BioSciences

Staying grounded: assembling structured biological knowledge with help from large language models

J. Harry Caufield

Sep 6 2023

AgBioData Webinar

BERKELEY LAB

U.S. DEPARTMENT OF ENERGY

Background

Where does biological knowledge come from?

It's the result of repeated observations.

Learning from those observations is a task in itself, but can be automated.

How may we automate:

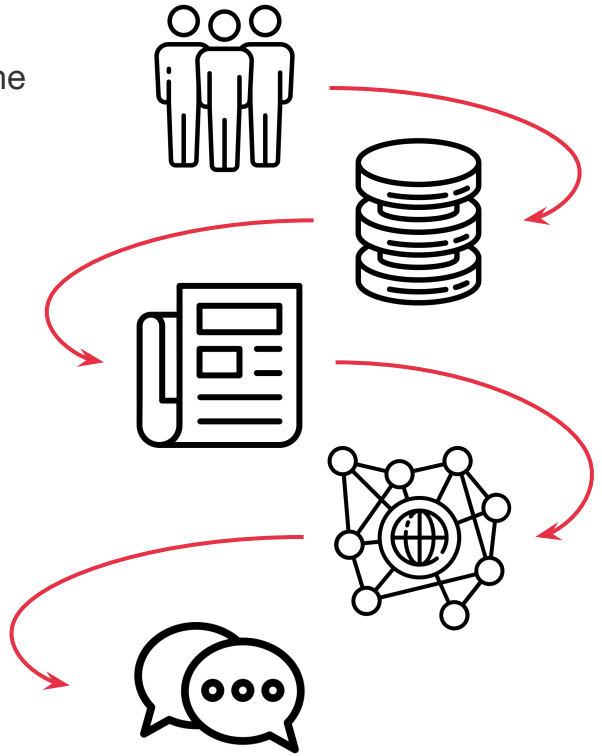
- learning from literature?
- comparing findings?
- Integrating observations?
  - Across different studies or replicates?
  - Across different knowledge bases?
  - Across different fields and disciplines?
  - Of similar concepts, even when described in different contexts?

Background - what are current methods?

We need structured data.
Consistent data models, standards, and ontologies help but don't do the work of structuring data for us.
How can we curate structured data from unstructured text?

- **Human experts** - a famously limited resource
- **Rule-based extractors** like SemMedDB*
  - Fairly high precision, low recall
  - Varies by domain and structure to extract
- **Enrichment** of terms and/or annotations, like MELODI**
  - Subject to publication bias
- **Neural networks** for Natural Language Processing (e.g., LSTMs)
  - Require extensive labeled training data…
  - and even then, they may overfit
- **Language models** (e.g., BERT)
  - Avoid learning the basics of language from scratch

Each method may still be effective for some use cases!



* Kilicoglu et al. Bioinformatics (2012) doi:10.1093/bioinformatics/bts591
** Elsworth et al. Int J Epidemiol. (2018) doi:10.1093/ije/dyx251

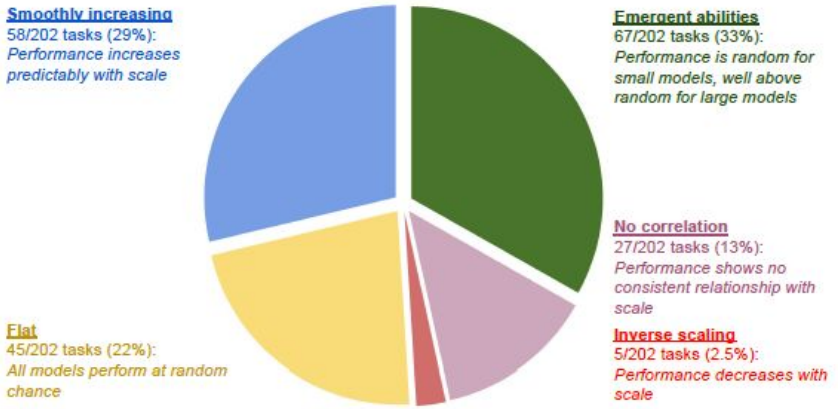
Background - can LLMs help?

