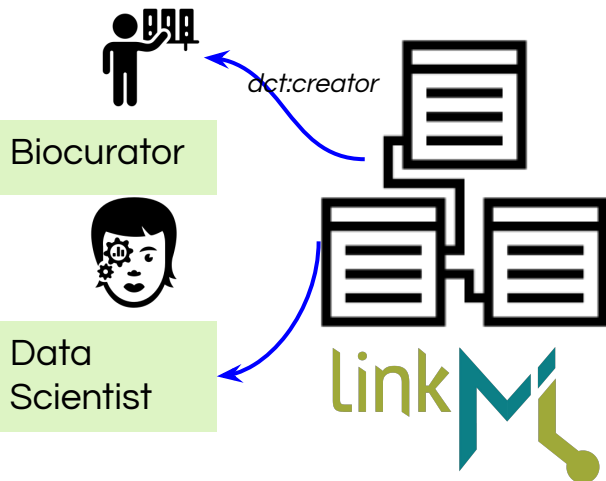
default_prefix: ex
default_curi_maps:
  - semweb_context

imports:
  - linkml:types

classes:
  Person:
    description: Minimal information about a person
    class_uri: sdo:Person
    attributes:
      id:
        identifier: true
        slot_uri: sdo:taxID
      first name:
        required: true
        slot_uri: sdo:givenName
        multivalued: true
      last name:
        required: true
        slot_uri: sdo:familyName
      knows:
        range: Person
        multivalued: true
        slot_uri: foaf:knows
```


LinkML: Easy for all

Create data models in simple YAML files,
optionally annotated using ontologies



<https://linkml.io>

<https://github.com/linkml/linkml>

```
id: https://example.org/linkml/hello-world
title: Really basic LinkML model
name: hello-world
version: 0.0.1
```

```
prefixes:
  linkml: https://w3id.org/linkml/
  sdo: https://schema.org/
  ex: https://example.org/linkml/hello-world/
```

```
default_prefix: ex
default_curi_maps:
  - semweb_context
```

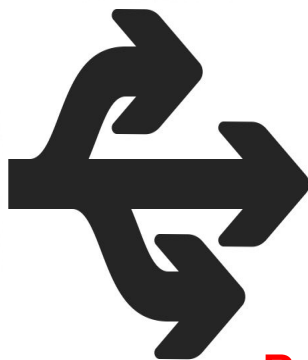
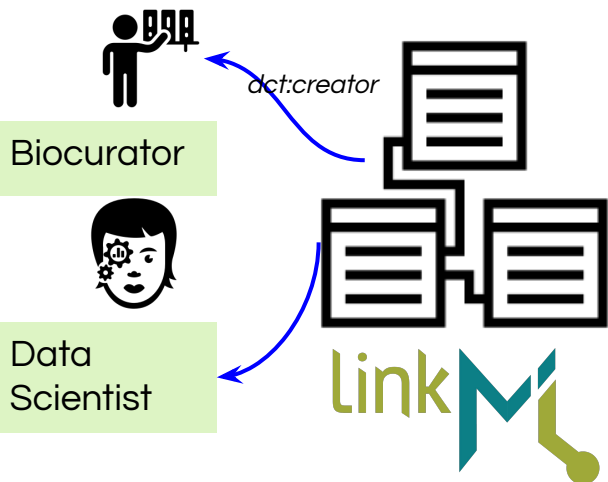
```
imports:
  - linkml:types
```

