KNowledge Acquisition and Representation Methodology

Hande Küçük McGinty

hande@ksu.edu

April 30, 2023

Who am I?

Assistant Prof. at the Department of Computer Science at K-State

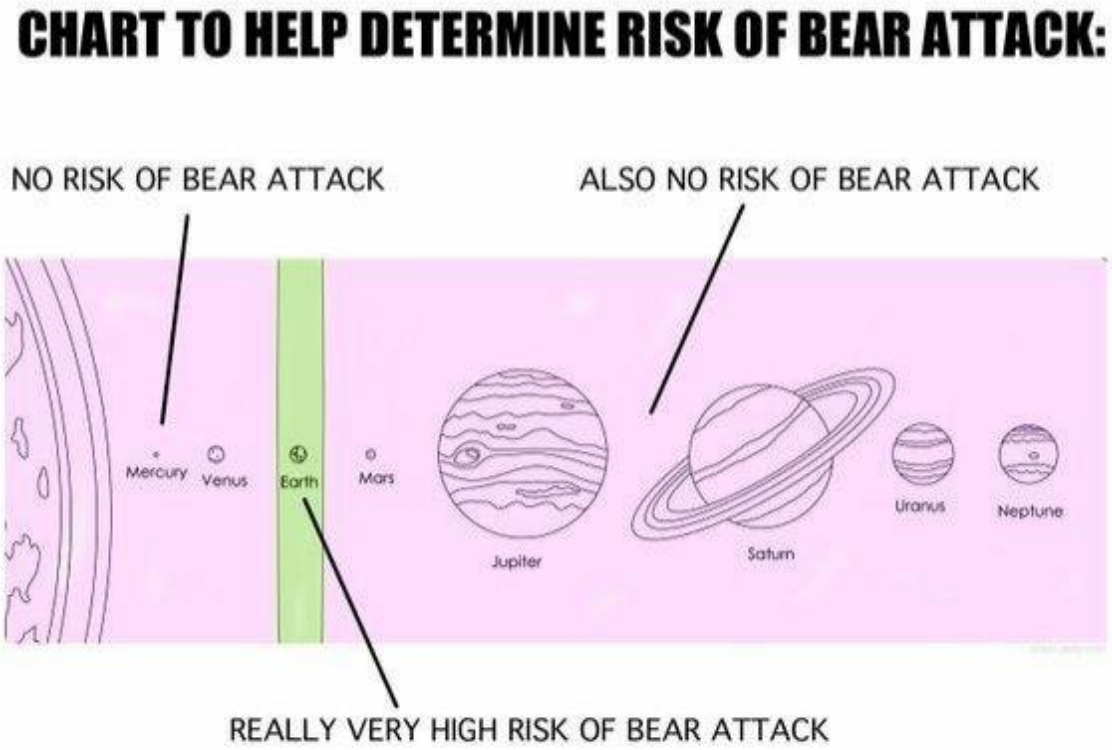
OHIO UNIVERSITY

Agricultural Research Service

CDD VAULT®
Complexity Simplified

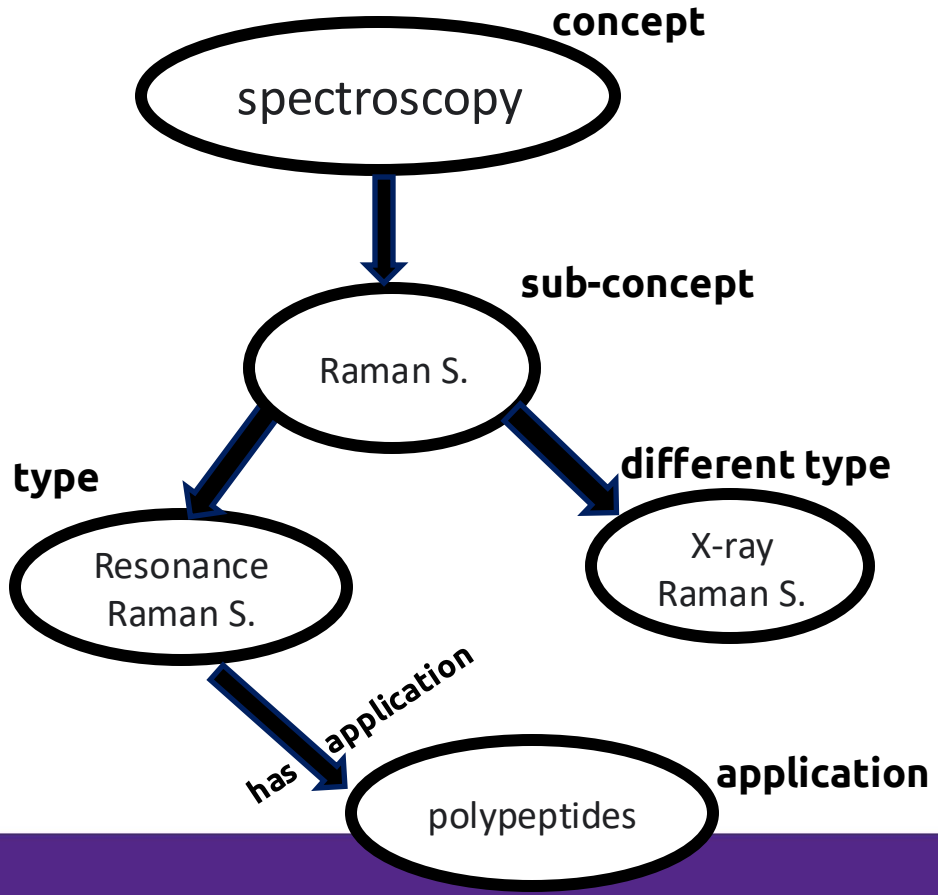
UNIVERSITY OF MIAMI

When statistical analysis doesn't tell us much..



What is an ontology?

- An ontology defines a set of concepts and relationships in a subject area to model it using formal logic.
- Ontologies show their concepts' properties and the relationships among them.



Why build ontologies / knowledge graphs?

- Ontologies aim to :
  - limit complexity,
  - align domain experts' vocabulary within themselves in addition to machines',
  - organize data into information, knowledge and improve problem solving within that domain.

Three Nationwide Projects

NIH LINCS PROGRAM

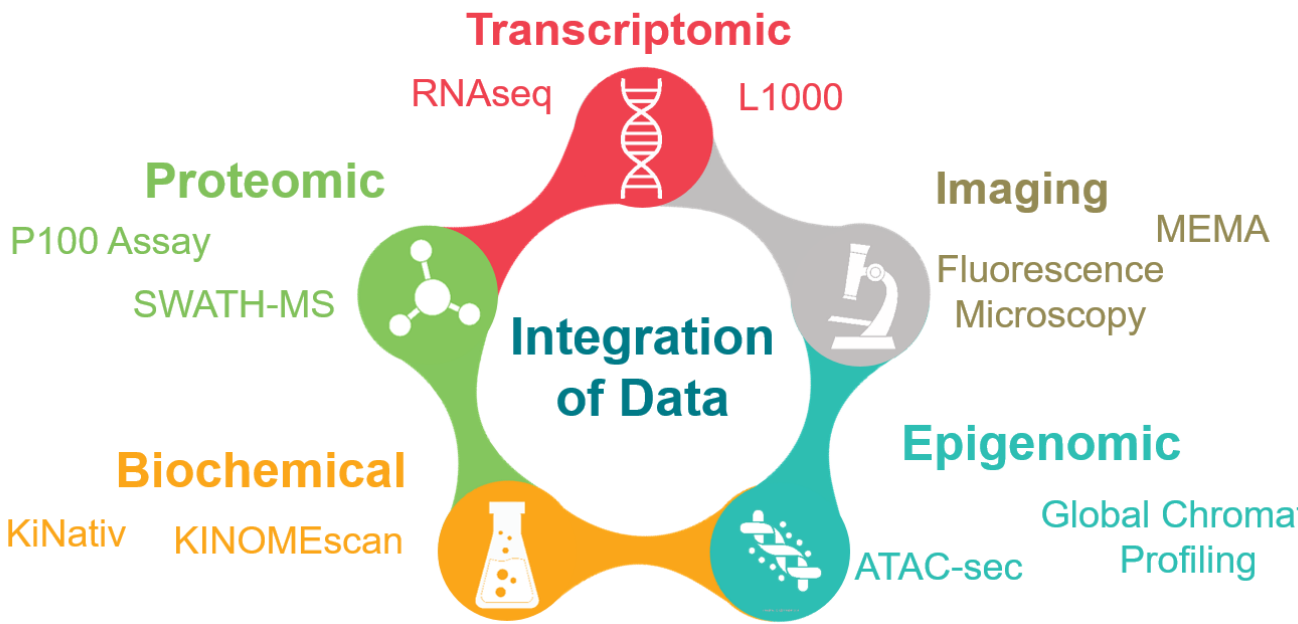
NIH Big Data to Knowledge(BD2K)

NIH grants

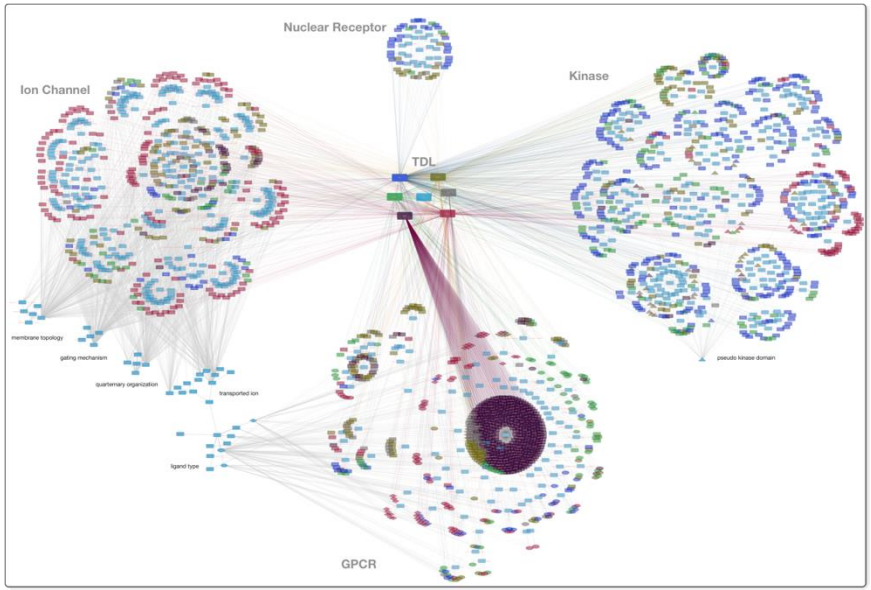
- 5U01HL111561-02 (LINCS Information FramEwork (LIFE) to Integrate and Analyze Diverse Data Sets)
- U54CA189205 (Illuminating the Druggable Genome Knowledge Management Center, IDG-KMC)
- U54HL127624 (BD2K LINCS Data Coordination and Integration Center, DCIC).