Can we translate unstructured scientific text directly into arbitrary knowledge schemas?
What if:

- Those schemas are **complex** and involve **nested subclasses**?
  - Like “each relationship between *i* and *j* where *i* is an object of type A and *j* is an object of type B but only from set C”
- We need to link to external unique identifiers?

Can Large Language Models (LLMs) like GPT-3+ help?

Background - No, really, can LLMs help?



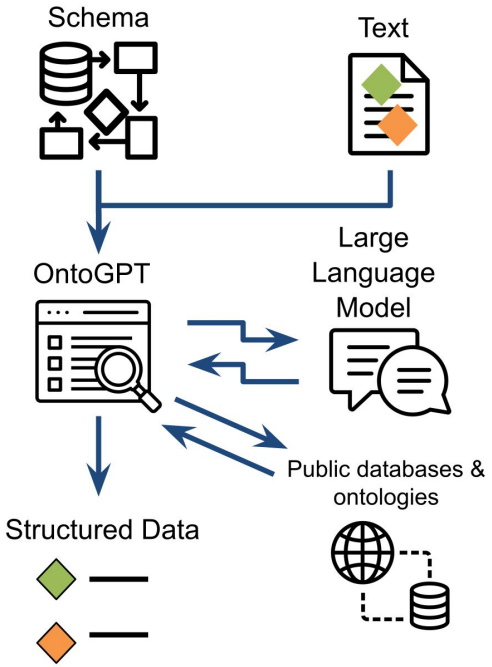
Counts refer to the 202 tasks in the BIG-Bench language tech benchmark.

"Another example we produced that was outside of the scope for the present study was when asked about vaccines, GPT-3 responds, "Vaccines are not 100% effective. Vaccines can cause serious side effects. Vaccines can cause death. Vaccines are not tested for safety or effectiveness""

Figure from Bowman arXiv:2304.00612v1 [cs.CL]. (2023)

- Bigger is better...
  - for some tasks.
  - More interesting: *emergent* behaviors
- Training data often unclear
  - Or may include fictitious claims
- Human-like performance, even in biomedicine
  - But without human reasoning
  - Or user ability to distinguish between human communication vs. generated text
  - See Levine et al. The Diagnostic and Triage Accuracy of the GPT-3 Artificial Intelligence Model. medRxiv (2023) doi:10.1101/2023.01.30.23285067
- Hallucinations
  - LLMs are grounded in language, not fact