```
classes:
  Person:
    description: Minimal information about a person
    class_uri: sdo:Person
    attributes:
      id:
        identifier: true
        slot_uri: sdo:taxID
      first name:
        required: true
        slot_uri: sdo:givenName
        multivalued: true
      last name:
        required: true
        slot_uri: sdo:familyName
    knows:
      range: Person
      multivalued: true
      slot_uri: foaf:knows
```

LinkML: Polyglot modeling

Create data models in simple YAML files,
optionally annotated using ontologies

Compile to other
frameworks



JSON-Schema

Python
Dataclasses

SQL DDL
TSVs

“Traditional”
Applications and
Infrastructure

{JSON}

<https://linkml.io>

<https://github.com/linkml/linkml>

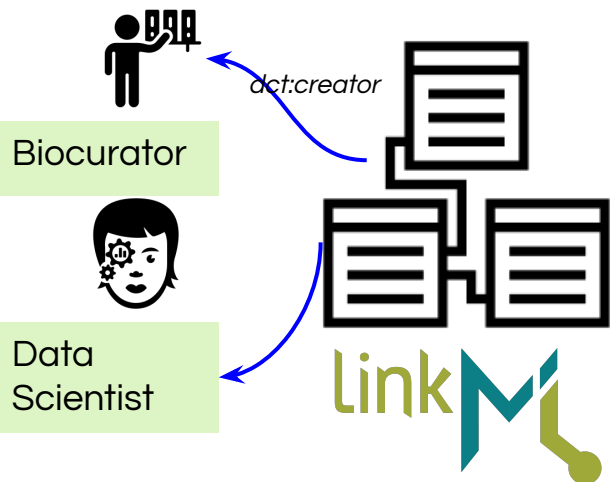


LinkML: Bridges semantic frameworks

Create data models in simple YAML files,
optionally annotated using ontologies

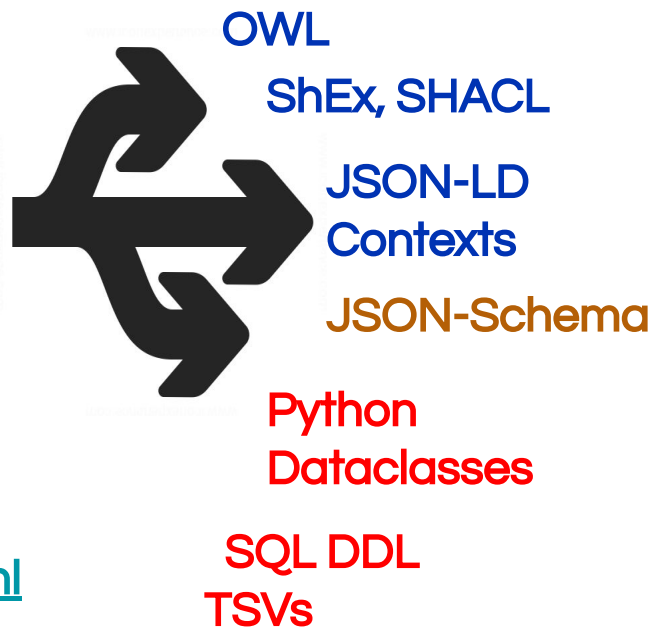
Compile to other
frameworks

Choose the right tools
for the job, no lock in



<https://linkml.io>

<https://github.com/linkml/linkml>



Semantic Web
Applications
And
Infrastructure



JSON-LD { }

“Traditional”
Applications and
Infrastructure

{JSON}



Annotate schemas with ontologies

```
id: https://example.org/linkml/hello-world
title: Really basic LinkML model
name: hello-world
license: https://creativecommons.org/publicdomain/zero/1.0/
version: 0.0.1
```

```
prefixes:
  linkml: https://w3id.org/linkml/
  sdo: https://schema.org/
  ex: https://example.org/linkml/hello-world/
```

```
default_prefix: ex
default_curi_maps:
  - semweb_context
```

```
imports:
  - linkml:types
```

```
classes:
  Person:
    description: Minimal information about a person
    class_uri: sdo:Person
    attributes:
      id:
        identifier: true
        slot_uri: sdo:taxID
      first_name:
        required: true
        slot_uri: sdo:givenName
        multivalued: true
      last_name:
        required: true
        slot_uri: sdo:familyName
    knows:
      range: Person
      multivalued: true
      slot_uri: foaf:knows
```

Export data to RDF and JSON-LD

>> Make the meaning of your schema more explicit

>> Data integration hooks



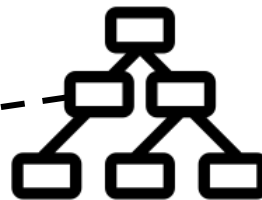
Easy ontology support via value sets

prefixes:

```
COB: http://purl.obolibrary.org/obo/COB\_
BFO: http://purl.obolibrary.org/obo/BFO\_
RO: http://purl.obolibrary.org/obo/RO\_
CHEBI: http://purl.obolibrary.org/obo/CHEBI\_
CHEMINF: http://semanticscience.org/resource/CHEMINF\_
SIO: http://semanticscience.org/resource/SIO\_
PUBCHEM.ELEMENT: https://pubchem.ncbi.nlm.nih.gov/element/
LANL.ELEMENT: https://periodic.lanl.gov/
```

enums:

```
nanostructure_morphology_enum:
  permissible_values:
    nanotube:
      meaning: CHEBI:50796
    nanoparticle:
      meaning: CHEBI:50803
    nanorod:
      meaning: CHEBI:50805
    nanotubosome:
      meaning: CHEBI:50806
    quantum dot:
      meaning: CHEBI:50853
    nanofibre:
      meaning: CHEBI:52518
    nanocrystal:
      meaning: CHEBI:52529
    nanoribbon:
      meaning: CHEBI:52530
    nanosheet:
      meaning: CHEBI:52531
    nanowire:
      meaning: CHEBI:52593
```



ChEBI

LinkM

Easy ontology support via value sets

☐ eukaryotic cell ☐ All t
 ☐ animal cell
 ☐ neural cell
 ☐ **neuron**
 ☐ CNS neuron (sensu Nematoda and Protostomia)
 ☐ CNS neuron (sensu Vertebrata)
 ☐ GABAergic neuron
 ☐ GABAergic interneuron
 ☐ GABAergic interplexiform cell
 ☐ Kolmer-Agduhr neuron
 ☐ Lugaro cell
 ☐ Martinotti neuron
 ☐ L4 sst GABAergic cortical interneuron (Mmus)
 ☐ T Martinotti neuron
 ☐ L5 T-Martinotti sst GABAergic cortical interneuron (Mmus)
 ☐ fan Martinotti neuron
 ☐ basket cell
 ☐ Ammon's horn basket cell
 ☐ cerebellum basket cell
 ☐ dentate gyrus of hippocampal formation basket cell
 ☐ neocortex basket cell
 ☐ cerebellar Golgi cell
 ☐ cerebral cortex GABAergic interneuron
 ☐ Ammon's horn basket cell
 ☐ L5/6 cck cortical GABAergic interneuron (Mmus)
 ☐ alpha7 GABAergic cortical interneuron (Mmus)
 ☐ caudal ganglionic eminence derived GABAergic cortical interneuron
 ☐ dentate gyrus of hippocampal formation basket cell
 ☐ lamp5 GABAergic cortical interneuron
 ☐ medial ganglionic eminence derived GABAergic cortical interneuron

enums:

NeuronTypeEnum:

reachable_from:

source_ontology: obo:cl

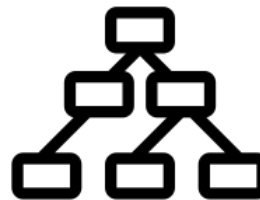
source_nodes:

- CL:0000540 ## neuron

include_self: false

relationship_types:

- rdfs:subClassOf



Model Serializations

LinkML Model (YAML)

Classes:

Person:

Slots:

age_in_years:
range: integer

ShEx, SHACL

```
<Person> CLOSED {  
  ( $<Person_tes> (  
    <age_in_years> @linkml:Integer ?  
  ;  
}
```

SQLDDL

```
CREATE TABLE "Person" (  
  age_in_years INTEGER  
)
```

Java

```
package org.person.model;  
  
import java.util.List;  
import lombok.*;  
  
@Data  
@EqualsAndHashCode(callSuper=false)  
public class Person { private Integer age_in_years; }
```

Python

```
@dataclass  
class Person(NamedThing):  
    _inherited_slots: ClassVar[List[str]] = []  
  
    age_in_years: Optional[int] = None
```

JSONSchema

```
{  
  "Person": {  
    "Properties": {  
      "age_in_years": {  
        "type": "integer"  
      }  
    }  
  }  
}
```



Instance data as JSON/YAML, RDF, or TSV

```
from examples.basic import Person
from linkml.dumpers import json_dumper, rdf_dumper

sam = Person("1172438", first_name=["Samual", "J"], last_name="Snooter")
ann = Person("17a3923", first_name="Jill", last_name="Jones", knows=[sam.id])

print(json_dumper.dumps(ann))
print(yaml_dumper.dumps(ann))
print(rdf_dumper.dumps(ann, contexts=" ../examples/jsonld/basic.context.jsonld"))
```

JSON output

```
{
  "id": "17a3923",
  "first_name": [
    "Jill"
  ],
  "last_name": "Jones",
  "knows": [
    "1172438"
  ],
  "@type": "Person"
}
```

YAML output