Bioassays in LINCS and Big Data To Knowledge Proj.

LINCS generates diverse multidimensional signatures



More data that should work concordantly..

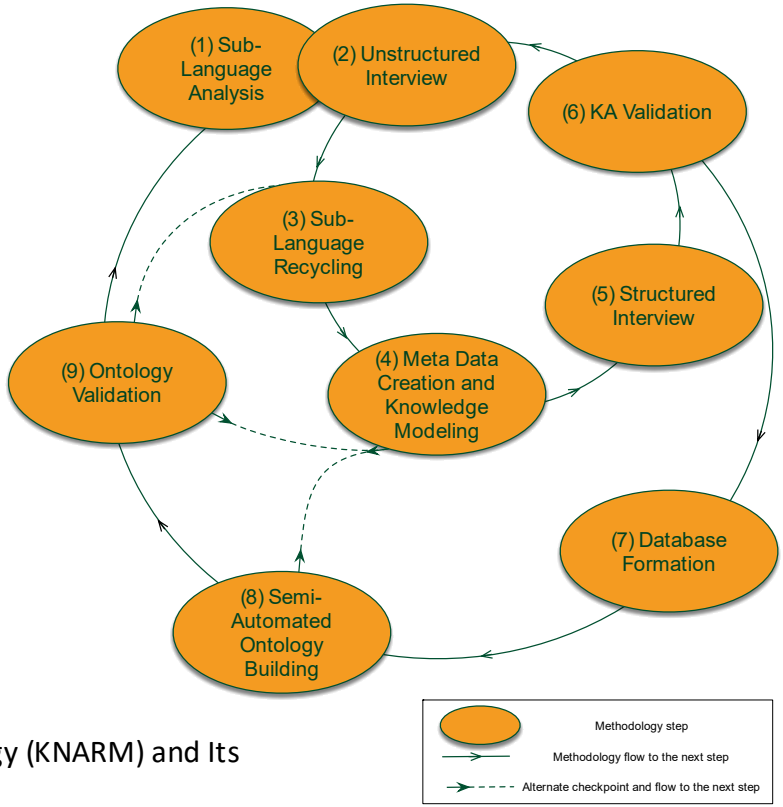


Family	Tbio	Tchem	Tclin	Tdark
Kinase	197	354	50	33
GPCR	131	127	96	52
Ion channels	106	84	128	25
NR	7	23	18	–

IDG Project Drug Target Proteins Target Development Levels

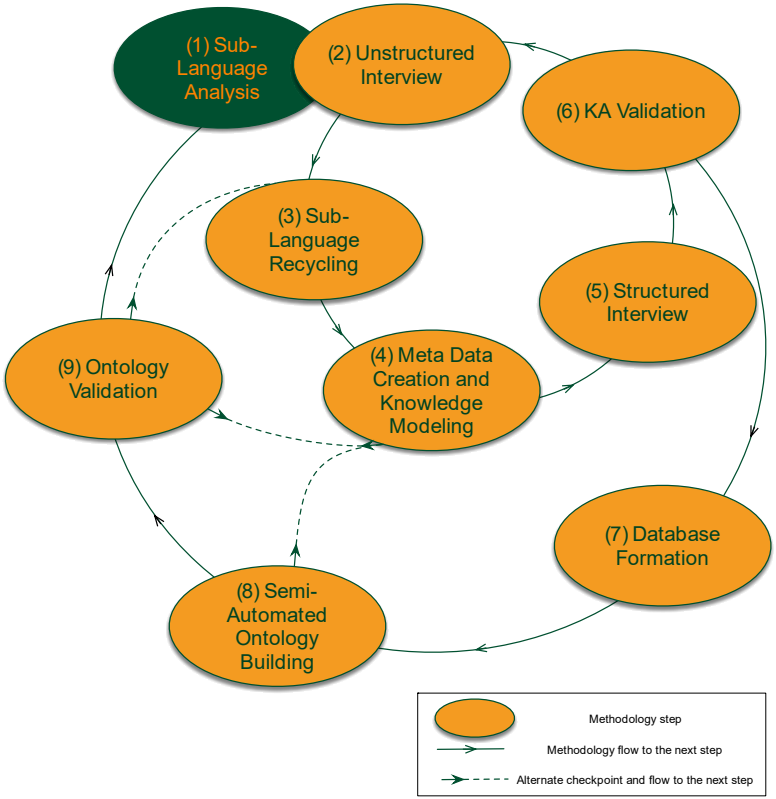
Hande Küçük McGinty, Lin, Yu, Saurabh Mehta, John Paul Turner, Dusica Vidovic, Michele Forlin, Amar Koleti et al. "Drug target ontology to classify and integrate drug discovery data." *Journal of biomedical semantics* 8, no. 1 (2017): 1-16.

KNowledge Acquisition and Representation Methodology (KNARM)



McGinty, Hande Küçük. "Knowledge Acquisition and Representation Methodology (KNARM) and Its Applications." PhD diss., University of Miami, 2018.

Sub-language Analysis



Literature review

* CDC Water Contamination Definitions (https://www.cdc.gov/healthywater/drinking/contamination.html)

* FDA Total Diet Study (https://www.fda.gov/food/fda-total-diet-study-tds/)

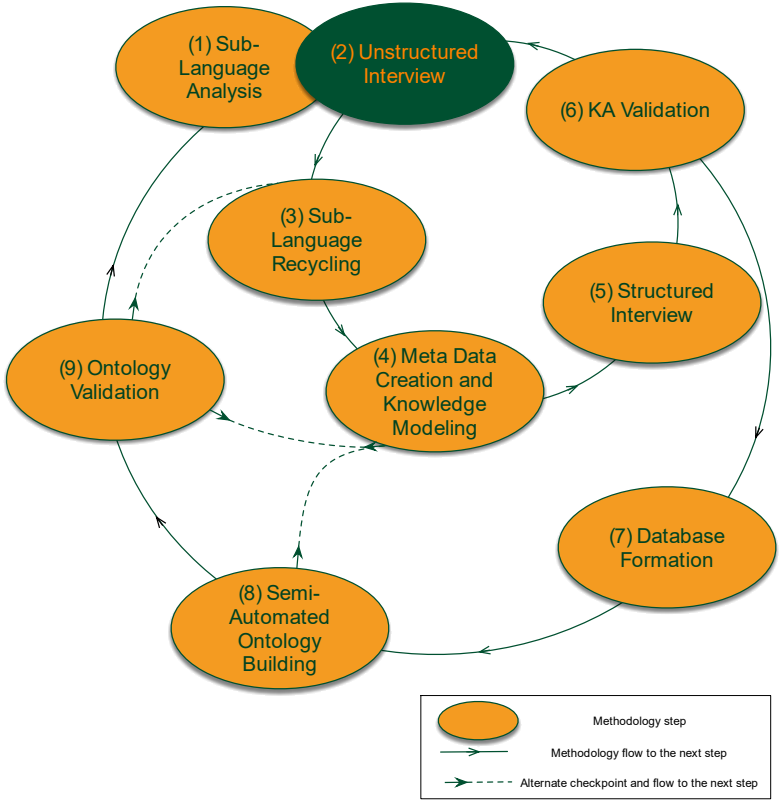
* Agency for Toxic Substances and Disease Registry (ATSDR) (https://www.atsdr.cdc.gov/pfas/PFAS-health-effects.html)

* ATSDR Report "Toxicological Profile for Perfluoroalkyls", May 2021, (https://www.atsdr.cdc.gov/toxprofiles/tp200.pdf)

What are the different datasets?

https://riversideca.gov/press/understanding-pfas

Unstructured Interview

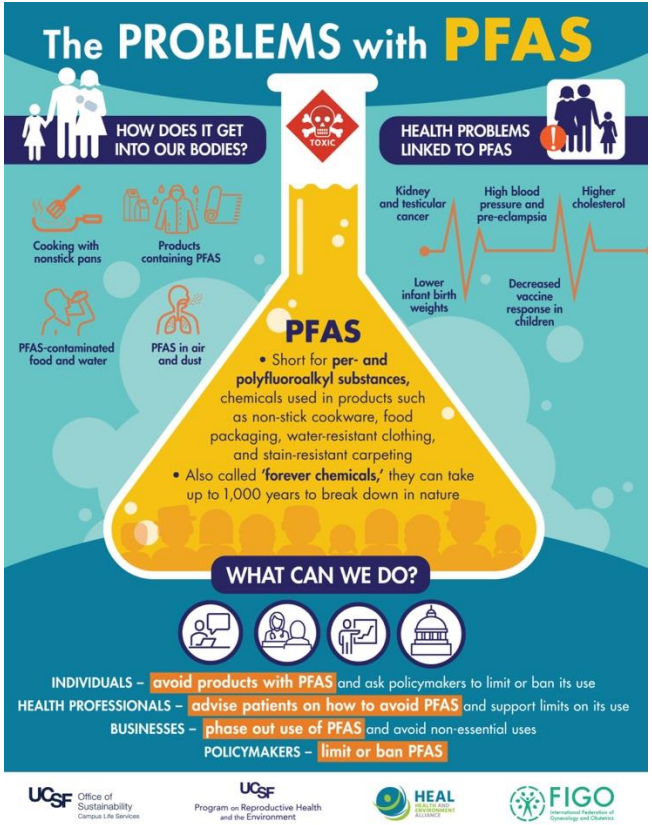


Understanding the questions

What may be some of our use-case problems:

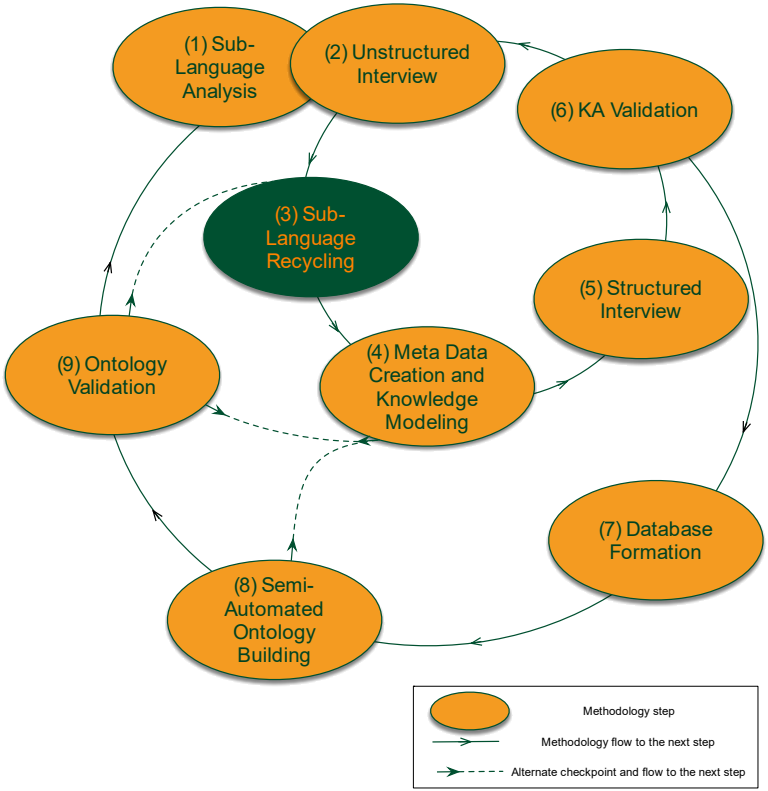
* Understand the size of the problem,
* Where and how to prioritize testing for PFAS,
* How PFAS contamination is affecting the environment and health of their communities,
* How to ensure safe food and water for their communities, and
* Where there are serious gaps in the knowledge about PFAS in the state/country.

Understanding the use cases and stakeholders..

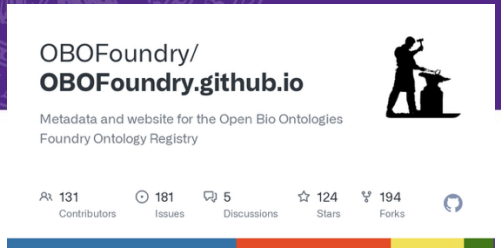


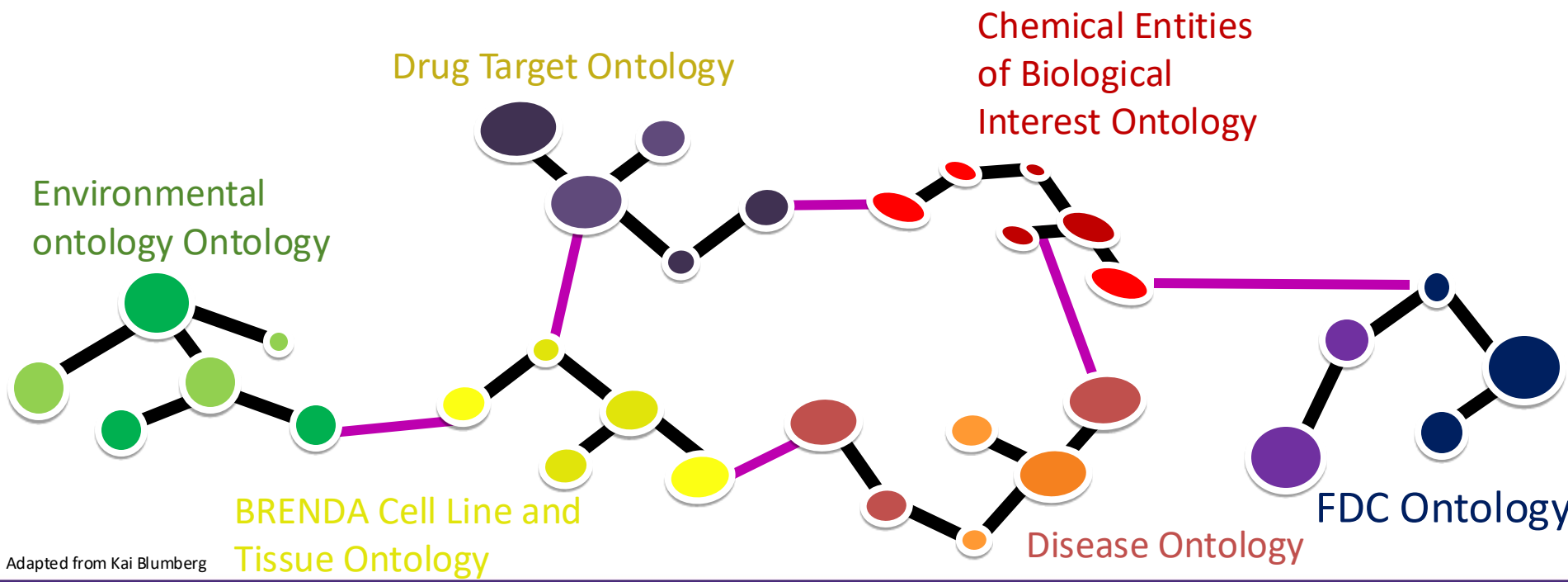
https://www.env-health.org/how-pfas-chemicals-affect-women-pregnancy-and-human-development-health-actors-call-for-urgent-action-to-phase-them-out/

Sub-Language Recycling



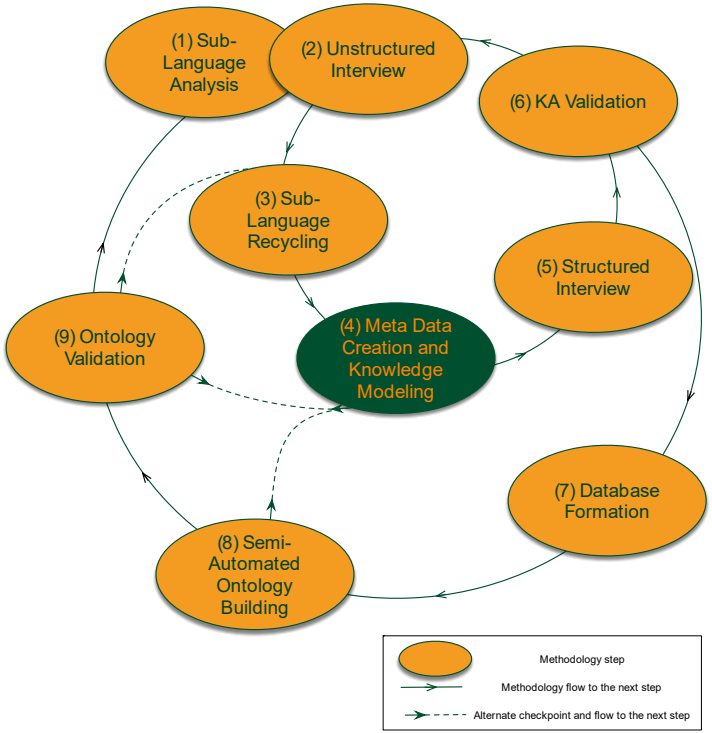
What are the surrounding efforts?



