

# **Overview of free-text to concept conversion for semantic search**

**projects**

**most with a focus on finding EMRs,  
for clinical-trial cohort selection ...**

**I ILLINOIS**

NCSA | National Center for  
Supercomputing Applications

# As a Knowledge-Engineer / Research-Programmer, I have had a few related projects:

- Early concept based search work:
  - Brightware/Mindbox.com: company split in two to focus on this (no visuals for this)
    - Helped design & made first install of automated email answering system used at a national scale
    - CBR match of concepts, with rules that matched on the concept hierarchy
    - Returned associated templates filled with info for the person and type of info on what queried
- Later biomedical concept based search work:
  - rctbank.UCSF.edu:
    - A Practical Method for Transforming Free-Text Eligibility Criteria into Computable Criteria
      - Focused on free-text to the query, to start
  - AlohaHealth.net: Matching subjects to study criteria
    - UCSF contact started his second clinical-trials based company
    - We do the concept based search, but focus using weights vs query logic
  - nca.UIUC.edu pilot study to improve nlm.nih.gov's SemRep from PI now at the iSchool
    - follow on and other-work around data management framework with flexible metadata
      - used for faceted search of metadata, discovery & matching for use
        - which could still benefit from NER/dedup/etc

*method*

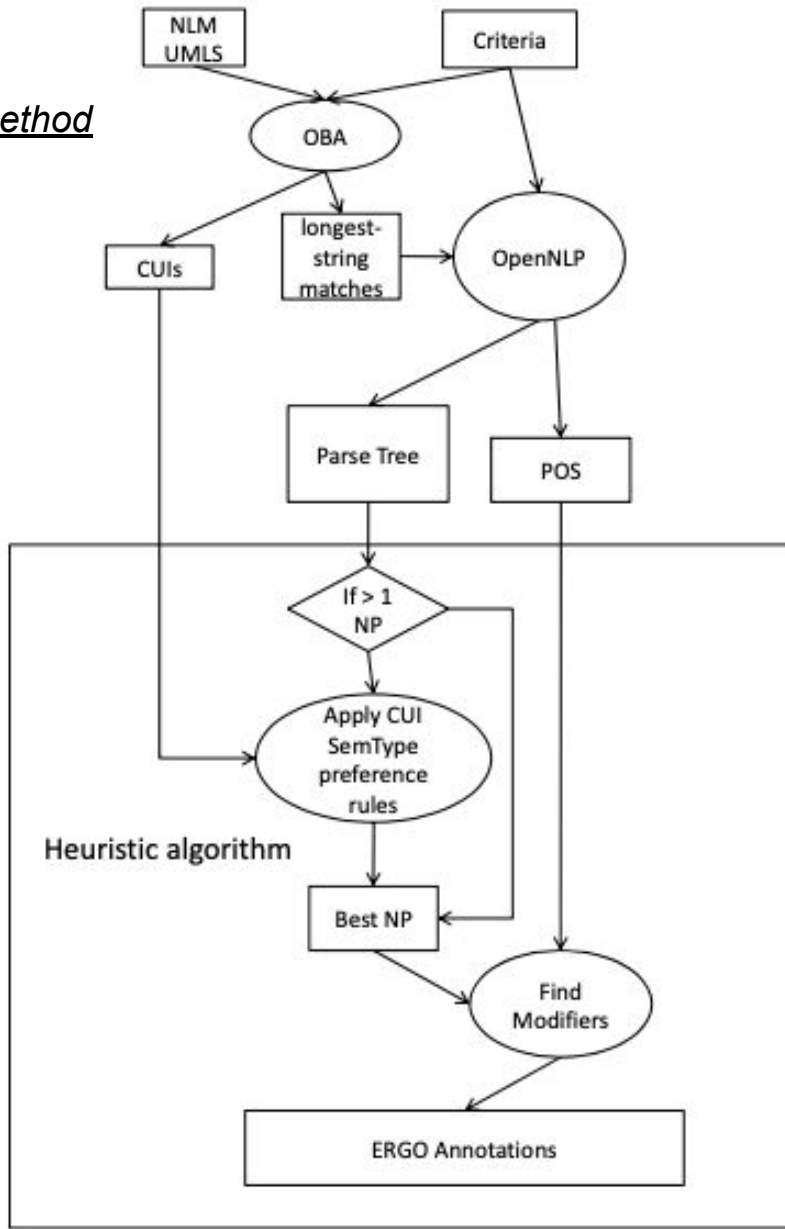


Figure 4. Steps in automated generation of ERGO Annotations.