```
id: 17a3923
first_name:
- Jill
last_name: Jones
knows:
- '1172438'
```

RDF output

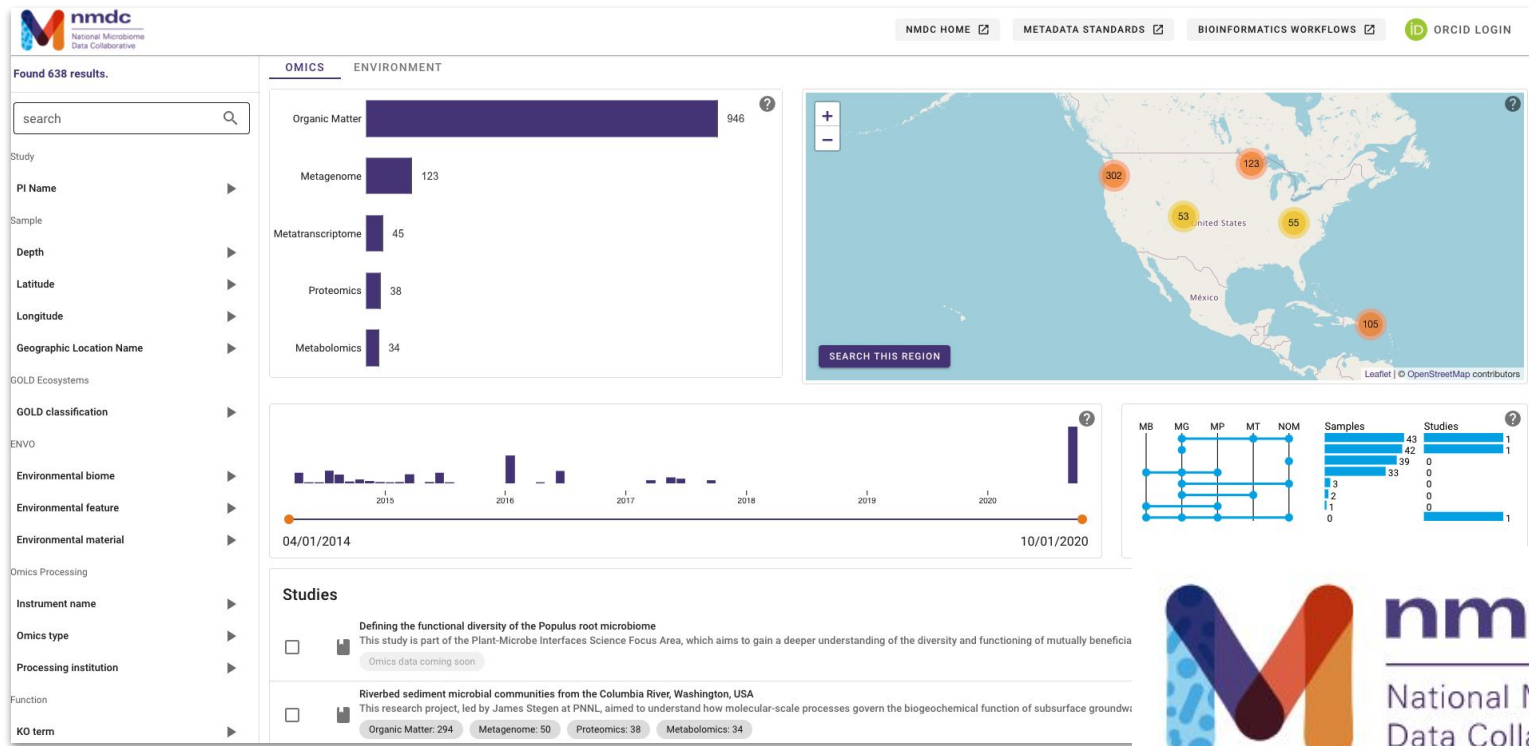
```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix sdo: <https://schema.org/> .