Approach - SPIRES

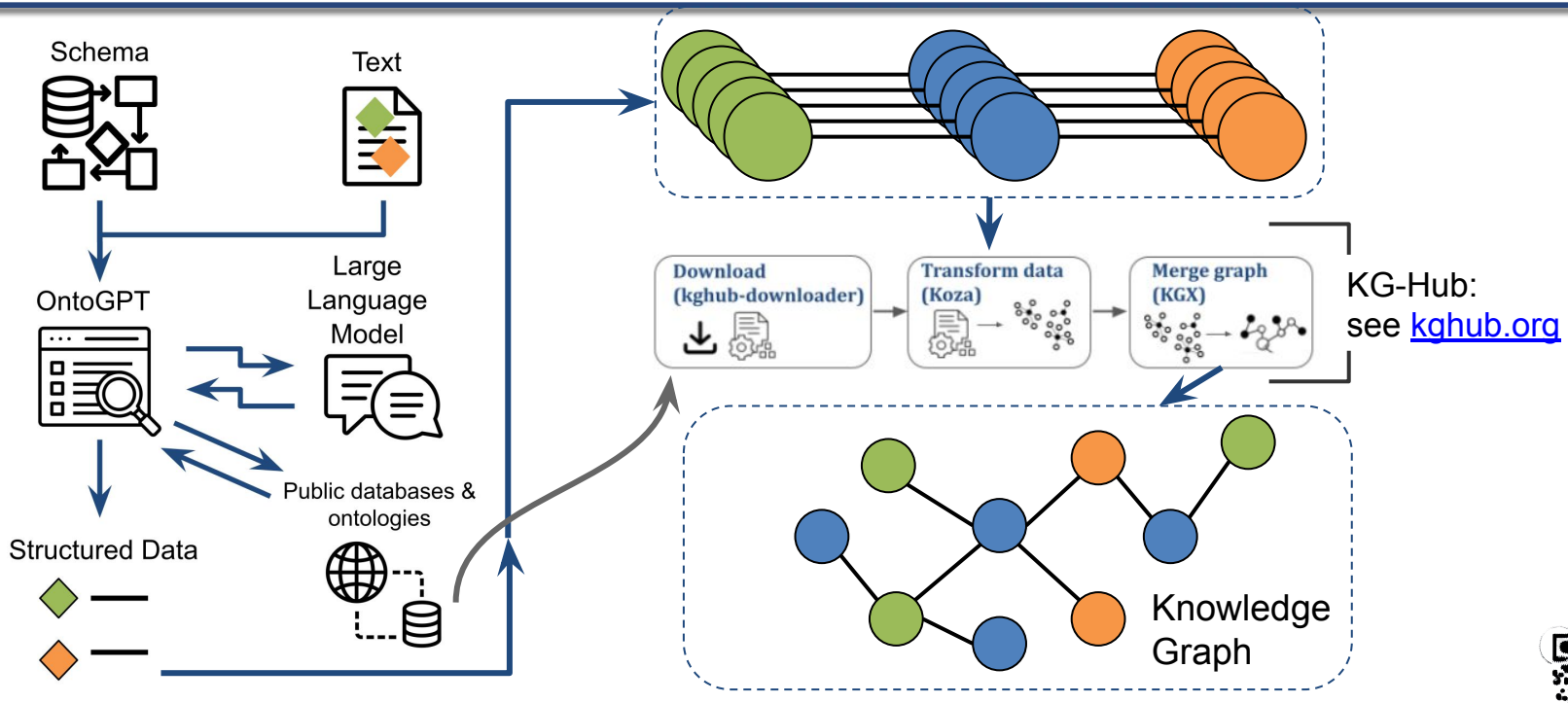


SPIRES: Structured Prompt Interrogation and Recursive Extraction of Semantics
(or, **information extraction grounded in reality**)

Populate knowledge schemas by providing structured prompts to a LLM. Parse the resulting response. Ground, map, and normalize the identifiers within using the **Ontology Access Kit** (https://github.com/INCATools/ontology-access-kit).

Available through **OntoGPT**:
https://github.com/monarch-initiative/ontogpt

Approach - SPIRES



See https://arxiv.org/abs/2304.02711 and https://github.com/monarch-initiative/ontogpt

Why not just tell the LLM to extract all relationships?

https://tpwd.texas.gov/publications/nonpwdpubs/young_naturalist/animals/predator_prey_relationship/index.phtml

Predator-Prey Relationship

In Nature's plan, there is no such thing as a good or bad animal. There are only those animals that kill to eat (predators) and those that are killed and eaten (prey). It is unfair for us to judge their natural survival actions by our own ideas of what is good and bad or to decide that all predators are villains.

When the coyote chases, catches, kills, and eats the cute, furry rabbit, it is not being intentionally cruel. It is only meeting its physical need for food with whatever is available at the time. Its meal could just as easily have been a mouse, rat, or other small creature if one of these other animals had been closer or easier to catch.

Whether it is a roadrunner swallowing a lizard, a blue crab eating a fish, an American Egret catching a perch, a shark investigating a surface disturbance that could be a meal, or a man rounding up a cow for slaughter, the predator kills to provide food for itself or its family.

Is there really any difference between a hungry mountain lion stalking, killing, and eating a deer and a hunter stalking, shooting, and taking a deer home for his family to eat? Both the mountain lion and the human hunter are predators, using the deer for food. You may think that since you are not a hunter, you are not a predator. How about the beef, pork, lamb, fish, or chicken you eat each day? Just because you do not kill the animal yourself does not make you less of a predator. You are still an animal eating the flesh of another animal to satisfy your hunger.

People usually think of predators as meat-eating mammals (carnivores), but the bird, reptile, amphibian, fish, and insect families also have their predators. Eagles, hawks, falcons, and owls are probably the best-known

```
raw_completion_output: |-
  subject: mountain lion
  predicate: stalking, killing, and eating
  object: deer

  subject: hunter
  predicate: stalking, shooting, and taking
  object: deer

  subject: mountain lion
  predicate: using
  object: deer

  subject: human hunter
  predicate: using
  object: deer

  subject: you
  predicate: think
  object: since you are not a hunter, you are not a predator

  subject_qualifier: you
  predicate: eat
  object_qualifier: beef, pork, lamb, fish, or chicken each
day
```

- We can improve the result by using a more defined relationship *and* by grounding to a consistent set of concepts.