# A Practical Method for Transforming Free-Text Eligibility Criteria into Computable Criteria

UCSF & Stanford-BiomedInfoResearch



*computable form*

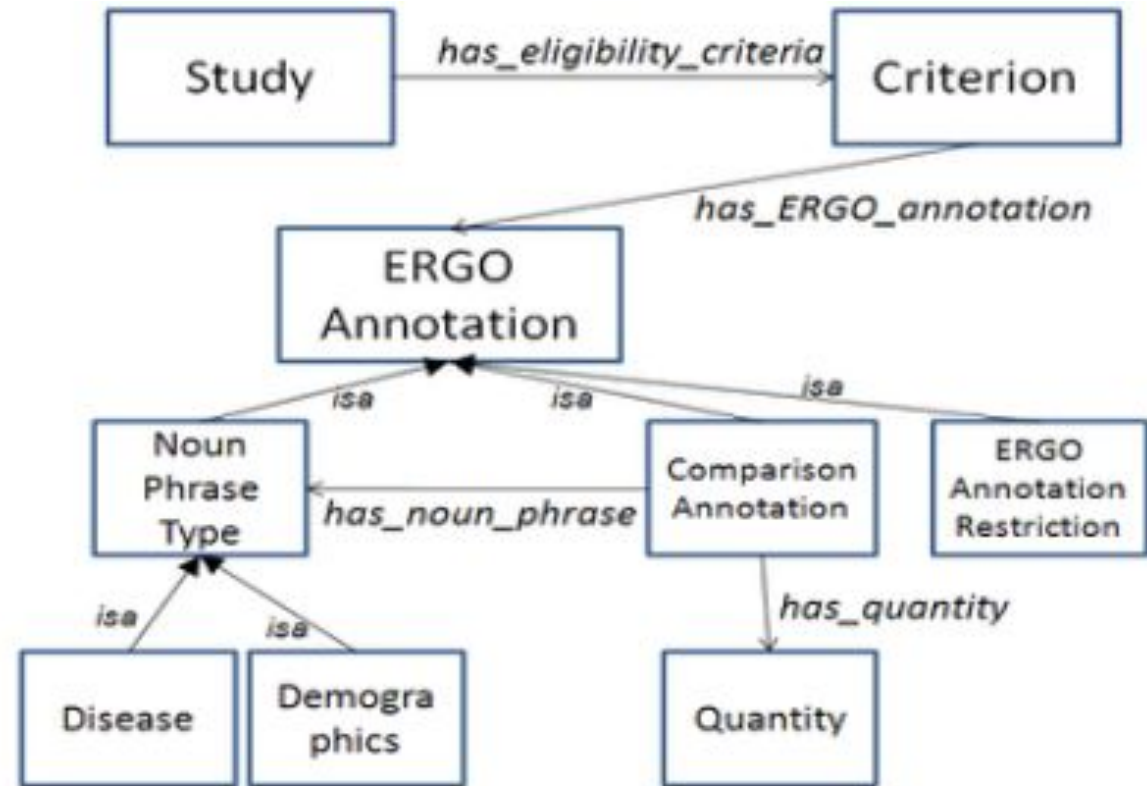
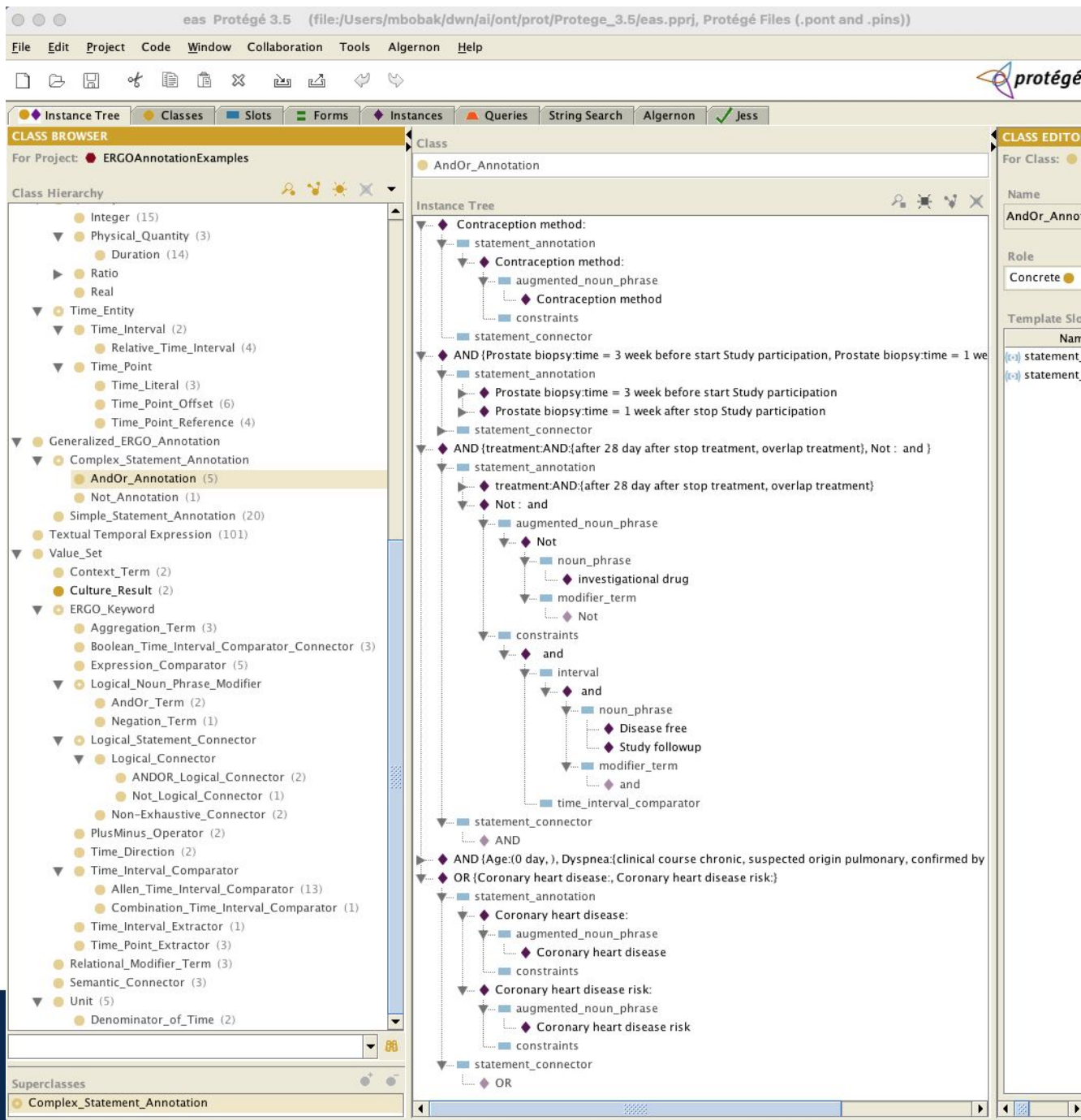
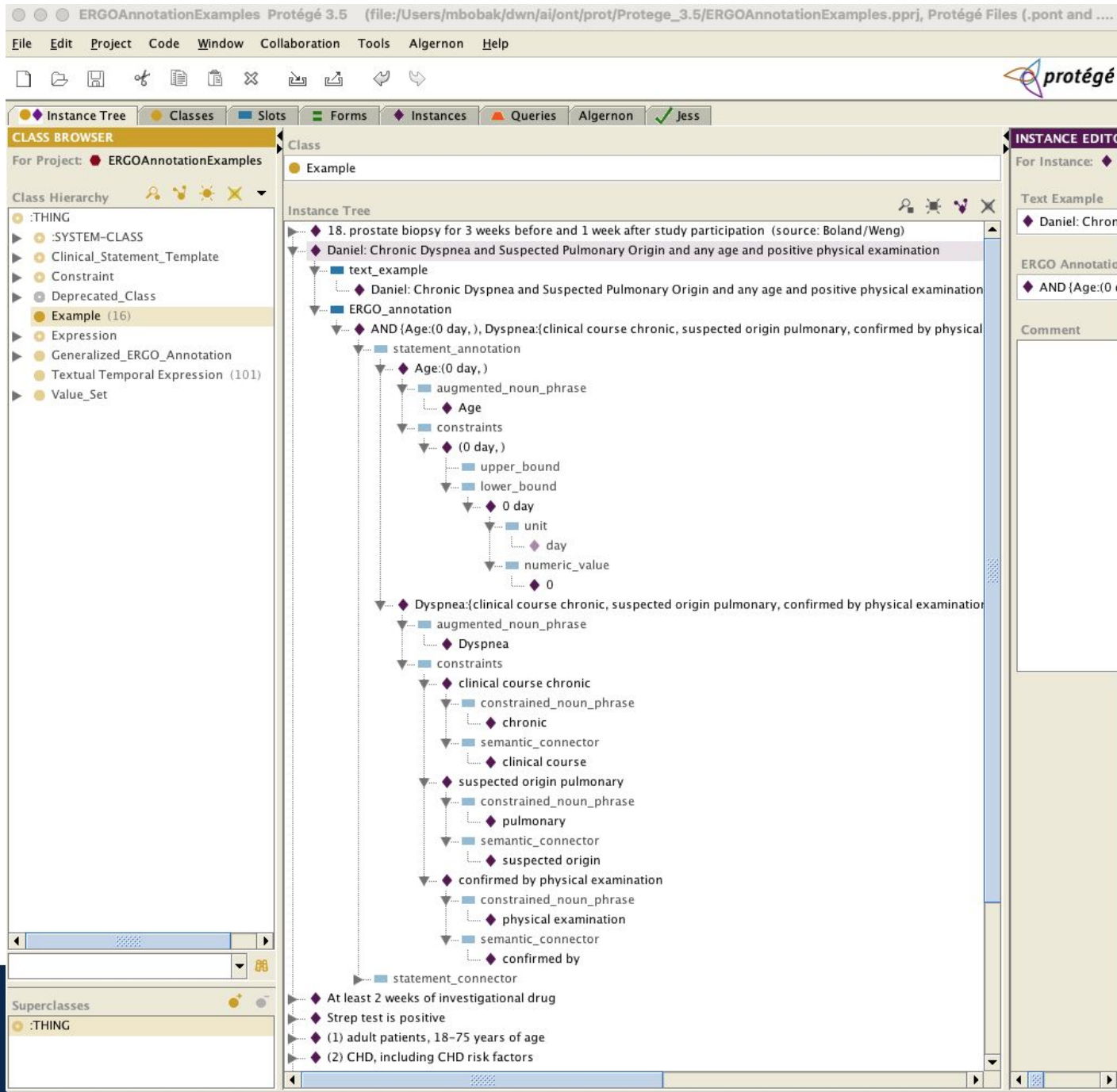


Figure 5. Predefined OWL ontology to illustrate how ERGO Annotations may be used to classify criteria and to search for.



# The Eligibility Rule Grammar and Ontology (ERGO)

- Input text transformed to connected instances
- These in/ex-clusion descriptions turned to queries
  - That would need annotated EMRs for search
- I iteratively worked on the code for the algorithm
  - comparing it's likely instance matches
  - with the hand scoring of the parts
  - till we could get most of the 1k statements



Examples using Protege 3.5 at: [sites.google.com/site/humanstudyome/home/ergo](https://sites.google.com/site/humanstudyome/home/ergo) of concept annotated in/ex-clusion criteria

- Learned many things along the way
  - enjoyed quality of NLM's metemap
  - needed some pre/post processing
- for the follow on searching tagged EMRs
  - importance of types of concept matches
    - by order of importance &/or
    - weighting the concept matches
- Pitched tagging EMR's so that radiologist could get feedback on their diagnosis.
  - had some i2b2 work on it
  - kept up with the datacenter group lead
    - who is the CEO of our startup
    - where we search for trail cohorts

# Explore relationships between criteria, sites and patient data

Phenotyping automates subject pool selection by providing accurate insights into qualified patients available at specific clinical sites.

add  
from  
EMRs

## Patient Phenotypes

Concept IDs representing a patient's medical history.

had

## Study Criteria

Inclusion Criteria  
Exclusion Criteria  
Trial description, etc.



## Matched Subjects

Patient phenotypes are matched and scored against weighted study criteria to create an aggregate score for each candidate subject.

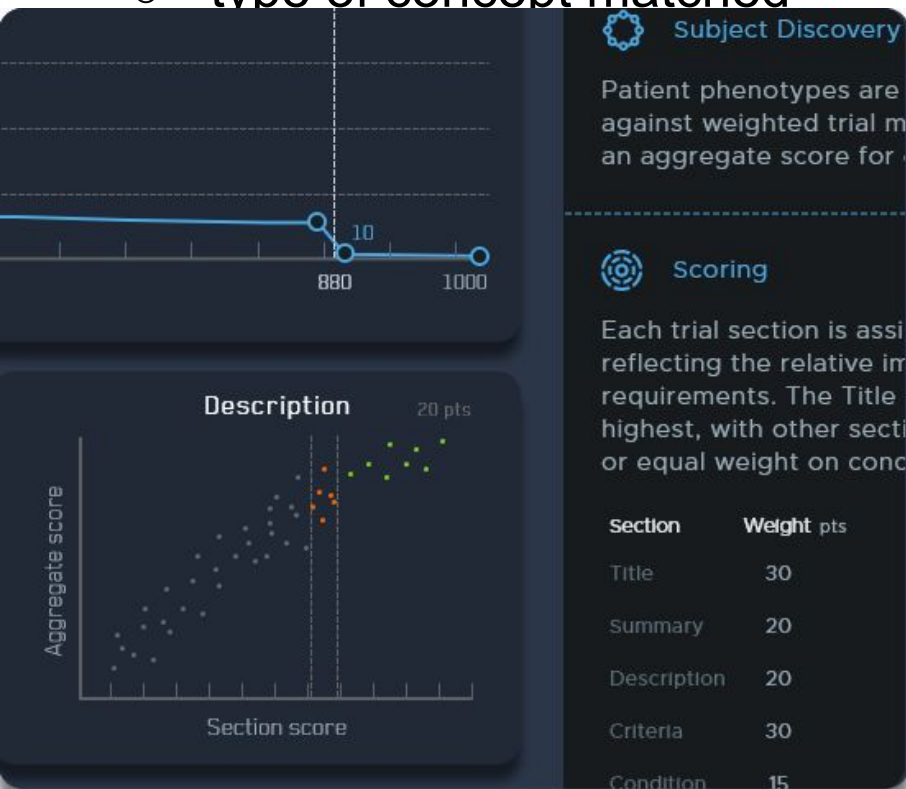
Iteration on trial criteria weights allows for a wider discovery of candidates.