<https://example.org/linkml/hello-world/17a3923> a sdo:Person ;
  foaf:knows <https://example.org/linkml/hello-world/1172438> ;
  sdo:familyName "Jones" ;
  sdo:givenName "Jill" .
```


LinkML adoption



Environmental microbiome data

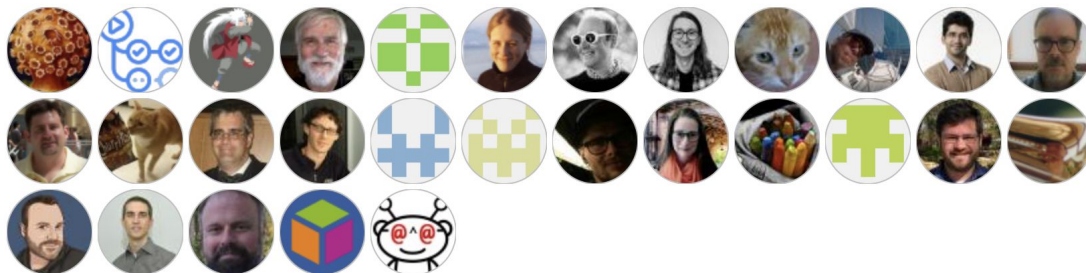


Open, active, helpful, inclusive community

Contributor Covenant Code of Conduct

Our Pledge

In the interest of fostering an open and welcoming environment, we as contributors and maintainers pledge to making participation in our project and our community a harassment-free experience for everyone, regardless of age, body size, disability, ethnicity, sex characteristics, gender identity and expression, level of experience, education, socio-economic status, nationality, personal appearance, race, religion, or sexual identity and orientation.



<https://github.com/linkml/linkml/graphs/contributors>

August 23, 2022 – September 23, 2022

Period: 1 month

Overview

24 Active pull requests

57 Active issues

24

Merged pull requests

0

Open pull requests

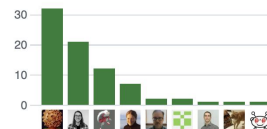
25

Closed issues

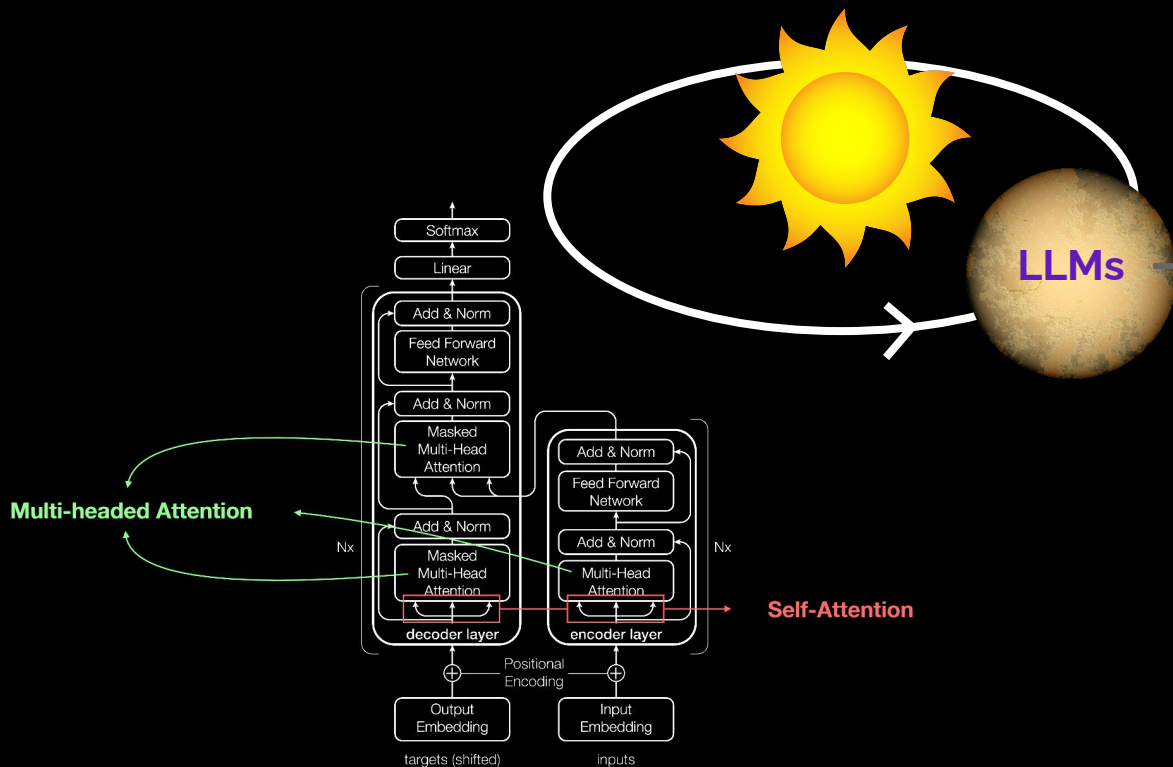
32

New issues

Excluding merges, **9 authors** have pushed **37 commits** to main and **76 commits** to all branches. On main, **154 files** have changed and there have been **29,371 additions** and **35,395 deletions**.



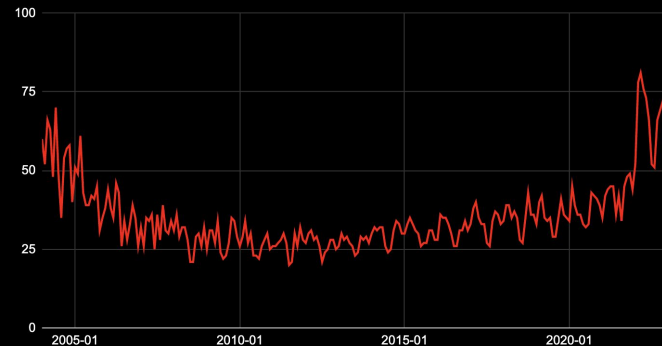
Planet Large Language Model



Large Language Model planet

- Model of *words*
- Temperature: **ultra hot**

language model: (Worldwide) vs. Month



LLMs lack semantic understanding



Gary Marcus
@GaryMarcus

1980s: AI companies spend vast amounts of \$ hiring humans to write rules to patch brittle symbolic systems.