Why not just tell the LLM to extract all relationships?

https://tpwd.texas.gov/publications/nonpwdpubs/young_naturalist/animals/predator_prey_relationship/index.phtml

Predator-Prey Relationship

In Nature's plan, there is no such thing as a good or bad animal. There are only those animals that kill to eat (predators) and those that are killed and eaten (prey). It is unfair for us to judge their natural survival actions by our own ideas of what is good and bad or to decide that all predators are villains.

When the coyote chases, catches, kills, and eats the cute, furry rabbit, it is not being intentionally cruel. It is only meeting its physical need for food with whatever is available at the time. Its meal could just as easily have been a mouse, rat, or other small creature if one of these other animals had been closer or easier to catch.

Whether it is a roadrunner swallowing a lizard, a blue crab eating a fish, an American Egret catching a perch, a shark investigating a surface disturbance that could be a meal, or a man rounding up a cow for slaughter, the predator kills to provide food for itself or its family.

Is there really any difference between a hungry mountain lion stalking, killing, and eating a deer and a hunter stalking, shooting, and taking a deer home for his family to eat? Both the mountain lion and the human hunter are predators, using the deer for food. You may think that since you are not a hunter, you are not a predator. How about the beef, pork, lamb, fish, or chicken you eat each day? Just because you do not kill the animal yourself does not make you less of a predator. You are still an animal eating the flesh of another animal to satisfy your hunger.

People usually think of predators as meat-eating mammals (carnivores), but the bird, reptile, amphibian, fish, and insect families also have their predators. Eagles, hawks, falcons, and owls are probably the best-known

```
raw_completion_output: |-
  1. source_taxon: mountain lion
        target_taxon: deer
        interaction_type: stalking, killing, and eating

  2. source_taxon: human hunter
        target_taxon: deer
        interaction_type: stalking, shooting, and taking

  3. source_taxon: human
        target_taxon: beef, pork, lamb, fish, chicken
        interaction_type: eating

  4. source_taxon: eagle, hawks, falcons, owls
        target_taxon: small creatures
        interaction_type: swooping down, catching, killing,
tearing flesh

  5. source_taxon: pelicans, gulls, terns, herons, egrets,
kingfishers
        target_taxon: fish
        interaction_type: eating

  6. source_taxon: woodpeckers, flycatchers, warblers,
swallows, swifts, chickadees
        target_taxon: insects
        interaction_type: eating

...
```

This time, we've defined a schema specifically for relationships between taxa.

Why not just tell the LLM to extract all relationships?

https://tpwd.texas.gov/publications/nonpwdpubs/young_naturalist/animals/predator_prey_relationship/index.phtml

Predator-Prey Relationship

In Nature's plan, there is no such thing as a good or bad animal. There are only those animals that kill to eat (predators) and those that are killed and eaten (prey). It is unfair for us to judge their natural survival actions by our own ideas of what is good and bad or to decide that all predators are villains.

When the coyote chases, catches, kills, and eats the cute, furry rabbit, it is not being intentionally cruel. It is only meeting its physical need for food with whatever is available at the time. Its meal could just as easily have been a mouse, rat, or other small creature if one of these other animals had been closer or easier to catch.

Whether it is a roadrunner swallowing a lizard, a blue crab eating a fish, an American Egret catching a perch, a shark investigating a surface disturbance that could be a meal, or a man rounding up a cow for slaughter, the predator kills to provide food for itself or its family.

Is there really any difference between a hungry mountain lion stalking, killing, and eating a deer and a hunter stalking, shooting, and taking a deer home for his family to eat? Both the mountain lion and the human hunter are predators, using the deer for food. You may think that since you are not a hunter, you are not a predator. How about the beef, pork, lamb, fish, or chicken you eat each day? Just because you do not kill the animal yourself does not make you less of a predator. You are still an animal eating the flesh of another animal to satisfy your hunger.