In startup  
with friend  
from  
UCSF  
Rob  
Wynden  
to  
continue  
this work.

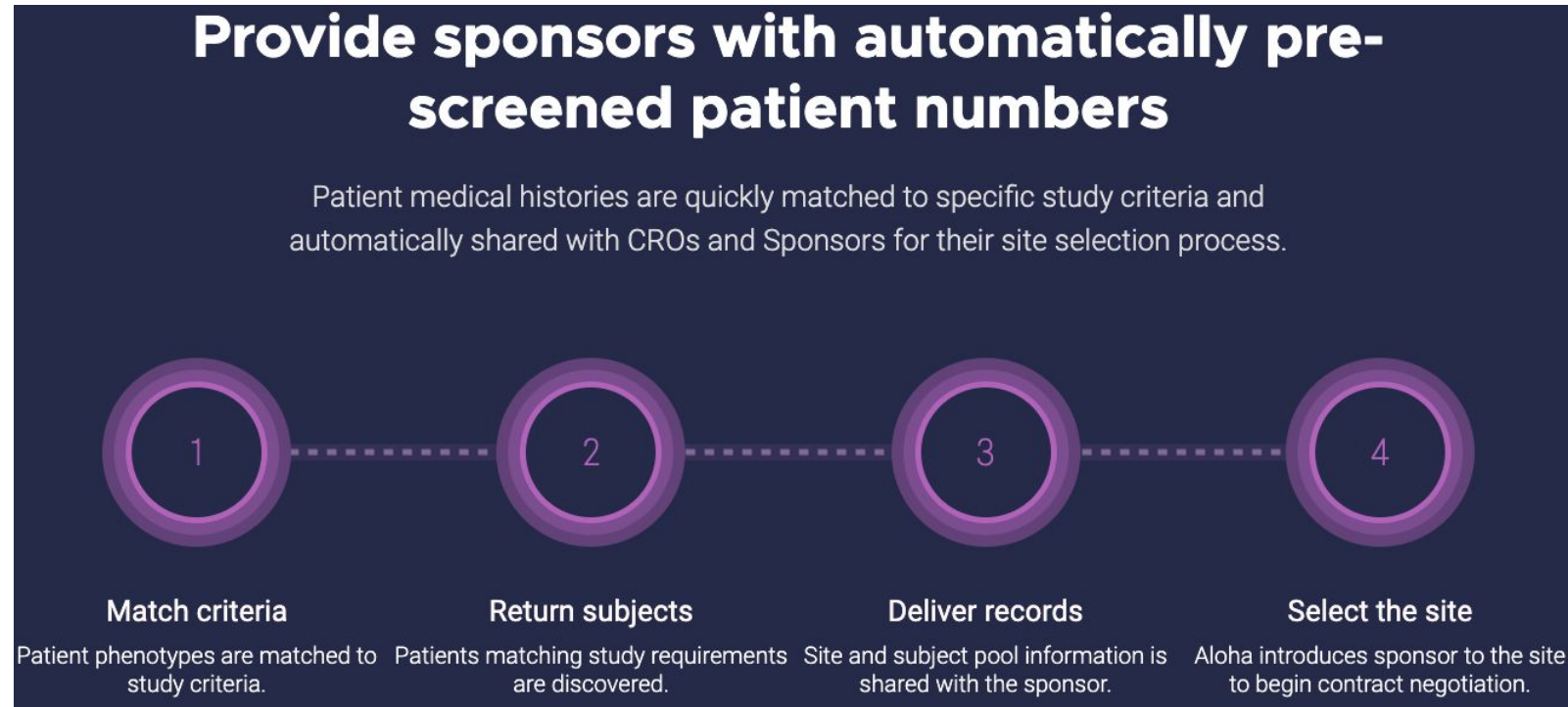
It's called

Aloha  
Health  
.net

- Score annotation matches of
  - study criterion against
  - a set of EMRs at a site
  - to return a cohort
- Patient's concept match weighted
  - by the part of the study/ EMR that the concept comes from
  - type of concept matched

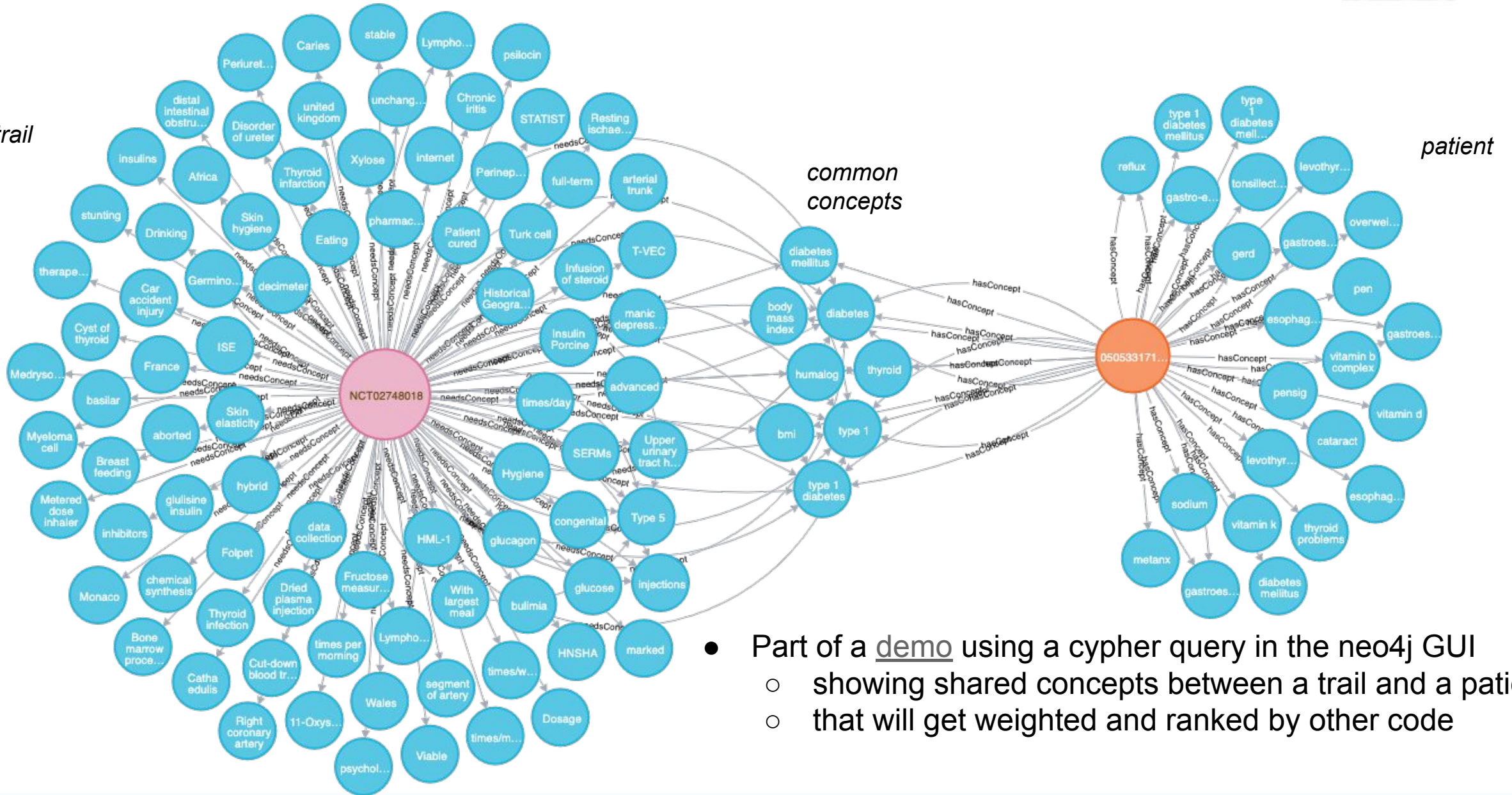


- Less ERGO like annotation logic, and more weighted concept sets



- We have python code for the annotation and matching score
- along with, first SPARQL then neo4j's cypher query abilities
  - where you can interactively explore a site's patients

trail



common  
concepts

patient

- Part of a demo using a cypher query in the neo4j GUI
  - showing shared concepts between a trail and a patient
  - that will get weighted and ranked by other code



- TOOLS
- Terms of Service
  - Batch Access to Tools
  - Interactive Access to Tools
  - Web API Access
  - Medical Text Indexer (MTI)
  - Phrase2MeSH
  - MeSH on Demand (MeSH link)
  - MetaMap
  - MetaMap Lite
  - Custom Taxonomy Builder
  - MTI ML (Machine Learning Package)
  - SPECIALIST Lexicon Information and Tools

TOOLS

## SemRep

SemRep is a UMLS-based program that extracts three-part propositions, called semantic predications, from sentences in biomedical text. Predications consist of a subject argument, an object argument, and the relation that binds them. For example, from the sentence in (1), SemRep extracts the predications in (2).