2020s: AI companies spend vast amounts of \$ hiring humans to patch brittle neural systems.



Alexandr Wang @alexandr_wang

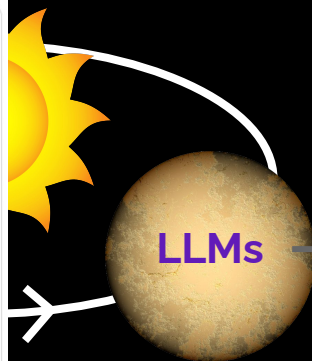
we're starting to see top companies spend the same amount on **RLHF** and compute in training ChatGPT-like LLMs

for example, OpenAI hired >1000 devs to **RLHF** their code models

crazy—but soon companies will start spending \$ hundreds of Ms or \$ billions on **RLHF**, just as w/compute

1:54 AM · Feb 2, 2023

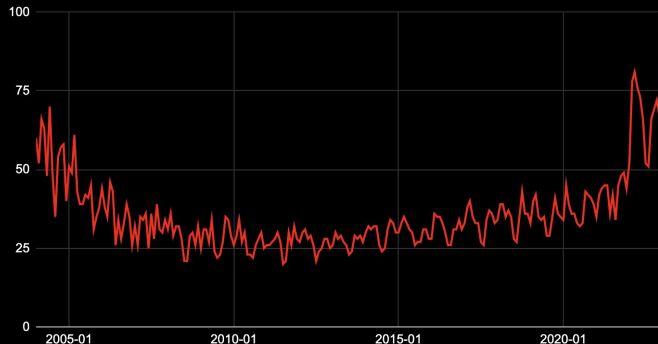
102 Likes 27 Retweets



Large Language Model planet

- Model of *words*
- Temperature: **ultra hot**

language model: (Worldwide) vs. Month



Multi-

tion

targets (shifted)

inputs

OntoGPT: using LLMs with semantic structures

Ontology layer onto OpenAI GPT-3 API

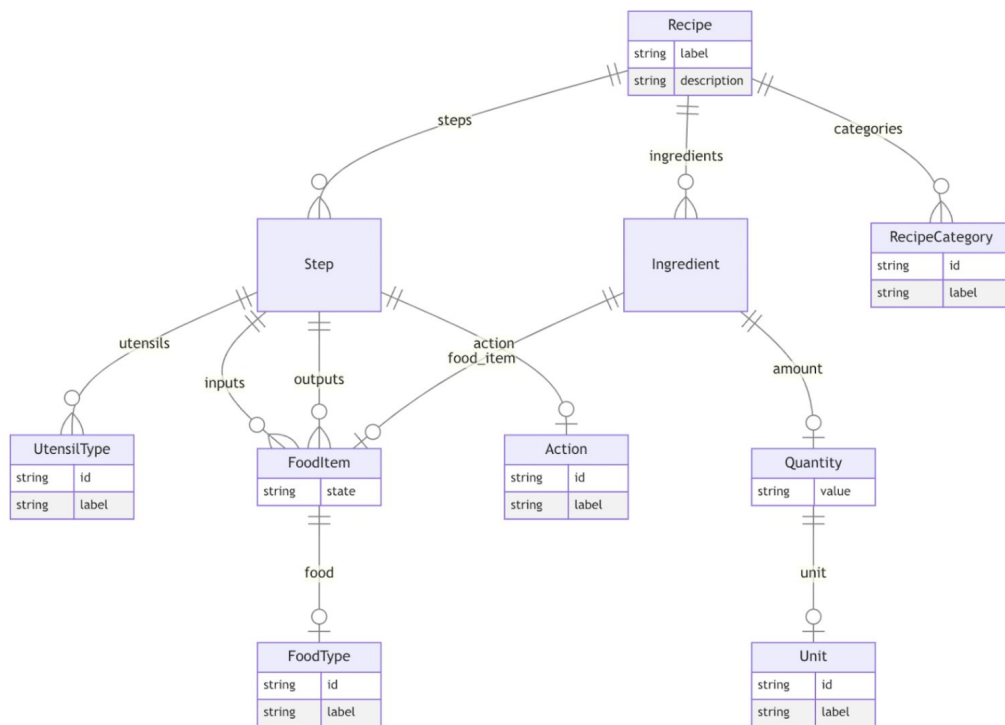
- Integrates prompt engineering, semantic data models, ontology annotation, and OWL translation

Components:

- SPIRES: Stochastic Parrot Interrogation and Recursive Extraction of Semantics
- HALO: Hallucinating Latent Ontologies
- Authoring KGs with copilot

<https://github.com/monarch-initiative/ontogpt>

SPIRES: Stochastic Parrot Interrogation and Recursive Extraction of Semantics



Input: LinkML Schema

- Nested/Compositional
- Annotated with ontology metadata (e.g. FOODON)

Schema can be nested; e.g

- Recipe has Step
- Step has utensils, actions, input, output
- Input has foodtype, quantity

SPIRES: Stochastic Parrot Interrogation and Recursive Extraction of Semantics

On medium heat melt the butter and sauté the onion and bell peppers.
Add the hamburger meat and cook until meat is well done.