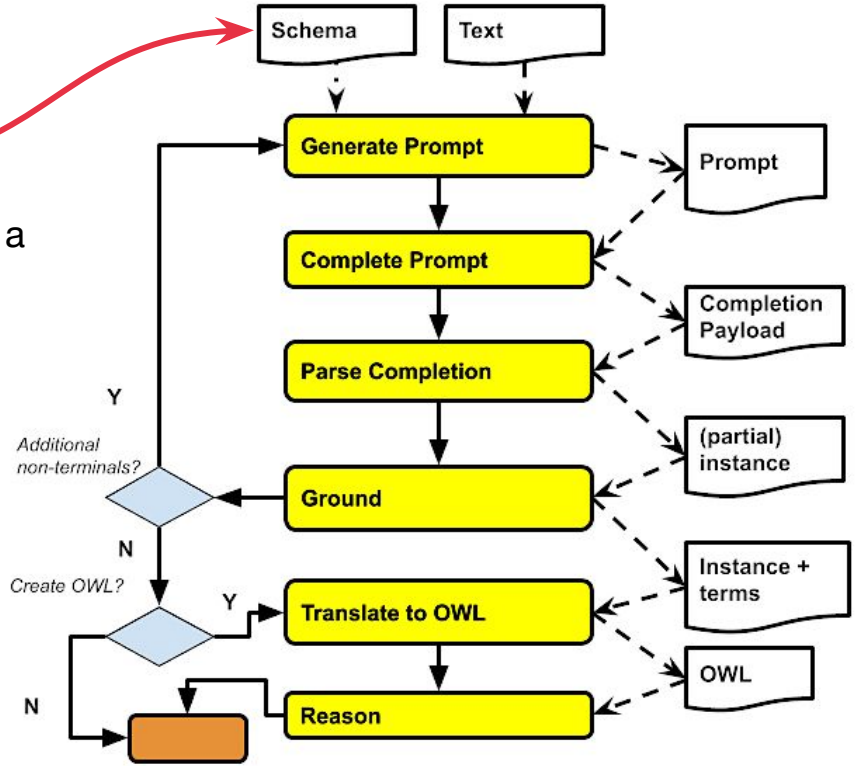
People usually think of predators as meat-eating mammals (carnivores), but the bird, reptile, amphibian, fish, and insect families also have their predators. Eagles, hawks, falcons, and owls are probably the best-known

```
named_entities:
...
  - id: NCBITaxon:9850          Cervidae
      label: deer
  - id: AUTO:human%20hunter
      label: human hunter
  - id: AUTO:beef
      label: beef
  - id: AUTO:pork
      label: pork
  - id: AUTO:lamb
      label: lamb
  - id: NCBITaxon:117565        Class Myxini - not quite the closest option
      label: fish
  - id: NCBITaxon:9031          Gallus gallus
      label: chicken
  - id: AUTO:stalking
      label: stalking
  - id: AUTO:killing
      label: killing
  - id: GO:0007631              "Feeding behavior"
      label: eating
```

Here, we've only used NCBITaxon and Gene Ontology - so we're still missing some domain knowledge, but now have identifiers for concepts.