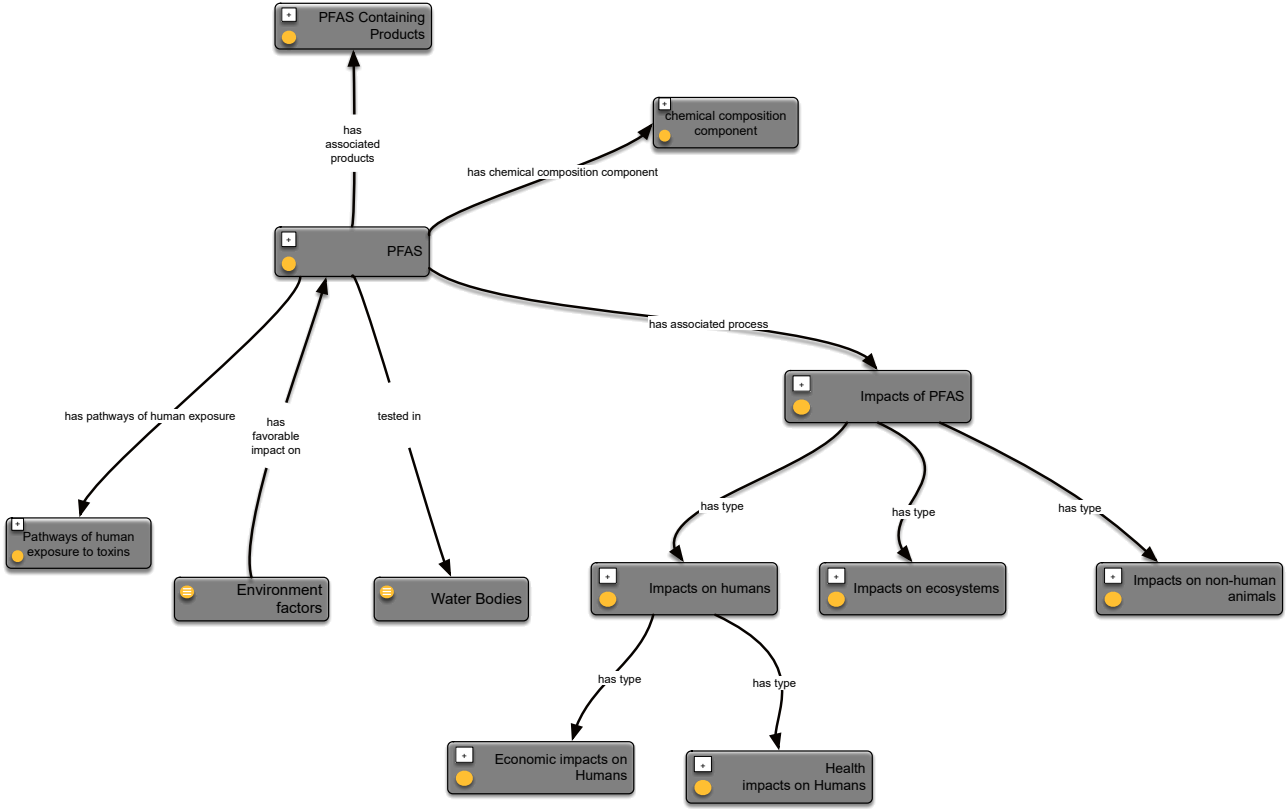


Adapted from Kai Blumberg

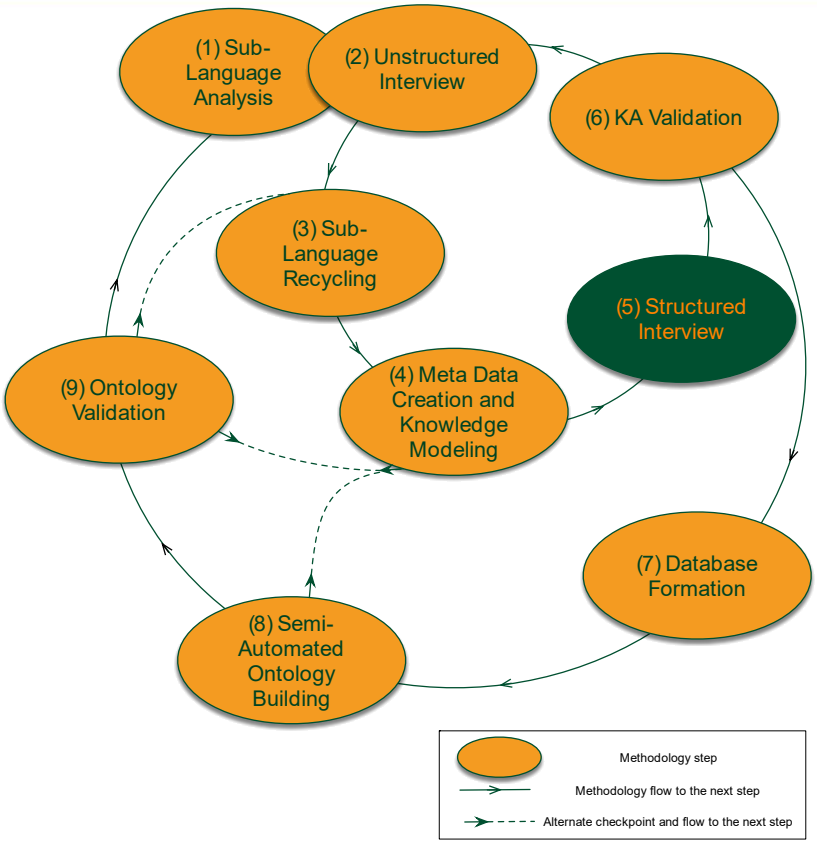
Meta Data Creation and Knowledge Modeling



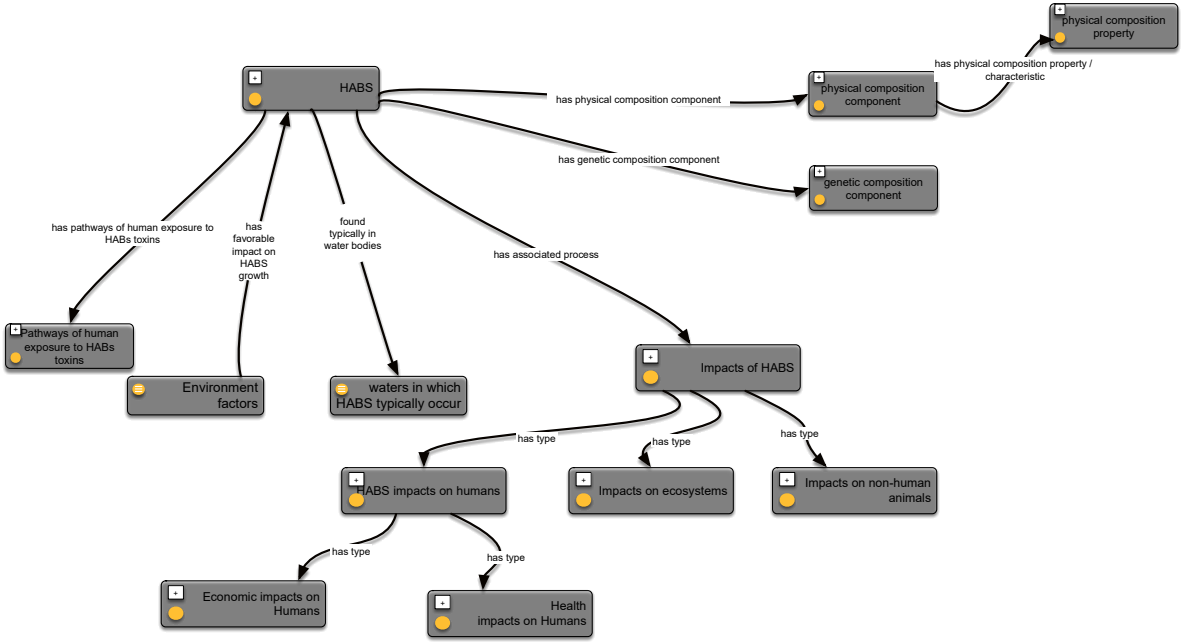
Modeling Example



Structured Interview

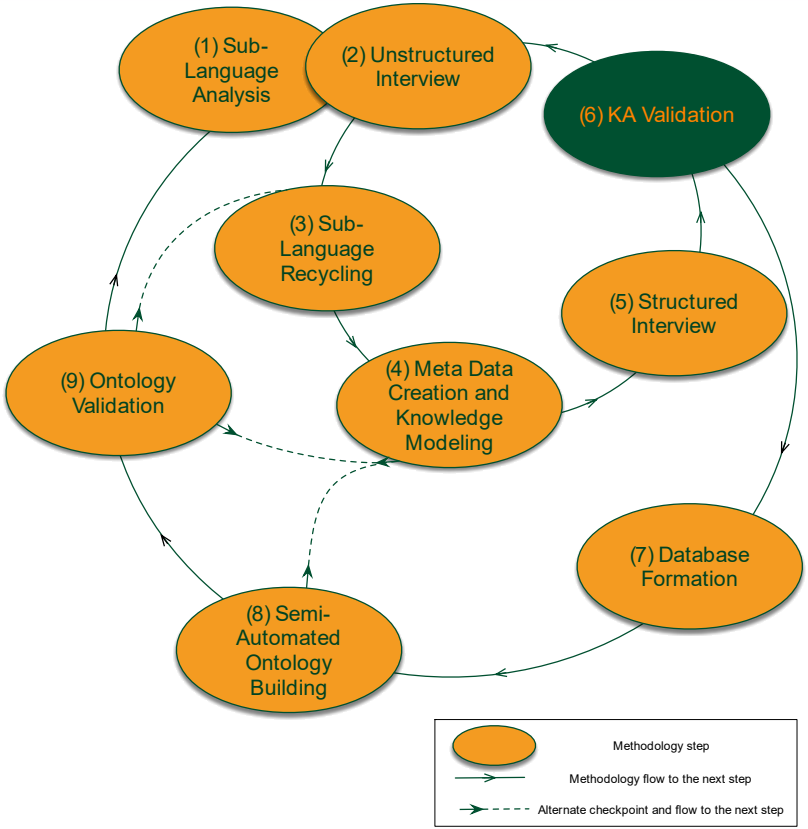


Because of the time restrictions – reused research..