Add the tomato sauce, salt, pepper and garlic powder.
Salt, pepper and garlic powder can be adjusted to your own tastes.
Cook noodles as directed.
Mix the sauce and noodles if you like, I keep them separated

INGREDIENTS

UNITS: US

1 small onion (chopped)
1 bell pepper (chopped)
2 tablespoons garlic powder
3 tablespoons butter
1 teaspoon salt
1 teaspoon pepper
2 (15 ounce) cans tomato sauce
1 (16 ounce) box spaghetti noodles
1 - 1 1/2 lb hamburger meat.

Utens
string
string

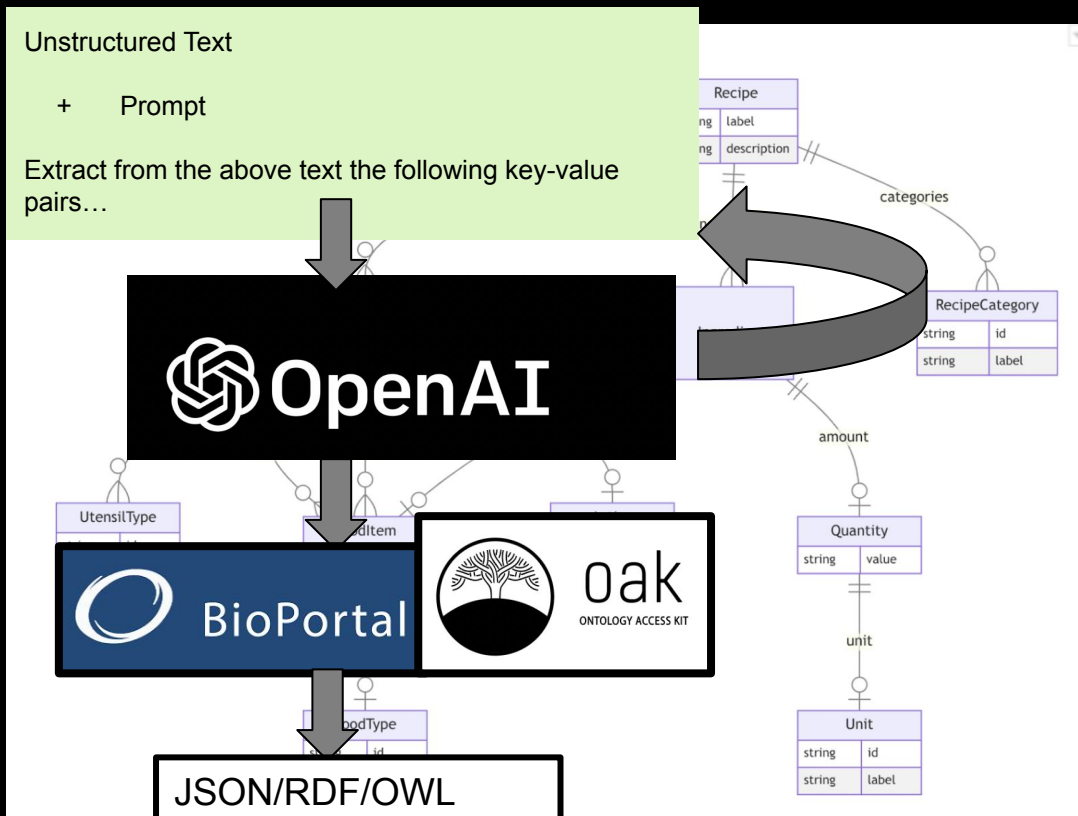
Input: LinkML Schema

- Nested/Compositional
- Annotated with ontology metadata (e.g. FOODON)

Input: Unstructured text

- E.g recipe, abstract

SPIRES: Stochastic Parrot Interrogation and Recursive Extraction of Semantics



Input: LinkML Schema

- Nested/Compositional
- Annotated with ontology metadata (e.g. FOODON)

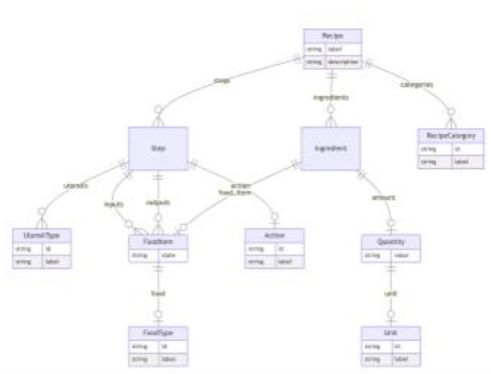
Input: Unstructured text

- E.g recipe, abstract

Process:

- Recursively query GPT3 via prompts
- Annotate using OAK and Bioportal
- (Map results to OWL)

SPIRES: Stochastic Parrot Interrogation and Recursive Extraction of Semantics



label: Simple Spaghetti
description: A tomato sauce spaghetti dish with hamburger meat and vegetables.
category:
- dbpedia:Main_course
- dbpedia:Italian_cuisine
ingredients:
- food_item: FOODON:03301704 ## onion (whole, raw)
quantity: 1
- food_item: FOODON:00003485 ## sweet red bell pepper (whole)
quantity: 1
- food_item: FOODON:03301844 ## garlic powder
quantity: 2
unit: "[tbs_us]"
- food_item: FOODON:03310351 ## butter
quantity: 3
unit: "[tbs_us]"
- food_item: FOODON:00001649 ## black or white pepper product
quantity: 1
unit: "[tps_us]"
...