1. We used hemofiltration to treat a patient with digoxin overdose that was complicated by refractory hyperkalemia.
2. Hemofiltration-TREATS-Patients  
Digoxin overdose-PROCESS\_OF-Patients  
hyperkalemia-COMPLICATES-Digoxin overdose  
Hemofiltration-TREATS(INFER)-Digoxin overdose

The subject and object arguments of each predication are concepts from the UMLS Metathesaurus and their binding relationship (in uppercase) is a relation from the UMLS Semantic Network. For a detailed description of SemRep, see [1,2].

Holders of a UMLS license can run SemRep interactively or in batch mode using the SKR Scheduler. SemRep is also available as a stand-alone program on the Linux platform.

## References

1. Kilicoglu H, Rosemblat G, Fiszman M, Shin D. [Broad-coverage biomedical relation extraction with SemRep](#). BMC Bioinformatics 2020;21:1-28.

Got pilot grant to help ex NLM staff now with a lab at the iSchool, who is continuing the SemRep work

SemRep goes beyond NER of the entities to include finding the relationships between them

which can be viewed in the brat rapid annotation tool

The screenshot displays the brat rapid annotation tool interface. The browser address bar shows the URL: `lhce-brat.nlm.nih.gov/index.xhtml#/SKR/Factuality/Rec.../SKR/Factuality/Reconcile_50/10048494`. The tool shows eight sentences, each with various entities and relationships annotated. The entities are represented by colored boxes: pink for food-related terms (fndg), purple for clinical terms (clna), green for physiological terms (ftcn, qnco, tmco, orch, nsba, aapp, spco, b poc, idcn), blue for food/diet terms (food), yellow for hormone terms (horm, chvs), and cyan for other terms (podg, orga). Relationships are shown as red boxes labeled "(FA) PROCESS OF" and purple boxes labeled "(FA) TREATS". Arrows indicate the direction of the relationships, such as "Subject" and "Object".

1 Dietary salt intake, blood pressure and the kidney in hypertensive patients with non-insulin dependent diabetes mellitus.

2 The mechanisms responsible for hypertension in NIDDM patients are only partially understood.

3 Increased sensitivity to dietary salt intake and to vasoconstrictor hormones are among the mechanisms proposed.

4 We have studied 19 hypertensive NIDDM patients 7 salt-sensitive and 12 salt-resistant while they were ingesting a diet with 20 mEq/day of Na<sup>+</sup> for 9 days and while they were ingesting a diet containing 250 mEq/day of Na<sup>+</sup> for 14 days.

5 During the last 4 days of each dietary regimen, they received 60 mg/day of slow-release nifedipine.

6 Blood pressure response to increasing doses of norepinephrine and angiotensin II was studied at the end of each of the four phases of the study.

7 High salt intake increased blood pressure and decreased heart rate in these patients.

8 High salt intake also increased the vascular response to norepinephrine but not to angiotensin II in NIDDM hypertensive subjects.

# Many applications: SemRep annotated MEDLINE

## Access to SemRep/SemMedDB/SKR Resources

The SKR project maintains a database of 96.3 million [SemRep](#) predications extracted from all MEDLINE citations. This database supports the [Semantic MEDLINE web application](#), which integrates PubMed searching, SemRep predications, automatic summarization, and data visualization. The application is intended to help users manage the results of PubMed searches. Output is visualized as an informative graph with links to the original MEDLINE citations.

To access any of the SemRep/SemMedDB/SKR Data Sets or the SemMedDB Database, users must have accepted the terms of the [UMLS Metathesaurus License Agreement](#), which requires users to respect the copyrights of the constituent vocabularies and to file a brief annual report on their use of the UMLS. Users must also have activated a [UMLS Terminology Services \(UTS\)](#) account. For information on how to use UTS authentication, please click [here](#).

For details of the licenses, please see the [UMLS Metathesaurus License Agreement](#) and [How to License and Access the Unified Medical Language System \(UMLS\) Data](#).

### SemRep Source Code



### Semantic MEDLINE Database (SemMedDB)



The Semantic MEDLINE Database (SemMedDB) is a repository of semantic predications (subject-predicate-object triples) extracted by SemRep, a semantic interpreter of biomedical text. SemMedDB currently contains information about approximately 96.3 million predications from all of PubMed citations (about 29.1 million citations) and forms the backbone of the [Semantic MEDLINE application](#).

For details about the SemMedDB schema, click [here](#).

To Download the SemMedDB Database click [here](#).

To learn more about Semantic Medline click [here](#).

## The follow on for the pilot: RCTCheck LM Model+Clowder data management

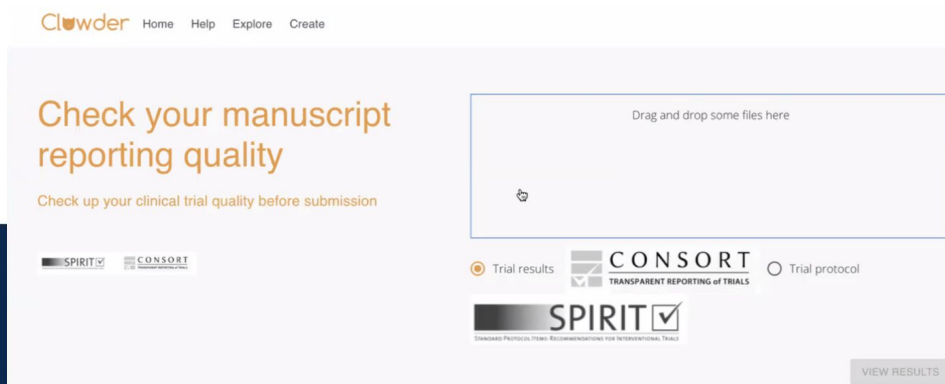


### Randomized Controlled Trials (RCT)s

- Can suffer from poor reporting quality
- Problems with design, execution, or reporting of the trial process can lead to unreliable finding, excessive cosad, and potentially harm to patients

CONSORT: Consolidating Standard of Reporting Trials  
SPIRIT: Standard Protocol Items: Recommendations for Interventional Trials

To help journals enforce/verify: LM Model starting with PubMedBERT, trained on a PubmMed dataset



## Biomed free-text conceptual annotation, applications:

UCSF: to make a conceptual query; AHN adds tagged patients; UIUC relationship tagging

UCSF: Annotated in/ex-clusion criteria, but the connection logic of the query was not fully automated

- I used MMTx & Metamap, and got the source to be posted at NLM; So I could more easily alter the algo
- Didn't use early SemRep; Started with NLP libs to get the Noun\_Phrases, modifiers/connection/etc.

Aloha Health: has looser concept connections, but includes patients, and contextual weights

- Use our own code for UMLS (SNOMED/Radlex/..) annotations of the criterion and EMRs and matching
- I would like to get back to extending open algorithms/code, on a live data warehouse

UIUC: easier to use SemRep allows for easier text to Knowledge-Graph, and maybe structured queries

- There is some of the easier code-base and the related NER extensions that I would consider using now
- The pilot grant did go forward using the data management [clowderframework.org](http://clowderframework.org): RCTCheck
  - I extended the framework with the ability to make the datasets discoverable
  - Also used it for a PoC for GeoCODES data & tool: discovery, matching & use; informing it's V2

UIC: Hoping to learn more about the potential range of the role today



Questions (now)

&

I have some more slides that I made after the pilot grant  
that I could go through

and

have more on FAIR (meta)data storage, search, and  
matching for use

in other slide sets too

The pilot grant did go forward using the clowder-framework; which I extended to make  
it's datasets more discoverable, and could benefit from another FAIR dataset  
discovery & use application of mine as well

# NCSA faculty fellowship (pilot grant) with iSchool on turning free-text into Knowledge-Graph triples

Mike Bobak

**I** ILLINOIS

NCSA | National Center for  
Supercomputing Applications

# NCSA faculty fellowship with iSchool 2021-2022

- Takes free-text to Knowledge-Graph triples (entities & relationships between them)
- Takes work of the professor from nlm.nih SemRep and get an easier to maintain port
- Started in a collection of languages incl. Prolog, then Java port, now in Python
- Has already helped in putting in for a NIH grant to take the work even further
- Makes use of NLM's MetaMap-Lite (MML) which does the Named-Entity-Recognition
- Then sets of rules are used to find relationships between the entities
- MML matching ability generated from any ontology, with synonyms in each class
- Also an aim to make it easier to generalize beyond the biomedical domain

I worked on:

- Getting the java then python code bases running on a new machine, update everything to python3
- Started some simple logging, to: catch errors, test for changes in output  
incl. some in brat format to more easily view the parse/relationships within the sentences
- Move away from socketed connections to either local calls or REST based service calls  
or  
Move services either to REST based calls, or to local execution.
- Updated process to pull synonym references from ontologies for NER in other domains
  - Updated python code to produce datafilebuilder input and run that into metamap
  - also found a simple python library to pull then match from an ontology
- Use of owlready2.pymedtermino2 for concept relationship [/ subsumption] tests
- Some looking at further work
  - List of next steps / use in possible grants, one of which is now active using clowder storage



**Motivation: of machine interpretability of knowledge from free-text**

**Things-not-strings** via: free-text -to-> Knowledge-Graph triples (entities w/relationships) helps achieve the goal of machine-interpretability [KGs need connected things]

[blog.google/products/search/introducing-knowledge-graph-things-not-strings](http://blog.google/products/search/introducing-knowledge-graph-things-not-strings)

# Introducing the Knowledge Graph: things, not strings

1. Find the right thing Language can be ambiguous
2. Get the best summary With the Knowledge Graph, Google can better understand your query
3. Go deeper and broader

Finally, the part that's the most fun of all—the Knowledge Graph can help you make some unexpected discoveries.