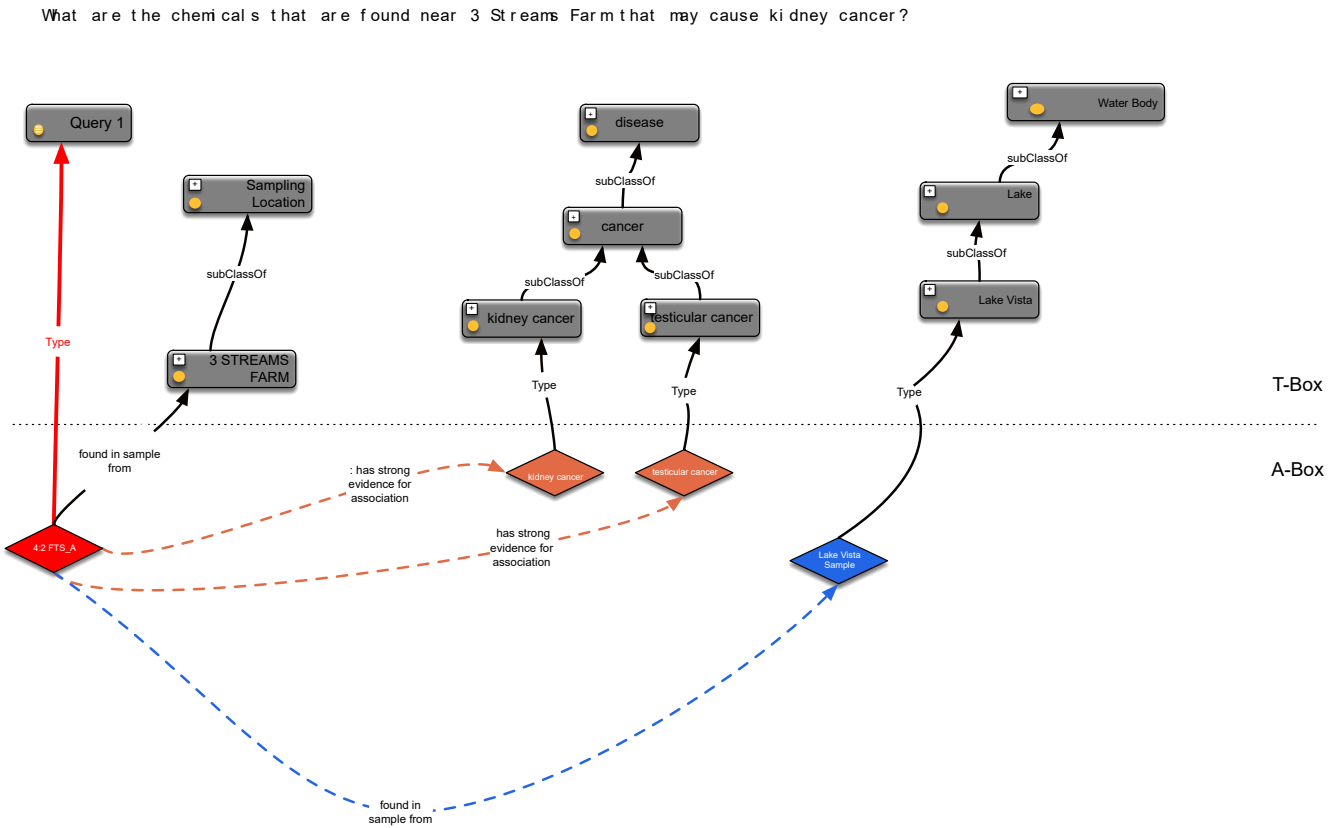


Leveraging Existing Literature on the Web and Deep Neural Models to Construct a Knowledge Graph Focused on Water Quality and Health Risks, Nikita Gautam, David Shumway, Megan Kowalcyk, Sarthak Khanal, Doina Caragea, Cornelia Caragea**, Hande McGinty** and Samuel Dorevitch, 2023, theWebConf, 11:50 AM – 12:00 PM – May 3rd, @AT&T Hotel and Conference Center Classroom #115

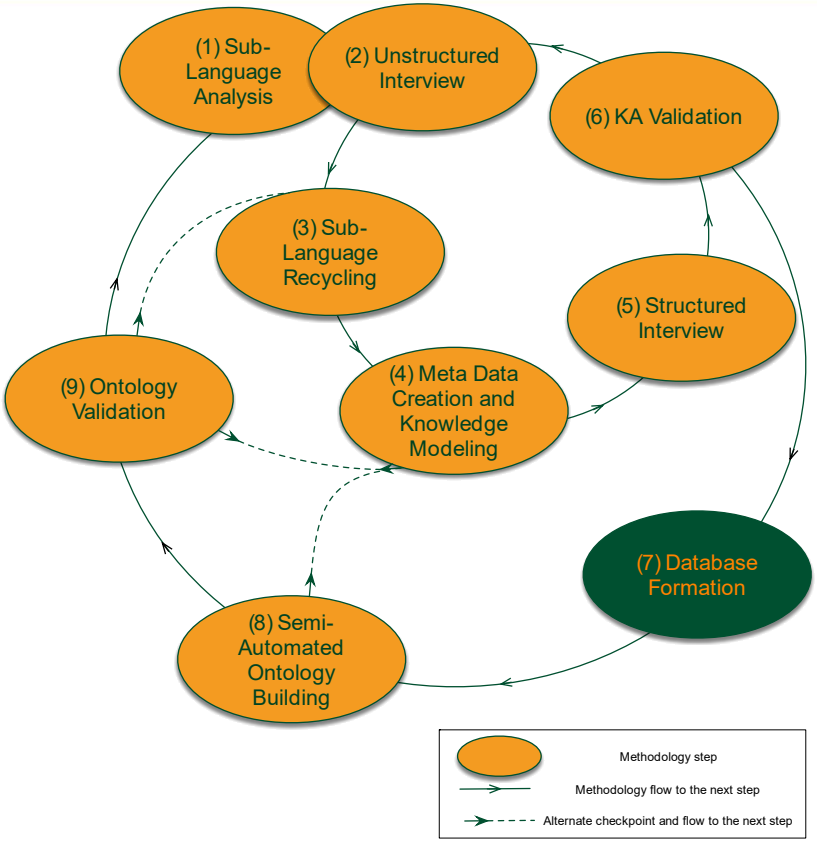
Knowledge Acquisition Validation



Ensure that we can answer questions of interest



Database Formation



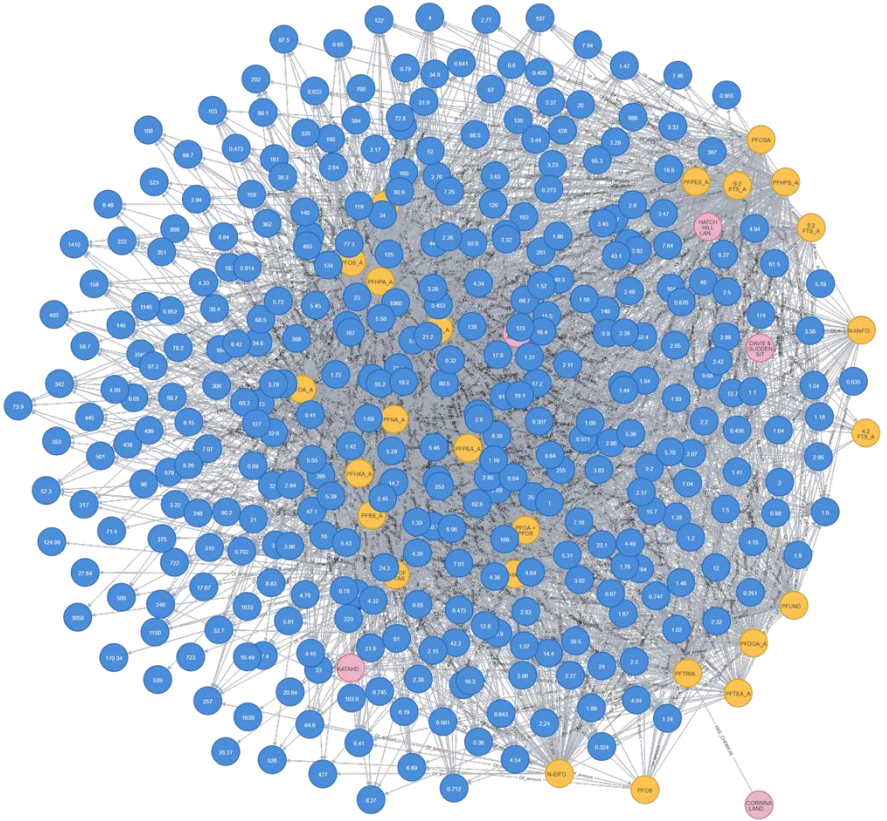
Sustainable Data Storage

MAINE PFAS DATA (2007-2022)

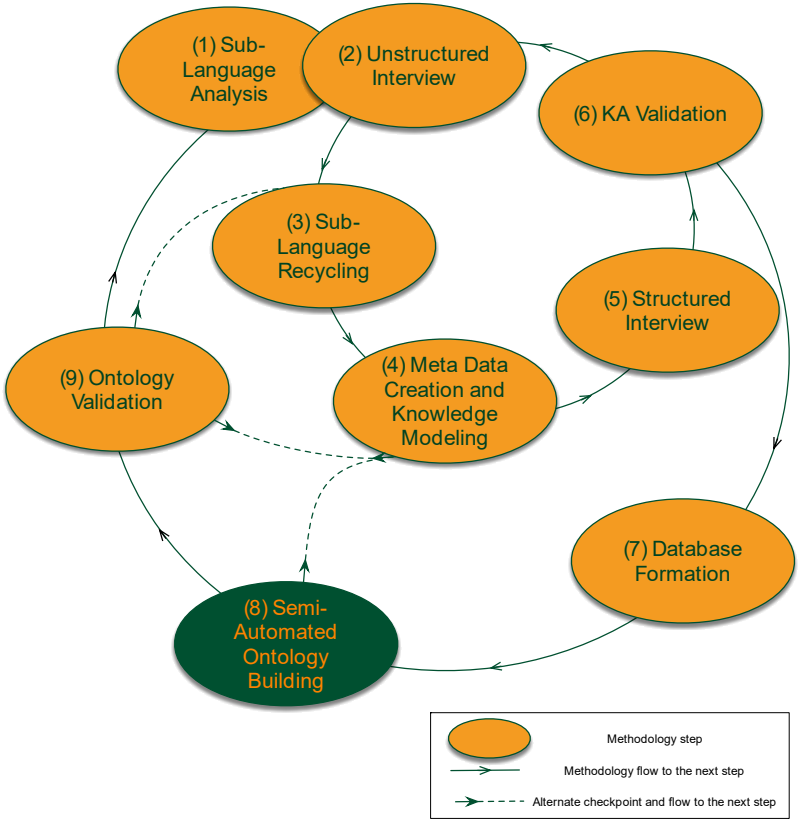
CURRENT SITE NAME	Town	SAMPLE POINT SEQ	SAMPLE DATE	SAMPLE TYPE	PARAMETER	CONCENTRATION	UNITS	LAB QUALIFIER	RL	TS
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	HFPO-DA_A		ng/L	U	43	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFODA_A		ng/L	U	3.44	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFOS_A	281	ng/L		1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFUNDA_A		ng/L	U	1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	N-MeFOSAA		ng/L	U	1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFPEA_A	32.6	ng/L		1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFPES_A		ng/L	U	1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	6:2 FTS_A		ng/L	UJ	1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	N-EtFOSAA	2.89	ng/L	J	1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFHXA_A	34.8	ng/L		1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFDOA_A		ng/L	U	1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFOA_A	113	ng/L		1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFDA_A	5.9	ng/L		1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFOS_A		ng/L	U	1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFHXS_A	6.27	ng/L		1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFBA_A	12.7	ng/L		1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFBS_A		ng/L	U	1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFHPA_A	26.9	ng/L		1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFHPS_A	3.37	ng/L		1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFNA_A	12.3	ng/L		1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFTEA_A		ng/L	U	1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	8:2 FTS_A		ng/L	U	1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFHXDA_A		ng/L	U	3.44	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFNS_A		ng/L	U	1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFTRIA_A		ng/L	U	1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFOSA		ng/L	U	1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	4:2 FTS_A		ng/L	UJ	1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	ADONA_A		ng/L	U	1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFOA + PFOS	394	ng/L		1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	SUM OF 6 PFAS	445	ng/L		1.72	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39569	12/30/21	L	HFPO-DA_A		ng/L	U	62	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39569	12/30/21	L	PFODA_A		ng/L	U	4.96	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39569	12/30/21	L	PFUNDA_A	0.953	ng/L	J	2.48	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39569	12/30/21	L	N-MeFOSAA	8.36	ng/L	J	2.48	N
PIXELLE ANDROSCOGGIN JAY LANDFILL	JAY	39569	12/30/21	L	PFPEA_A	428	ng/L	J	2.48	N