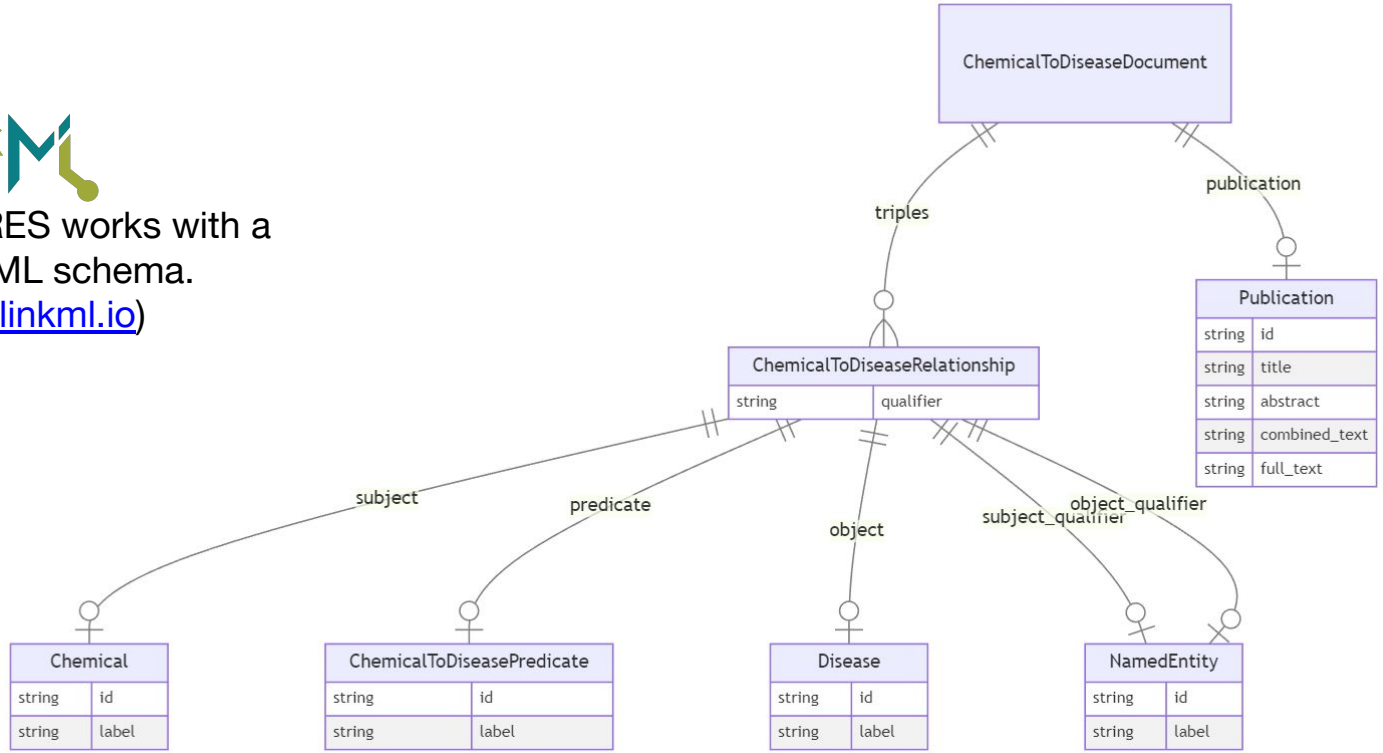
Approach - SPIRES

SPIRES works with a LinkML schema.
(see linkml.io)



Approach - SPIRES

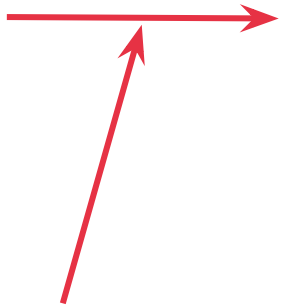
SPIRES works with a LinkML schema.
(see linkml.io)



Approach - SPIRES

A description of β-Catenin's role and interactions with HSV-1
You et al. (2020) J Virol
PMID: 31801859

Title: β-Catenin Is Required for the cGAS/STING Signaling Pathway but Antagonized by the Herpes Simplex Virus 1 US3 Protein
Text: The cGAS/STING-mediated DNA-sensing signaling pathway is crucial for interferon (IFN) production and host antiviral responses. Herpes simplex virus I (HSV-1) is a DNA virus that has evolved multiple strategies to evade host immune responses. Here, we demonstrate that the highly conserved β-catenin protein in the Wnt signaling pathway is an important factor to enhance the transcription of type I interferon (IFN-I) in the cGAS/STING signaling pathway,...



GO-Causal Activity Model
GO-CAM
Template

```
extracted_object:
  genes:
        - HGNC:2514
        - HGNC:10420
        - HGNC:5417
        - AUTO:ISG
  organisms:
        - NCBITaxon:10298
  gene_organisms:
        - gene: HGNC:10420
        organism: NCBITaxon:10298
  activities:
        - GO:0006351
        - GO:0016301
        - AUTO:replication
        - GO:0048151
        - GO:0051170
        - AUTO:inhibition
  gene_functions:
        - gene: HGNC:2514
        - gene: HGNC:10420
  cellular_processes:
        - AUTO:IFN%20production
        - GO:0051607
        - GO:0006955
        - GO:0045087
        - GO:0016032
  pathways:
        - GO:0140896
        - GO:0016055
```