# Metadata for Machines (M4M)

There are several application areas for machine interpretable knowledge

e.g.



Short [workshops](#) that create high-priority machine-actionable metadata for the specific needs of particular communities of practice.



# Named-Entity-Recognition & Linking

“Paris is the capital of France”



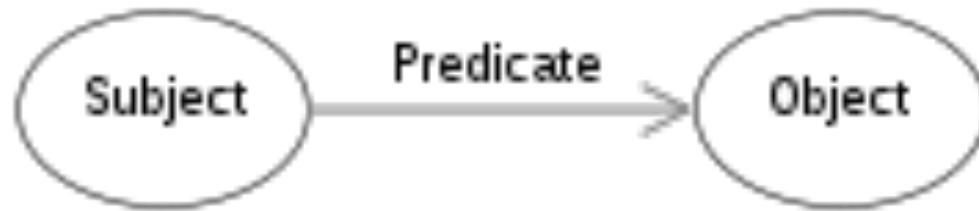
wikipedia.org/wiki/Paris

wikipedia.org/wiki/Capital\_city\_of



wikipedia.org/wiki/France

Knowledge-Graph triples are made of URI/things,  
w/some literal objects



wikipedia.org/wiki/France

wikipedia.org/wiki/Capital\_city

wikipedia.org/wiki/Paris

literals are eg. text numbers, or any xml type; but can only be in terminal Objects  
dbp:Paris dbp:Population 2161000^^xsd:int

# We use MetaMap-Lite for Entity-Linking

## How it works:

- `input text ->`
- `sentence/line segmentation ->      tokenization ->      part-of-speech tagging ->`
- `token window generation ->                      term normalization ->`
- `concept dictionary lookup ->`
- `negation detection ->`
- `result presentation`

# Example MML match:

```
"Papillary Thyroid Carcinoma is a Unique Clinical Entity"  
  "Papillary Thyroid Carcinoma is a Unique Clinical"  
  "Papillary Thyroid Carcinoma is a Unique"  
  "Papillary Thyroid Carcinoma is a"  
  "Papillary Thyroid Carcinoma is"  
  "Papillary Thyroid Carcinoma" --> match  
    "is a Unique Clinical Entity"  
    "is a Unique Clinical"  
    "is a Unique"  
    "is a"  
    "is"  
      "a Unique Clinical Entity"  
      "a Unique Clinical"  
      "a Unique"  
      "a"  
        "Unique Clinical Entity"  
        "Unique Clinical"  
        "Unique" --> match  
          "Clinical Entity"  
          "Clinical" --> match  
            "Entity" --> match
```

# Entity Linking output to the brat rapid annotation tool

1 Dietary salt intake, blood pressure and the kidney in hypertensive patients with non-insulin dependent diabetes mellitus.

2 The mechanisms responsible for hypertension in NIDDM patients are only partially understood.

3 Increased sensitivity to dietary salt intake and to vasoconstrictor hormones are among the mechanisms proposed.

4 We have studied 19 hypertensive NIDDM patients 7 salt-sensitive and 12 salt-resistant while they were ingesting a diet with 20 mEq/day of Na<sup>+</sup> for 9 days and while they were ingesting a diet containing 250 mEq/day of Na<sup>+</sup> for 14 days.

5 During the last 4 days of each dietary regimen, they received 60 mg/day of slow-release nifedipine.

6 Blood pressure response to increasing doses of norepinephrine and angiotensin II was studied at the end of each of the four phases of the study.

7 High salt intake increased blood pressure and decreased heart rate in these patients.

8 High salt intake also increased the vascular response to norepinephrine but not to angiotensin II in NIDDM hypertensive subjects.

# Expanding Beyond BioMedical domain

Ontologies with predicate *hasExactSynonym*,  
w/literal objects being that text that can be harvested  
to make MML handle new domains.

I plan to use it for GeoCODES, & can think of many others it could be used in



- Get the java then python code bases running on a new machine, update everything to python3
- Start some simple logging, suggest use to catch errors, test for changes in output  
incl some in brat to more easily view the parse/relationships within the sentences
- Move away from socketed connections to either local calls or REST based service calls.
- Update process to pull synonym references from ontologies for NER in other domains
- Use of owlready2.pymedtermino2 for concept relationship tests

<https://isda.ncsa.illinois.edu/~mbobak/>

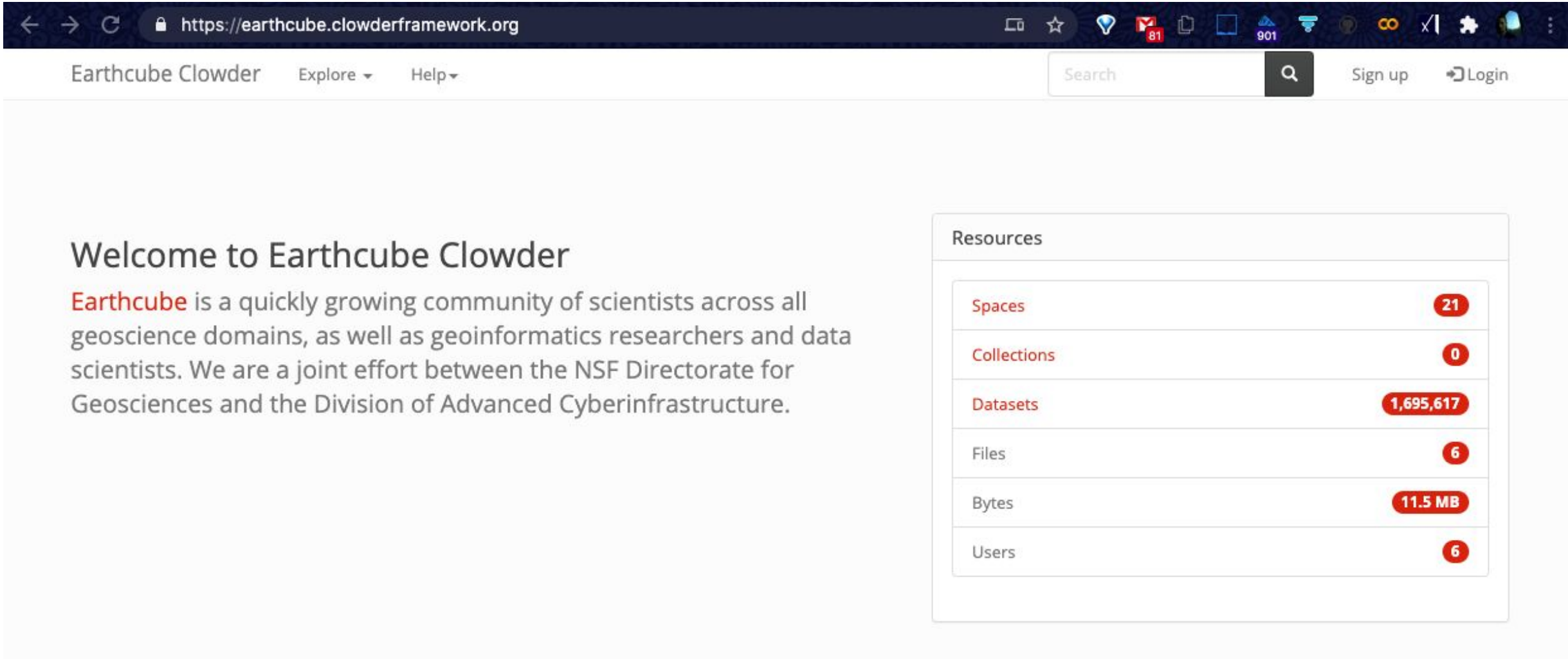
for February-June:

- Process/documentation for regular UMLS updates
  - Metamorphosys
  - Can we rely on MetaMap Lite files?
- Process/documentation for adapting MetaMap Lite to non-UMLS vocabularies/ontologies
  - What is required in the vocabulary/ontology? What is good-to-have?
  - Data File Builder
  - Tips/tricks
- Overall infrastructure
  - Should we consider running MetaMap Lite and other server processes in a different way?
  - Logging
  - Unit tests
  - Serialization/deserialization

after this, extra slides, this is just a very rough, 1st draft

gives you some feel of possible software reuse, and some of my other more recent projects

Clowder is used in the pilot follow on NIH grant & I will annotate this EC free-text too



The screenshot shows the Earthcube Clowder website. The browser address bar displays <https://earthcube.clowderframework.org>. The navigation bar includes "Earthcube Clowder", "Explore", "Help", a search bar, "Sign up", and "Login".