DB view of Samples, Chemicals and Concentration Values

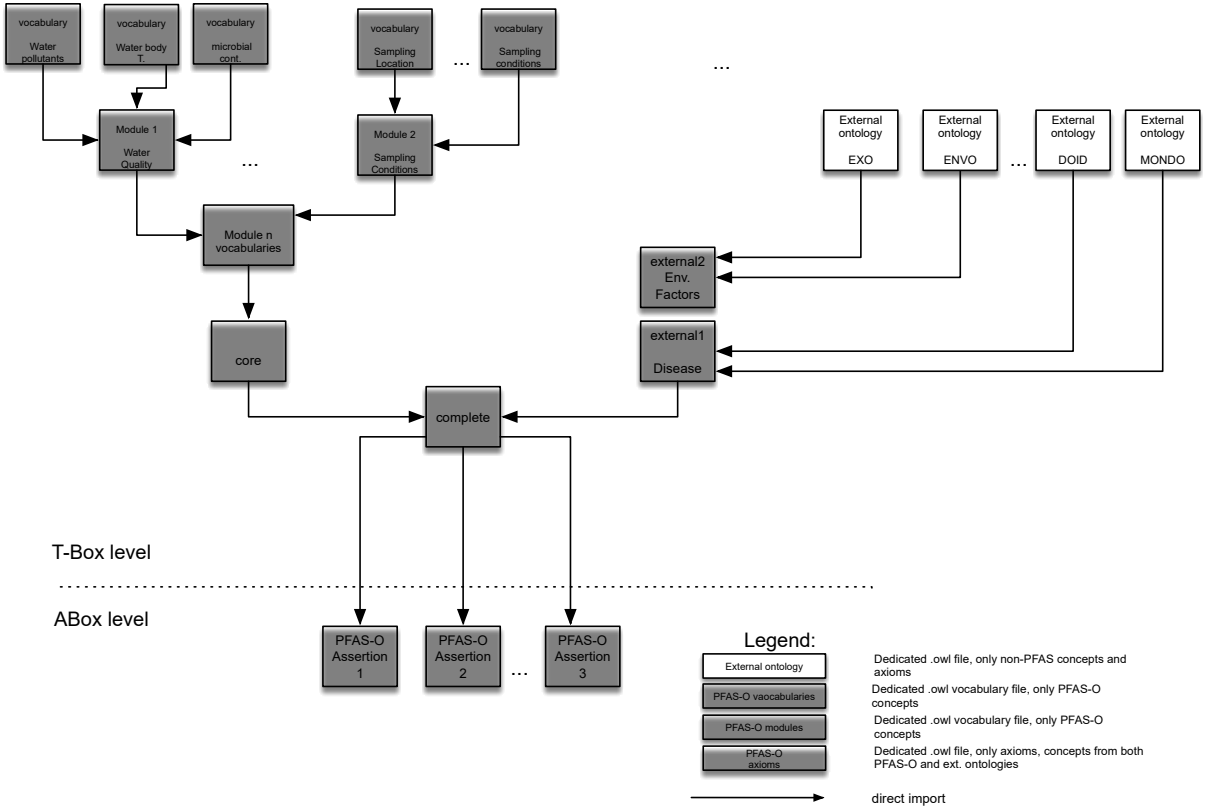
```
LOAD CSV WITH HEADERS FROM 'file:///Book.csv'
AS row

MERGE (loc:Location {name: row.Location})

MERGE (Chem:Chemical {name: row.Chemical})

MERGE (cont:Content {name: row.Content})

MERGE (loc)-[:HAS_CHEMICAL]->(Chem)

MERGE (Chem)-[:Of_amount]->(cont)

RETURN loc, cont, Chem
```