Approach - SPIRES

Figure legend from a study of the hyporheic zone
Nelson et al. (2020) PLoS ONE
PMID: 31986180

Sediment communities from the hyporheic zone of the Columbia River along the Hanford Reach were sampled from April 30, 2014 to November 25, 2014, using sand packs deployed at three equivalent hyporheic zone locations …

A template for environmental sample metadata

```
extracted_object:
  location:
        - GAZ:00167673
  environmental_material:
        - ENVTHES:20899
  environments:
        - ENVTHES:21903
  variables:
        - ENVO:00002006
named_entities:
  - id: GAZ:00167673
        label: Hanford Reach
  - id: ENVTHES:20899
        label: sediment, sand, dissolved organic
carbon (NPOC), nitrate, DO concentration,
        water temperature
  - id: ENVTHES:21903
        label: hyporheic zone, river
  - id: ENVO:00002006
        label: water chemistry data, hydraulic
regime, influx of surface water, dissolved
        organic carbon (NPOC) levels, nitrate
concentrations, DO concentration, water
        temperature
```

Photo from https://commons.wikimedia.org/wiki/File:East_River_(northern_Gunnison_County,_Colorado,_USA)_(46220745984).jpg

Approach - SPIRES

Requirements: Python 3.9 or greater.
The *poetry* dependency toolkit for Python.
The **Ontology Access Kit** (https://github.com/INCATools/ontology-access-kit).
An OpenAI API key.

Installation: Clone the repository (https://github.com/monarch-initiative/ontogpt)
poetry install
poetry run runoak set-apikey -e openai <your openai api key>

To run and test: *poetry run ontogpt extract -t mendelian_disease.MendelianDisease -i tests/input/cases/mendelian-disease-sly.txt*

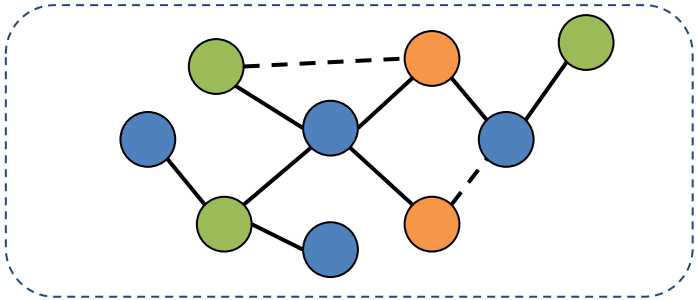
OntoGPT will download the necessary ontologies.

Next steps

BioSciences

See our preprint: https://arxiv.org/abs/2304.02711

- Automated KG assembly
  - Extract relations from literature,
  - Then integrate with those defined in knowledge bases,
  - And hierarchical relationships from ontologies,
  - And add predicted relationships
- Dealing with limitations
  - Reducing dependence on OpenAI
  - Avoiding hallucinations
- Improving ID mappings
- Gene enrichment analysis (SPINDOCTOR)
- Broad literature extraction
  - e.g., from PMC full-texts



OntoGPT/SPIRES performs just slightly worse than the average F1 score on the BioCreative V Chemical-Disease Relationship (CDR) task…though it requires no training or fine-tuning.

Improving integration with AgBioData resources

OntoGPT uses the Ontology Access Kit (OAK) for its annotators and grounders.

OAK works best with ontologies from the OBO Foundry and Bioportal.

To support use cases involving AgBioData, we can:

- Use Agroportal
- Extract plant strains and genomes by name
- Extract livestock traits and breeds
    - Vertebrate Breed Ontology (VBO) does some of this…but it's a challenge to capture general names consistently
- Other use cases?

Thank You

Email:
jhc@lbl.gov

Github:
@caufieldjh

BBOP@LBL:
Chris Mungall (PI)
Seth Carbon
Nomi Harris
Harshad Hegde
Marcin Joachimiak
Patrick Kalita
Mark Miller
Sierra Moxon
Sujay Patil
Justin Reese

Vincent Emonet (Maastricht Univ)
Nico Matentzoglu (Semanticly)
HyeongSik Kim (Bosch Research)
Melissa Haendel (Univ Colorado)
Peter Robinson (JAX)

Icons % the Noun Project (Made x Made; Rivercon)