```
steps:
- action: chop
  inputs:
  - FOODON:03301704 ## onion (whole, raw)
  outputs:
  - _:ChoppedOnion
- action: chop
  inputs:
  - FOODON:00003485 ## sweet red bell pepper (whole)
  outputs:
  - _:ChoppedBellPepper
- action: saute
  utensils:
  - cooking pan
  inputs:
  - FOODON:03310351 ## butter
  - _:ChoppedOnion
  - _:ChoppedBellPepper
- action: add
  inputs:
  - FOODON:00001282 ## ground beef food product
  outputs:
  - _:CookedGroundBeef
- action: add
  inputs:
  - FOODON:03301217 ## tomato sauce
  - FOODON:00002221 ## salt product
  - FOODON:00001649 ## black or white pepper product
  - FOODON:03301844 ## garlic powder
  outputs:
  - FOODON:03304014 ## spaghetti sauce with meat
- action: cook
  inputs:
  - FOODON:03306312 ## pasta (dried)
  outputs:
  - _:CookedPasta
- action: mix;
  inputs:
  - FOODON:03304014 ## spaghetti sauce with meat
  - _:CookedPasta
  outputs:
  - _:MixedSpaghetti
```

Input: LinkML Schema

- Nested/Compositional
- Annotated with ontology metadata (e.g. FOODON)

Input: Unstructured text

- E.g recipe, abstract

Process:

- Recursively query GPT3 via prompts
- Annotate using OAK and Bioportal
- (Map results to OWL)

Output:

- Nested JSON/RDF
- (OWL TBox model)

Standard schemas in life sciences, 1990s

LSR Technology Adoption Roadmap



The **Life Sciences Research (LSR)** group is a consortium of people representing pharmaceutical companies, academic institutions, software vendors and hardware vendors from all over the world who are working together within the **Object Management Group (OMG)** to improve communication and interoperability among computational resources in life sciences research

Mission:

To improve the quality and utility of software and information systems used in Life Sciences Research through use of the Model Driven Architecture (MDA), OMG technologies such as **UML** and **CORBA**, and MDA-supported technologies such as **XML** and **EJB**

Standard schemas in life sciences, 1990s

Why should this standardisation effort succeed when others have failed?

It is based on a strong and very widely accepted technology (OMA and CORBA) and will follow a process (the OMG technology adoption process) that has already proven itself in helping to standardize object interfaces in a variety of areas.

Conclusions

- Embrace change
- Don't be wedded to any one technology stack
 - Why can't BigTable, Neo4j, etc have semantics?
- We need to scale up semantics to cover ALL the data
- This requires an open, inclusive approach

Acknowledgments

- OBO and Ontologies
 - OBO Operations Group, James Overton, **Nico Matentzogl**, Bjoern Peters
 - Gene Ontology Consortium, Ben Good, Seth Carbon, Laurent-Philippe Albou, Tremayne Mushayaham, Dustin Ebert, Peter d'Eustachio, David Hill, Pascale Gaudet, Paul D Thomas
- Biolink, Translator, and KG-Hub:
 - NCATS Translator Data Modeling Group, Chris Bizon, Nomi Harris, Mike Bada, Matt Brush, **Deepak Unni**, **Sierra Moxon**
 - **KG-Hub group**, Peter Robison, Melissa Haendel, David Osumi-Sutherland, Kevin Shafer, Marcin Joachimiak, Seth Carbon, Tim Putman, Luca Cappalletti, **Justin Reese**
- LinkML:
 - LinkML community and developer group, **Harold Solbrig**, Sierra Moxon, Mark Miller, Sujay Patil
- OntoGPT:
 - **Harry Caufield**, Harshad Hegde