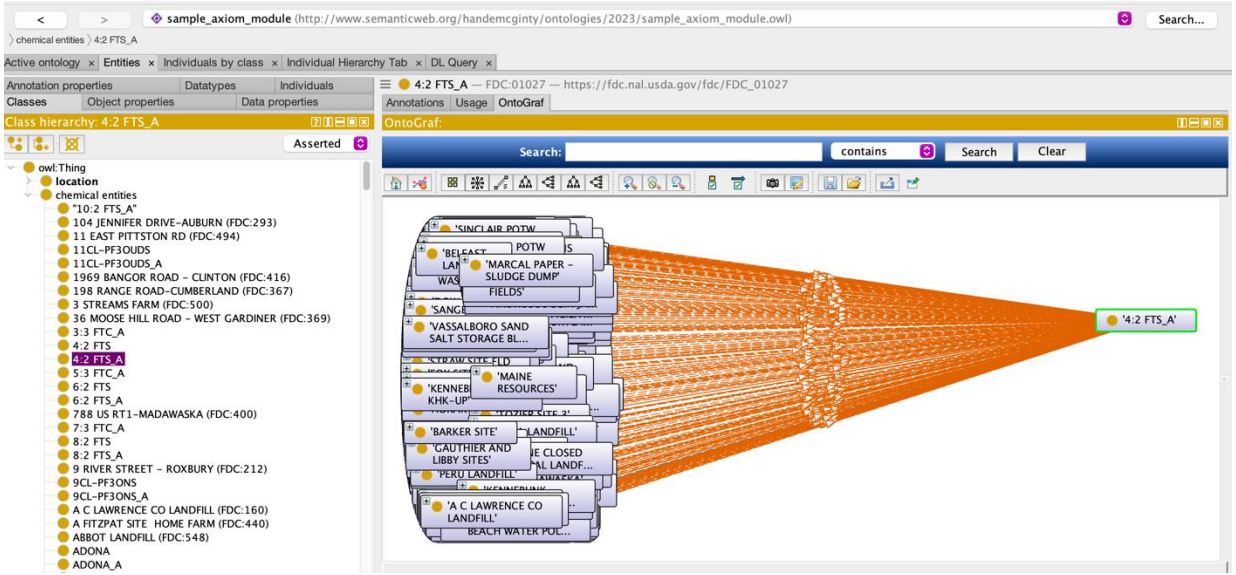
Semi-Automated Ontology Building



Semi-Automated Modular Architecture

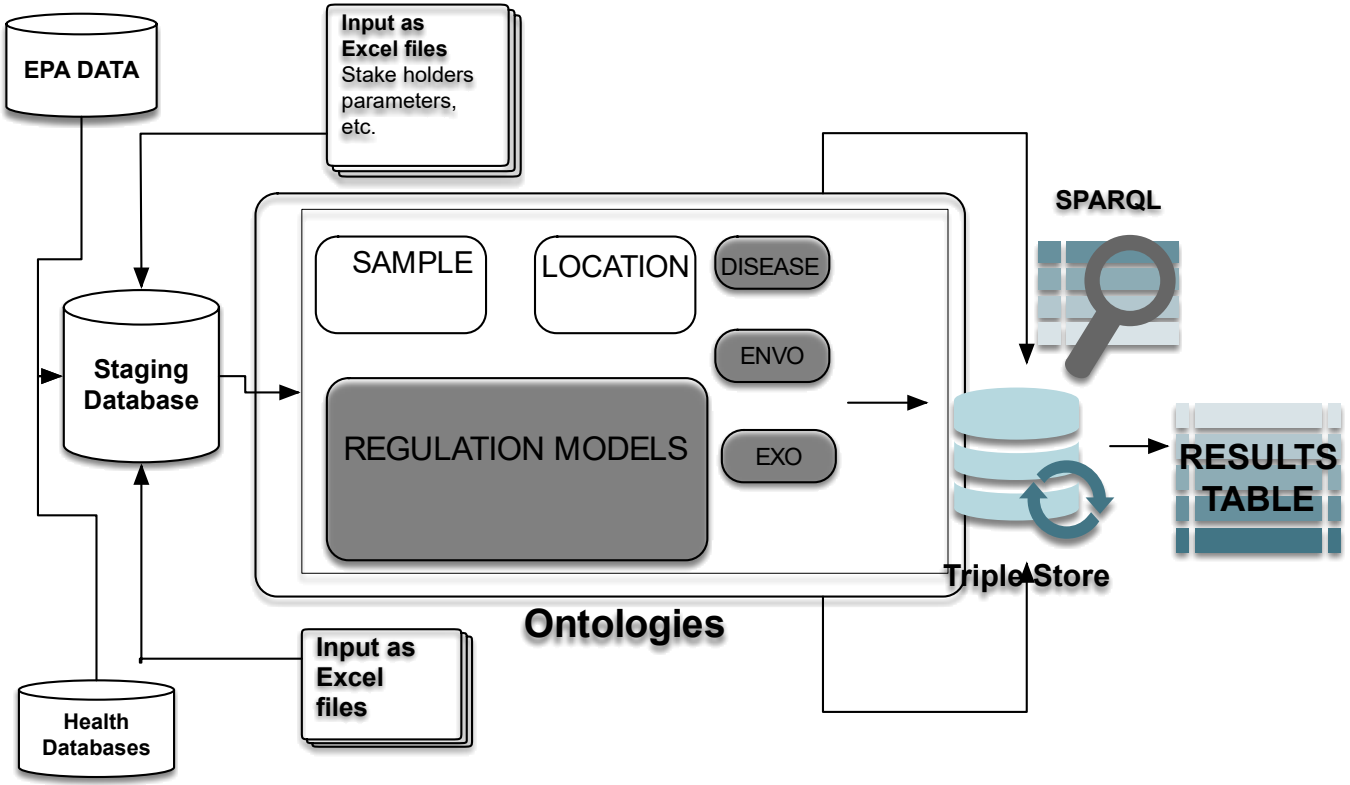


Semi-Automated Ontology Output

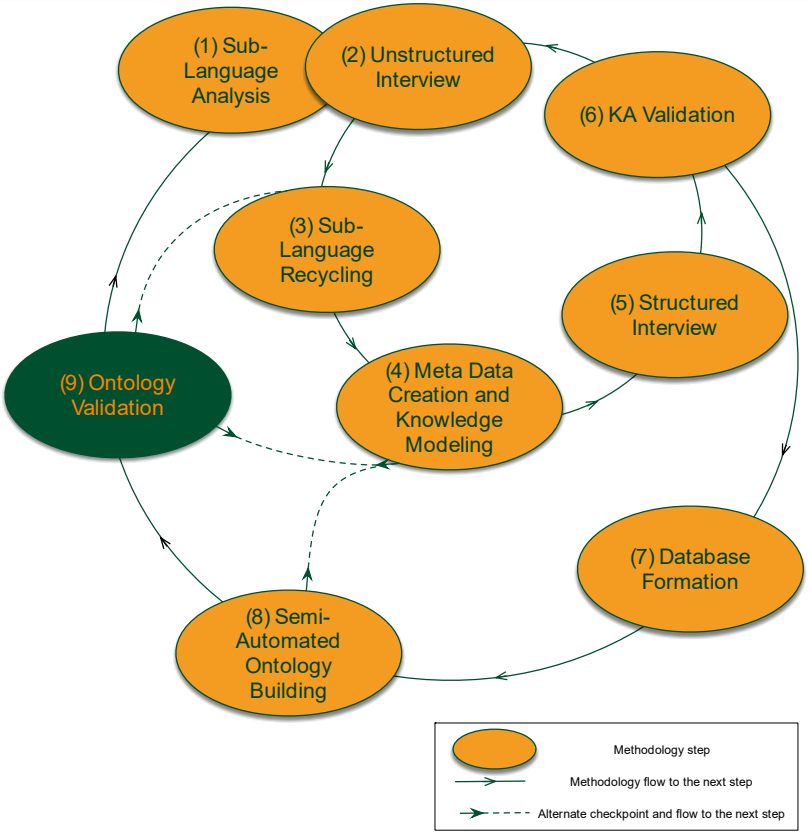



```
robot template --input /Users/yz/Desktop/csv_to_owl/output/parameter.owl \
  --template /Users/yz/Desktop/csv_to_owl/axiom_edit.csv \
  --prefix "id:https://fdc.nal.usda.gov/fdc/FDC" \
  --output /Users/yz/Desktop/csv_to_owl/output/axiom_edit.owl
```

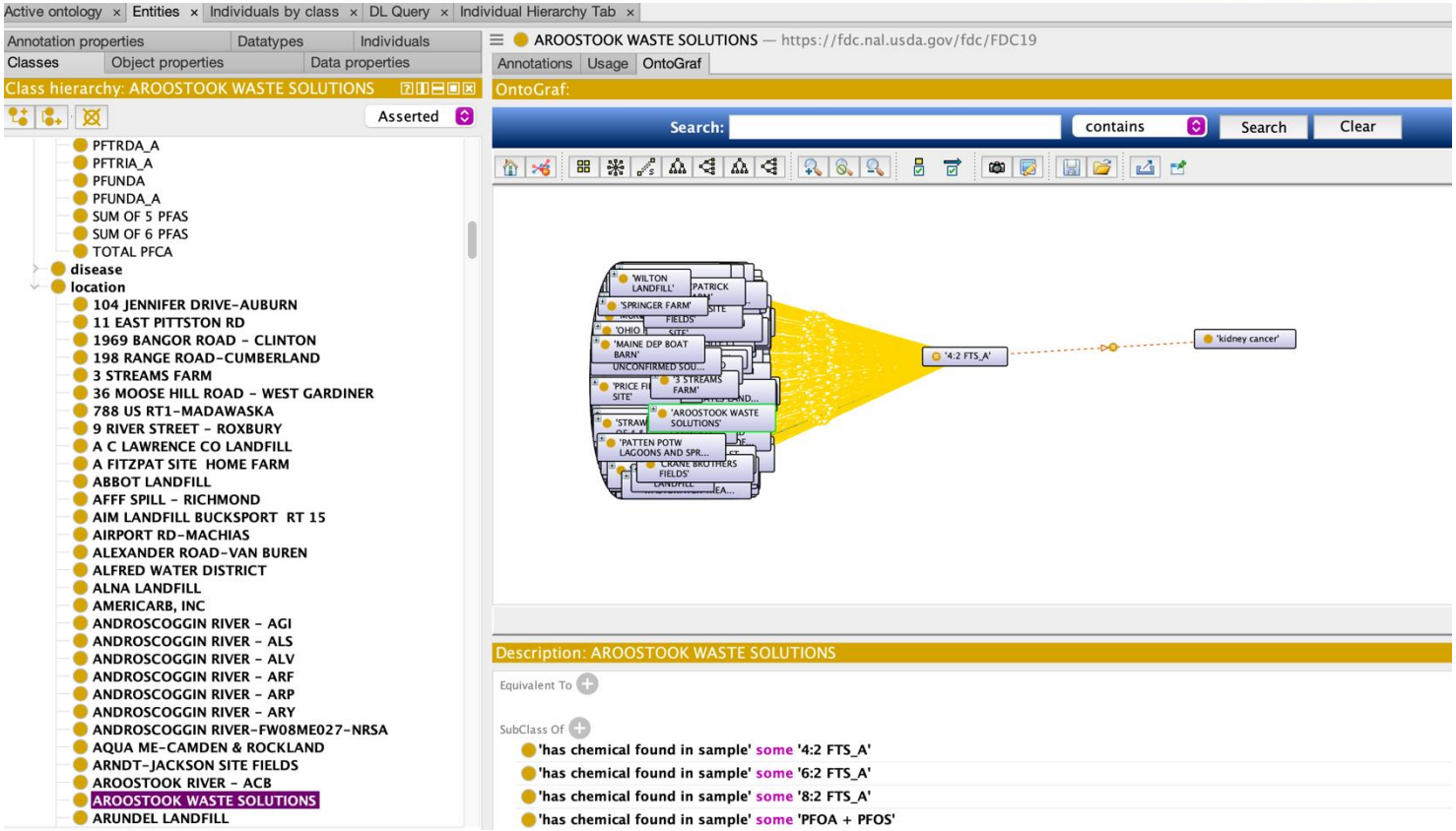
These Systems are complex, so we need a systematic approach to build them!