## Welcome to Earthcube Clowder

**Earthcube** is a quickly growing community of scientists across all geoscience domains, as well as geoinformatics researchers and data scientists. We are a joint effort between the NSF Directorate for Geosciences and the Division of Advanced Cyberinfrastructure.

Resources	
Spaces	21
Collections	0
Datasets	1,695,617
Files	6
Bytes	11.5 MB
Users	6



### opentopography

High-Resolution Topography Data and Tools

0 0 1

### MagIC

Magnetics Information Consortium (MagIC) Promoting Information technology Infrastructures for the International paleomagnetic, geomagnetic and rock magnetic community.

4136 0 1

### opencoredata

Open Core Data is an implementation of the RDA Digital Object Cloud. Open Core Data contains digital objects from the continental and ocena drilling research projects funded by the National Science Foundation. These objects are described using the structured data on the web patterns pro...

18171 0 1



### ssdb.iodp

The Site Survey Data Bank (SSDB) is a repository for site survey data submitted in support of International Ocean Discovery Program (IODP) proposals and expeditions. SSDB serves different roles for different sets of users

5344 0 1



### neotomadb

Neotoma Paleoecology Database and Community is an online hub for data, research, education, and discussion about paleoenvironments. Anyone with an Internet connection can access Neotoma.

11955 0 1



### ucar

OpenSky is the home for NCAR/UCAR research and historical materials as well as other collections



### unavco

Transforming understanding of Earth systems and hazards using geodesy.

5086 0 1



### hydroshare

HydroShare is CUAHSI's online collaboration environment for sharing data, models, and code.

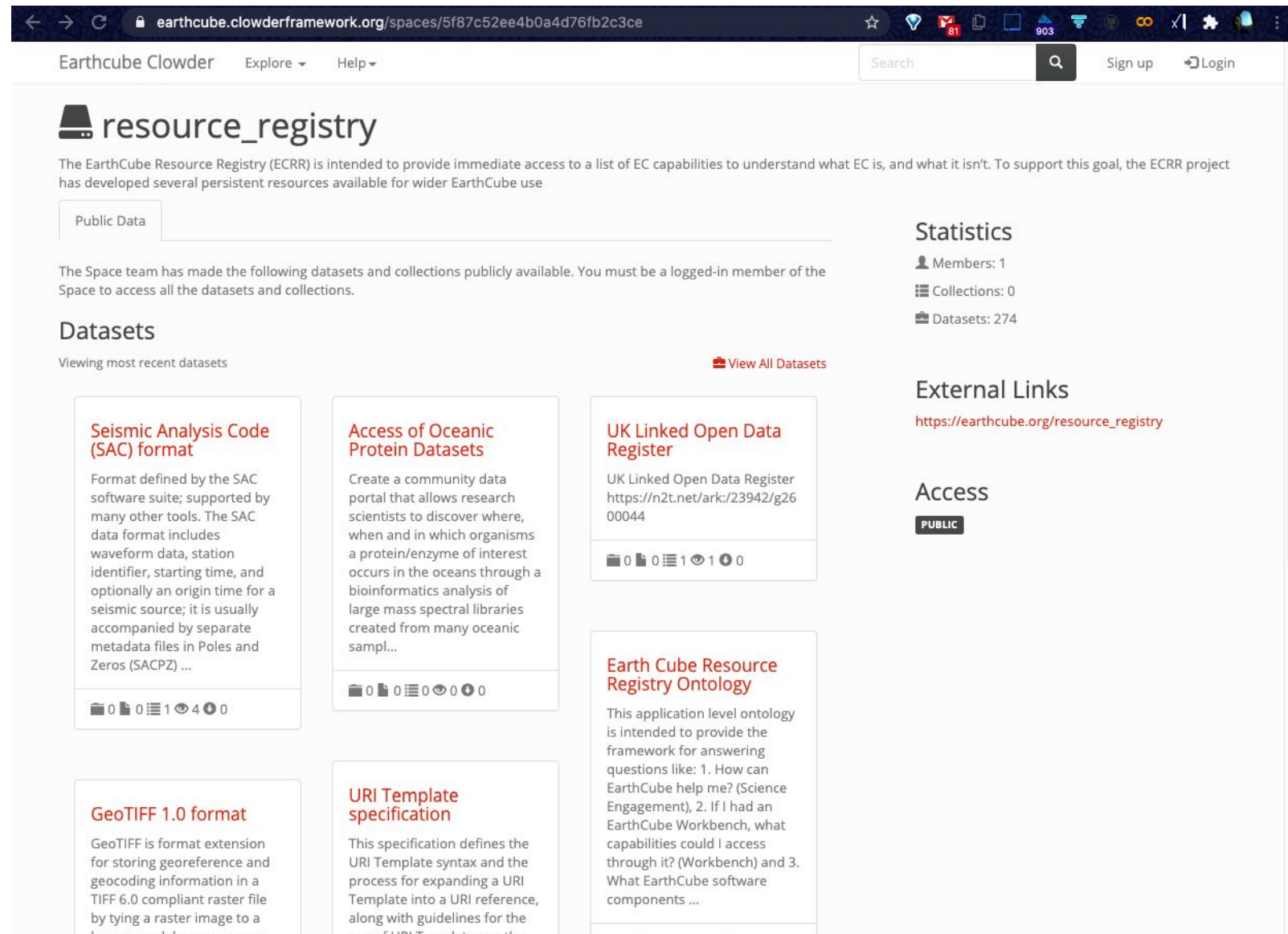
4185 0 1

## Clowder organization

- One *space* per data-facility
- *Datasets* hold metadata
- Also a Resources space:

## Allows for

- dataset & tool search
- metadata/annotation
- linking out to get the data
- & sometimes (assoc) tool/s



The screenshot shows the Earthcube Clowder interface for the 'resource\_registry' space. The page title is 'resource\_registry' and it includes a search bar and navigation links. The main content area is titled 'Public Data' and contains a description of the EarthCube Resource Registry (ECRR) project. Below this, there is a section for 'Datasets' with a 'View All Datasets' link. The datasets are displayed in a grid of cards, each with a title, description, and a small icon set at the bottom. The datasets shown are:

- Seismic Analysis Code (SAC) format**: Format defined by the SAC software suite; supported by many other tools. The SAC data format includes waveform data, station identifier, starting time, and optionally an origin time for a seismic source; it is usually accompanied by separate metadata files in Poles and Zeros (SACPZ) ...
- Access of Oceanic Protein Datasets**: Create a community data portal that allows research scientists to discover where, when and in which organisms a protein/enzyme of interest occurs in the oceans through a bioinformatics analysis of large mass spectral libraries created from many oceanic sampl...
- UK Linked Open Data Register**: UK Linked Open Data Register <https://n2t.net/ark:/23942/g2600044>
- Earth Cube Resource Registry Ontology**: This application level ontology is intended to provide the framework for answering questions like: 1. How can EarthCube help me? (Science Engagement), 2. If I had an EarthCube Workbench, what capabilities could I access through it? (Workbench) and 3. What EarthCube software components ...
- GeoTIFF 1.0 format**: GeoTIFF is format extension for storing georeference and geocoding information in a TIFF 6.0 compliant raster file by tying a raster image to a known model space or map
- URI Template specification**: This specification defines the URI Template syntax and the process for expanding a URI Template into a URI reference, along with guidelines for the use of URI Templates on the

On the right side of the page, there are sections for 'Statistics' (Members: 1, Collections: 0, Datasets: 274), 'External Links' (https://earthcube.org/resource\_registry), and 'Access' (PUBLIC).

# Clowder search results

# & a result's metadata(tab) tree listing





The screenshot shows the Earthcube Clowder search interface. The search bar contains the word "carbon". Below the search bar, there are four search results, each with a red briefcase icon, a title, a timestamp, and a description. The first result is "SensorML urn:sunburst:sensor:SAMI-CO2" with a timestamp of "Wed Nov 04 19:50:22 GMT 2020". The second is "Soil chemical properties, periodic" with a timestamp of "Tue Nov 17 15:54:46 GMT 2020". The third is "Root chemical properties" with a timestamp of "Tue Nov 17 15:54:46 GMT 2020". The fourth is "Sediment chemical properties" with a timestamp of "Tue Nov 17 15:54:46 GMT 2020".