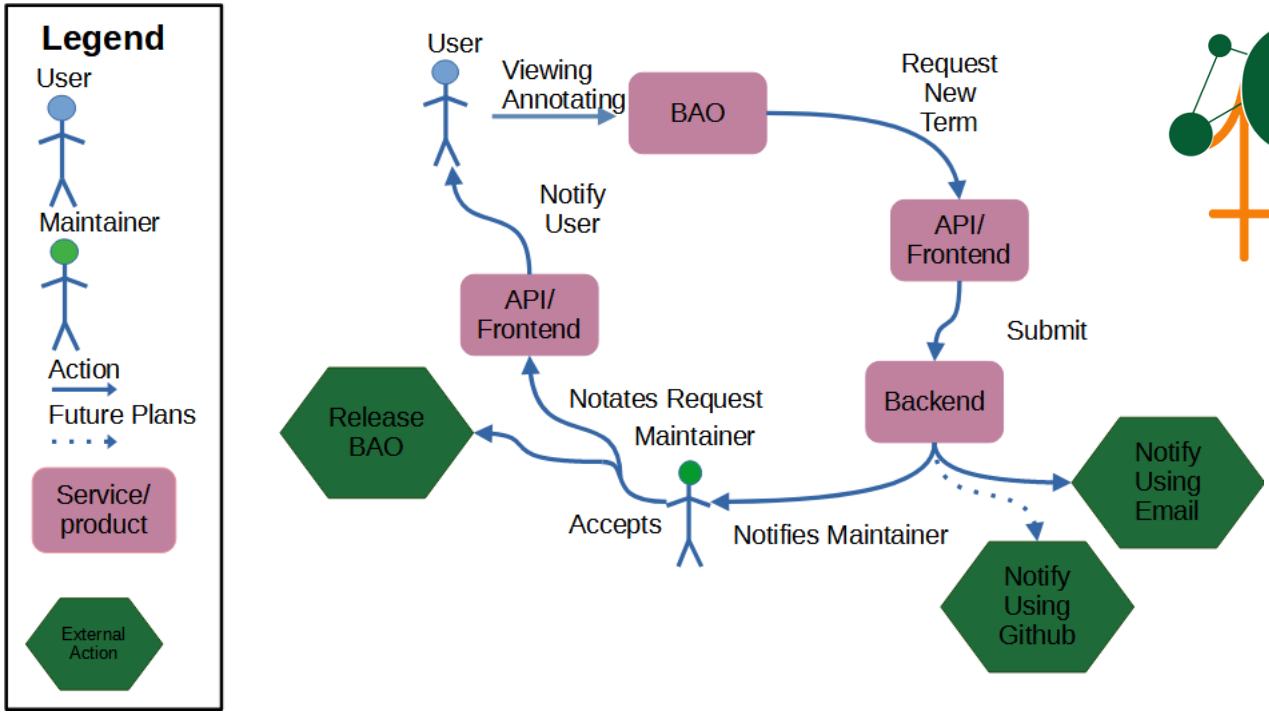
Ontology Validation



Quality Control



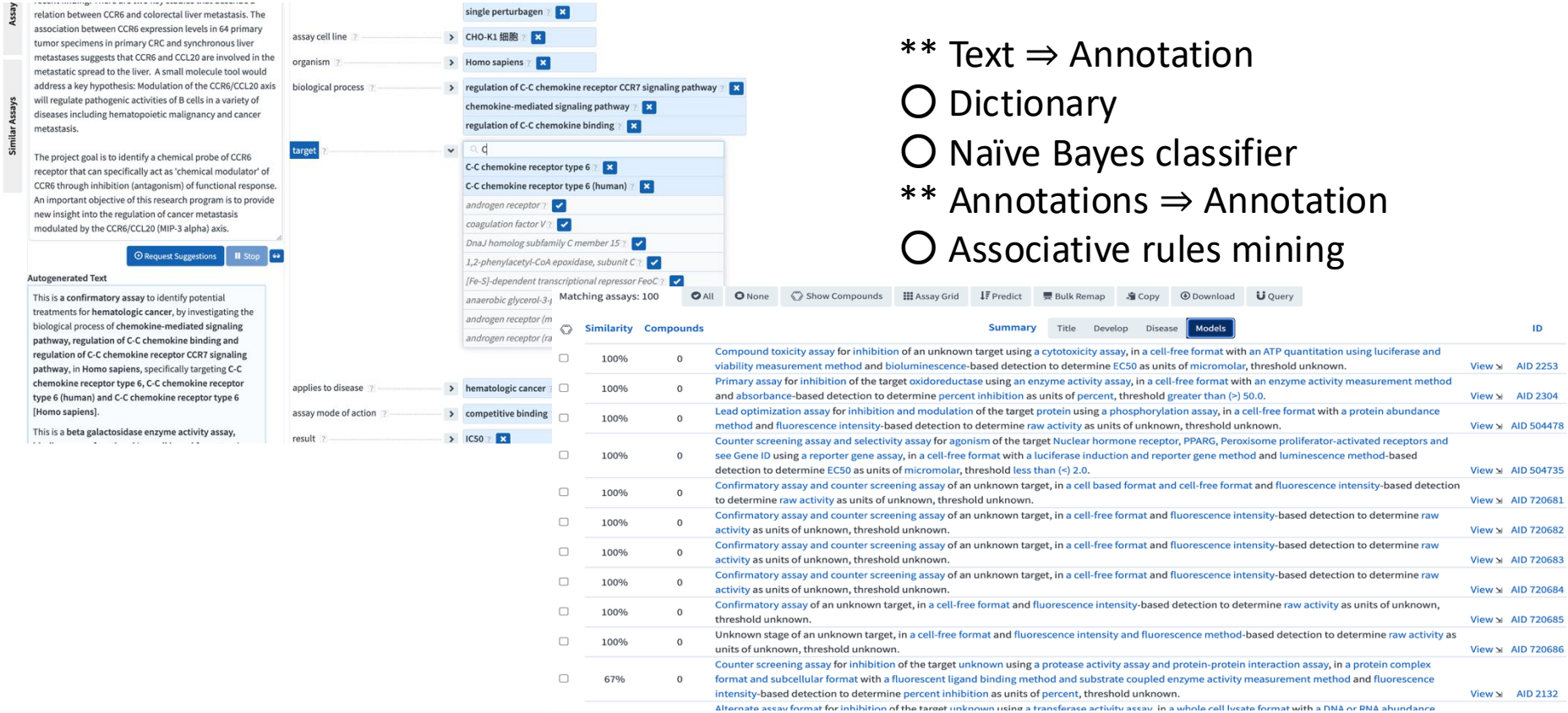
OntoloBridge



Hande Küçük McGinty, John Paul Turner, Alex M. Clark, Peter Gedeck, John Graybeal, Michael Dorf, Caty Chung, Mark Musen, Barry A. Bunin and Stephan Schürer;
OntoloBridge – A FAIR Semi-Automated Ontology Update Request System; (In preparation)

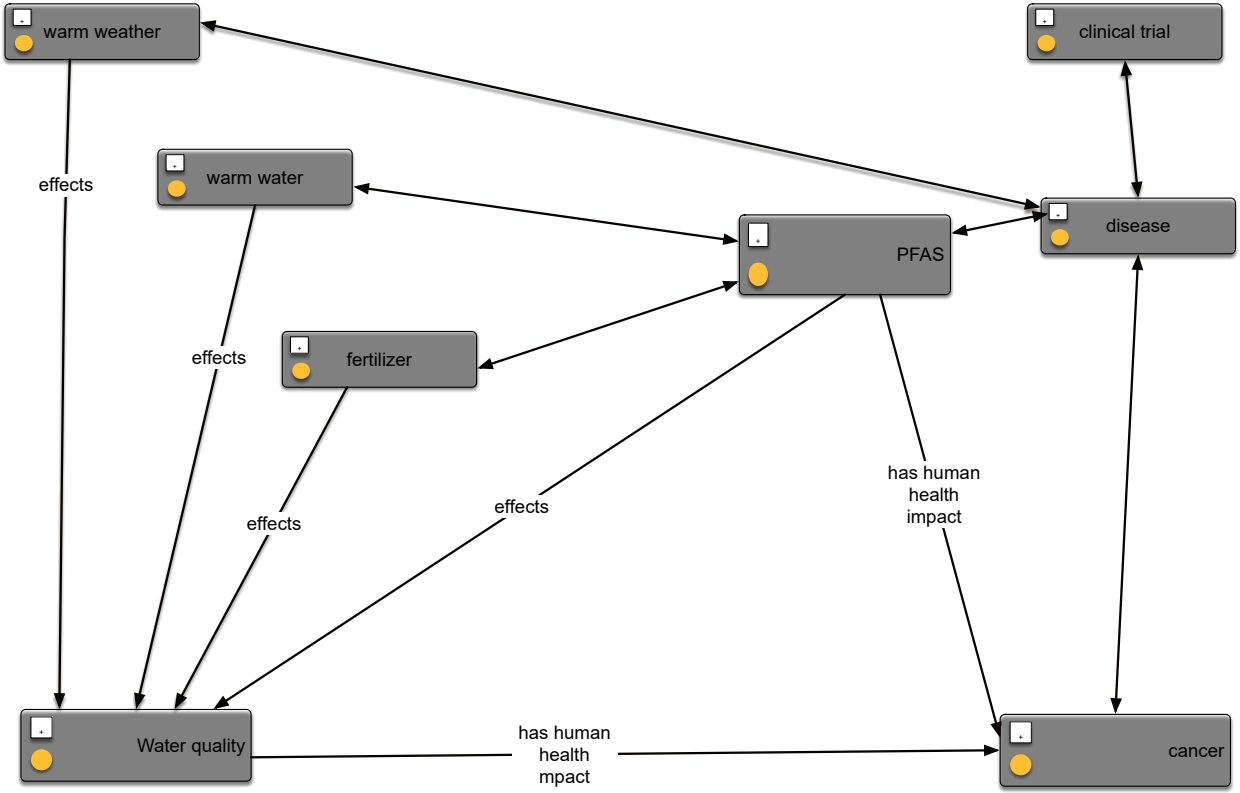
NIH - 1U01LM012630-01 (Unifying Templates, Ontologies and Tools to Achieve Effective Annotation of Bioassay Protocols (OntoloBridge))

Use it as a part of backend for Applications using Machine Operable Data



** Text ⇒ Annotation
○ Dictionary
○ Naïve Bayes classifier
** Annotations ⇒ Annotation
○ Associative rules mining

Scalable Knowledge Graph Generation



Acknowledgements

NIH, USDA and FDA funded projects

- RC2HG005668 (BioAssay Ontology and Software Tools to Integrate and Analyze Diverse Data Sets)
- 5U01HL111561-02 (LINCS Information FramEwork (LIFE) to Integrate and Analyze Diverse Data Sets)
- U54CA189205 (Illuminating the Druggable Genome Knowledge Management Center, IDG-KMC)
- U54HL127624 (BD2K LINCS Data Coordination and Integration Center, DCIC).
- 1U01LM012630-01 (Unifying Templates, Ontologies and Tools to Achieve Effective Annotation of Bioassay Protocols (OntoloBridge))
- 2R44TR000185-04 (BioAssay Express Phase 2)
- 3R43TR002528-01S1 (NIH I-CORPS Grant Supplement to SBIR Phase 1 Grant related with Chemical Mixtures)
- 2R44TR002528-02 (Digital Representation of Chemical Mixtures to Aid Drug Discovery and Formulation)

Teammates

- My students: Yinglung Zhang and Aryan Dalal
- Harrington Lab – Ohio U
- Musen Lab & BioPortal Team – Stanford
- Collaborative Drug Discovery Team
- Schurer Lab – Uni. Of Miami, Coral Gables, FL

Check us out again!

Monday, May 1st

Workshop: Knowledge Graphs for Sustainability - KG4S

Workshop

9:00 AM – 12:30 PM

University of Texas Gates Dell Complex - Room # 6.202

11:05-11:25 Short paper presentation:

Yinglun Zhang, Antonina Broyaka, Jude Kasten, Allen Featherstone, Cogan Shimizu, Pascal Hitzler and Hande Küçük McGinty. *Sustainable Grain Transportation in Ukraine Amidst War Utilizing KNARM and KnowWhereGraph*

Web4Good: Health

11:00 AM – 12:30 PM at AT&T Hotel and Conference Center Classroom #115

Web4Good

Wednesday, May 3rd

8420 Leveraging Existing Literature on the Web and Deep Neural Models to Construct a Knowledge Graph Focused on Water Quality and Health Risks

11:50 AM – 12:00 PM

Description

Authors: Nikita Gautam, David Shumway, Megan Kowalcyk, Sarthak Khanal, Doina Caragea, Cornelia Caragea, Hande McGinty and Samuel Dorevitch

Thank you

QUESTIONS?

HANDE@KSU.EDU