Earthcube Clowder Explore Help Search Sign up Login

## Search

carbon Search Syntax Help Metadata Search

### Results

-  **SensorML urn:sunburst:sensor:SAMI-CO2**  
Wed Nov 04 19:50:22 GMT 2020  
\* Measures the partial pressure of carbon dioxide pCO2 in water from 200-600  $\mu$ atm (ranges above 600 are available by request) \* Uses a highly precise and stable colorimetric reagent method \* Provide researchers with valuable in-situ time series data \* Depolyable to depths up to 600 meters \* Can be deployed in the ocean or in freshwater \* Long-term depolyments - can run for more than a year taking hourly measurements \* Can support up to 3 external instruments such as PAR, dissolved oxygen, chlorophyll fluorometer, or CTD \* Can support inductive modems or external loggers if required. \* Biofouling Package available for deployments in productive environments <https://xdomes.tamucc.edu/srr/sensorML/urn-sunburst-sensor-SAMI-CO2.html>
-  **Soil chemical properties, periodic**  
Tue Nov 17 15:54:46 GMT 2020  
Carbon and nitrogen concentrations from the top 30 cm of the profile. Data are reported by horizon (organic vs. mineral) within a soil core. <https://data.neonscience.org/data-products/DP1.10078.001>
-  **Root chemical properties**  
Tue Nov 17 15:54:46 GMT 2020  
Carbon and nitrogen concentrations in root biomass, either from periodic collections of surface soil (0-30 cm) or from one-time soil Megapit sampling in increments to 2 m depth. <https://data.neonscience.org/data-products/DP1.10102.001>
-  **Sediment chemical properties**  
Tue Nov 17 15:54:46 GMT 2020

The screenshot shows the metadata page for a dataset in Earthcube Clowder. The URL is "earthcube.clowderframework.org/datasets/5fa305fee4b097cab4a0021b". The page has tabs for "Files", "Metadata", "Extractions", "Visualizations", and "Comments (0)". The "Metadata" tab is selected. The metadata is displayed in a tree-like structure with expandable sections. The first section is "Extracted by http://clowder.ncsa.illinois.edu/extractors/deprecatedapi on Nov 4, 2020". The second section is "@type: Dataset". The third section is "isAccessibleForFree: true". The fourth section is "alternateName: urn:sunburst:sensor:SAMI-CO2". The fifth section is "description: \* Measures the partial pressure of carbon dioxide pCO2 in water from 200-600  $\mu$ atm (ranges above 600 are available by request) \* Uses a highly precise and stable colorimetric reagent method \* Provide researchers with valuable in-situ time series data \* Depolyable to depths up to 600 meters \* Can be deployed in the ocean or in freshwater \* Long-term depolyments - can run for more than a year taking hourly measurements \* Can support up to 3 external instruments such as PAR, dissolved oxygen, chlorophyll fluorometer, or CTD \* Can support inductive modems or external loggers if required. \* Biofouling Package available for deployments in productive environments". The sixth section is "includedInDataCatalog:" with sub-sections for "url: https://xdomes.tamucc.edu/srr/" and "@id: https://xdomes.tamucc.edu/srr/". The seventh section is "keywords: oceanography,CO2". The eighth section is "license: https://creativecommons.org/licenses/by/4.0/". The ninth section is "name: SensorML urn:sunburst:sensor:SAMI-CO2". The tenth section is "url: https://xdomes.tamucc.edu/srr/sensorML/urn-sunburst-sensor-SAMI-CO2.html". The eleventh section is "version: 2020-04-17 17:00:00". The twelfth section is "provider:" with sub-sections for "@type: Organization", "legalName: Regional Ocean Acidification: Northwestern Gulf of Mexico", "name: OAR Northwestern Gulf of Mexico", "url: http://hulab.tamucc.edu/OAP/OAP\_index.htm", and "@id: data.gcoos.org". The thirteenth section is "publisher:" with sub-section for "@type: Organization".

Earthcube Clowder Explore Help

Files Metadata Extractions Visualizations Comments (0)

## Metadata

- Extracted by <http://clowder.ncsa.illinois.edu/extractors/deprecatedapi> on Nov 4, 2020
- @type: Dataset
- isAccessibleForFree: true
- alternateName: urn:sunburst:sensor:SAMI-CO2
- description: \* Measures the partial pressure of carbon dioxide pCO2 in water from 200-600  $\mu$ atm (ranges above 600 are available by request) \* Uses a highly precise and stable colorimetric reagent method \* Provide researchers with valuable in-situ time series data \* Depolyable to depths up to 600 meters \* Can be deployed in the ocean or in freshwater \* Long-term depolyments - can run for more than a year taking hourly measurements \* Can support up to 3 external instruments such as PAR, dissolved oxygen, chlorophyll fluorometer, or CTD \* Can support inductive modems or external loggers if required. \* Biofouling Package available for deployments in productive environments
- includedInDataCatalog:
  - url: <https://xdomes.tamucc.edu/srr/>
  - @id: <https://xdomes.tamucc.edu/srr/>
- keywords: oceanography,CO2
- license: <https://creativecommons.org/licenses/by/4.0/>
- name: SensorML urn:sunburst:sensor:SAMI-CO2
- url: <https://xdomes.tamucc.edu/srr/sensorML/urn-sunburst-sensor-SAMI-CO2.html>
- version: 2020-04-17 17:00:00
- provider:
  - @type: Organization
  - legalName: Regional Ocean Acidification: Northwestern Gulf of Mexico
  - name: OAR Northwestern Gulf of Mexico
  - url: [http://hulab.tamucc.edu/OAP/OAP\\_index.htm](http://hulab.tamucc.edu/OAP/OAP_index.htm)
  - @id: data.gcoos.org
- publisher:
  - @type: Organization

## Later EC to future work:

- Linking data with tools ..
- Automatic launching of tools with data
- From search to use in a NoteBook
- Search on map & in NoteBook
- Search enhanced w/NER & more, see:
- <https://mbcode.github.io/ec>
- Getting these benefits in clowder via:
  - triple store sync with clowder
  - embedding science on schema
  - DCAT as a superset/furthering the gateway from schema.org to real science descriptions

### Transect data of coral species and other substrate types collected in the field using line transects in Palau and Yap in 2017 and in the Federated States of Micronesia in 2018

Website Cite Metadata

Type: Data

**Abstract:** As part of the reef-composition survey of Palau (7°30' N, 134°30' E) and Yap (9°32' N, 138°7' E), 10-meter long, 2 to 5-meter depth transects were conducted. Coral species along the transect were recorded along with substrate types and other organisms present. Surveys in Palau were conducted from June 2nd to June 24th, 2017, and from June 25th to July 6th, 2017 in Yap. In Pohnpei (6.2°N, 158.2°E) and Kosrae (5.3°N, 162.9°E) FSM, six 10-meter transects were used to measure the benthic composition for every centimeter, at each site of 48 sites. Corals were recorded to species level, except massive Porites and encrusting Montipora, which were recorded in the field as growth forms. All other organisms along each transect were identified to the highest possible taxonomic resolution.

**Creator:** Robert van Woesik

**Publisher:** Florida Institute of Technology

**Date:** 2020-09-08

Location



Downloads

Download TIFF

Download Shapefile

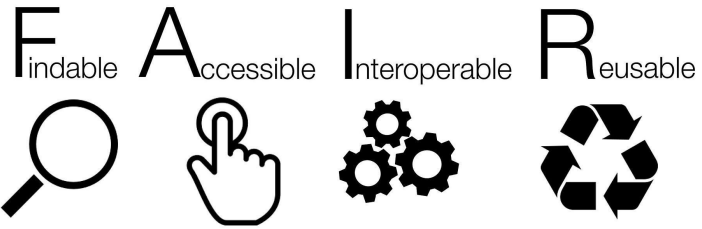
#### Related Data

- ▲ Coral densities and extension rates from scientific literature collected in the field or in laboratories
- ▲ Sea urchin size, density, and species from transects surveyed in Palau and Yap in 2017 and in the Feder...
- ▲ Parrotfish species, density counts, and fish length from field-video surveys in Palau and Yap in 2017...
- ▲ Transect data of coral species and other substrate types collected in the field using line transects in...
- ▲ Bacterial cell counts and Dissolved Organic Carbon (DOC) measurements from R/V Atlantis AT32, AT34...

#### Compatible Tool

- ▲ NetCDF classic format (netCDF)
- ▲ TopBraid Composer Free Edition
- ▲ LinkedEarth
- ▲ McIDAS grid file format (McIDASGrid)
- ▲ Application for Extracting and Exploring Analysis Ready Samples (AppEARS)

*Faster time to science*  
via metadata use  
to get more



resources

Can take questions later: @Mike Bobak



# GEOSCIENCE CYBERINFRASTRUCTURE FOR OPEN DISCOVERY IN THE EARTH SCIENCES (GEOCODES.EARTHCUBE.ORG)

Bobak, Mike; Coakley, Kevin; Fils, Doug; Gatzke, Lisa; Kirkpatrick, Christine; McHenry, Kenton; Richard, Steve; Valentine, David; Zaslavsky, Ilya; Zhang, Bing



GeoCODES.earthcube.org

The NSF EarthCube program funded several programs, and the program office<sup>[1]</sup> worked with and for the community through the GeoCODES effort. First, to adopt FAIR (Findable Accessible Interoperable Reusable) metadata principles<sup>[2]</sup> starting with schema.org<sup>[3]</sup> to annotate datasets across several repositories. Then, to construct tools that would crawl, index and host a search of the metadata for these resources<sup>[4]</sup>. Which includes discovery of other data and tools. Much of the metadata for these tool matches come from the office's Resource-Registry<sup>[5]</sup>, which jump-started the cataloging of this metadata, now allowing new entry at [addto.earthcube.org](http://addto.earthcube.org).

The Findable part of FAIR starts in the search, and we continue to work on easier Access Interoperation and Reuse of geoscience resources, e.g. data and tools, through links out to repositories and tool/services, and by combining them in computational notebooks such as [myBinder](http://myBinder) and [Google-Colab](http://Google-Colab).

We plan to continue the adoption of even more machine-actionable FAIR metadata standards. Through entity-finding annotation and feedback to the repositories. With this we will more easily be able to match/find even more related resources and help in automating bringing them together in a computational notebook/workspace. Enabling much more of the FAIR acronym than just Fair/search.

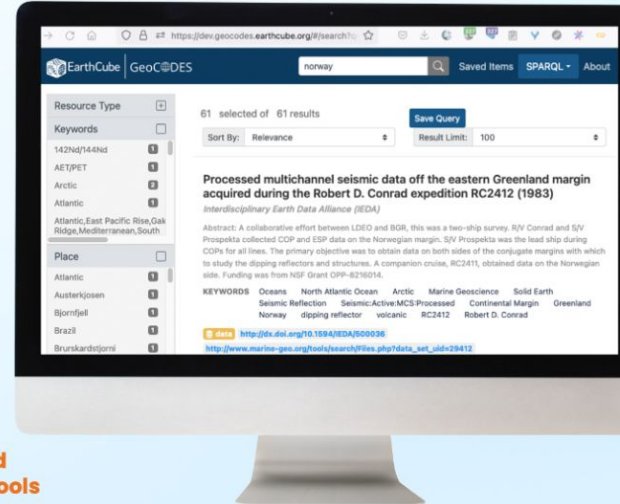
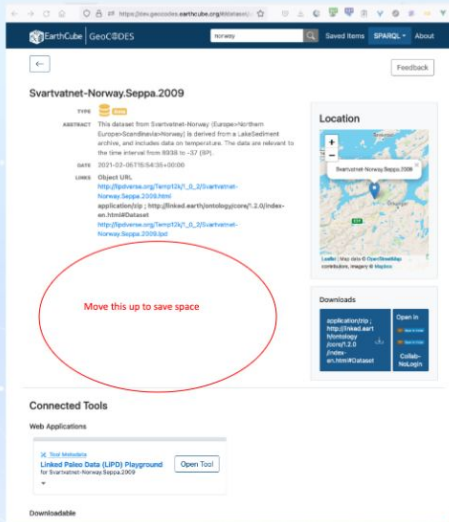
This started as a prototype, but has built up a more reliable infrastructure, and is not only soliciting feedback on usability, and asking for new resources, we also hope to get other contributions, and make this more easily reused and extended by the wider community.



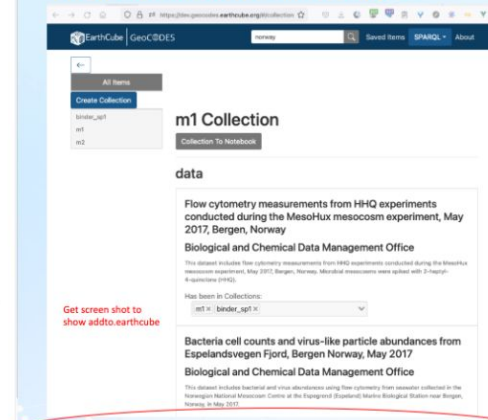
Search geoscience data repositories and related resources.

Faceted by resource-type, keyword, time, place, publisher..

Dataset page: with related resources: datasets and tools



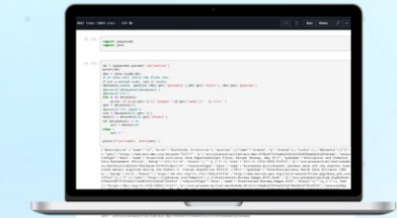
Can pick and open collections of resources in the same Notebook:



not shown: [addto.earthcube.org](http://addto.earthcube.org) for new resource-registry entry

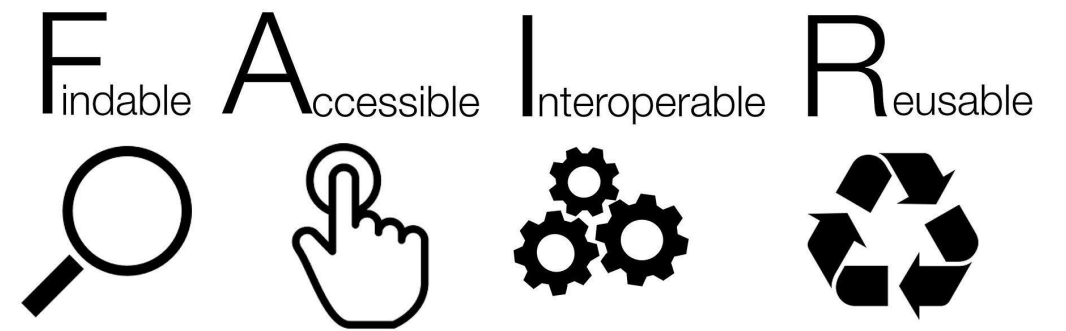


Open Resources within a Notebook

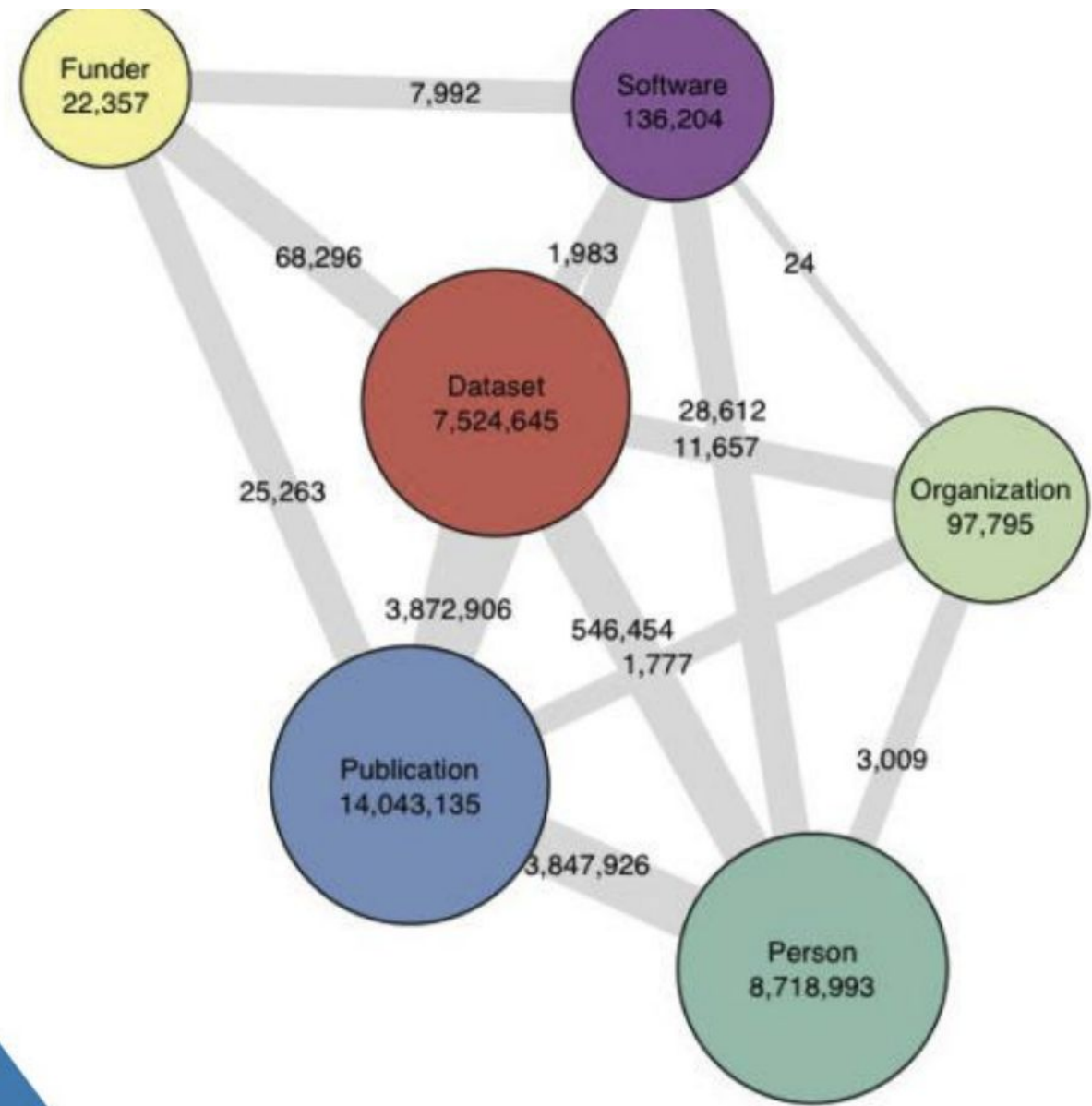
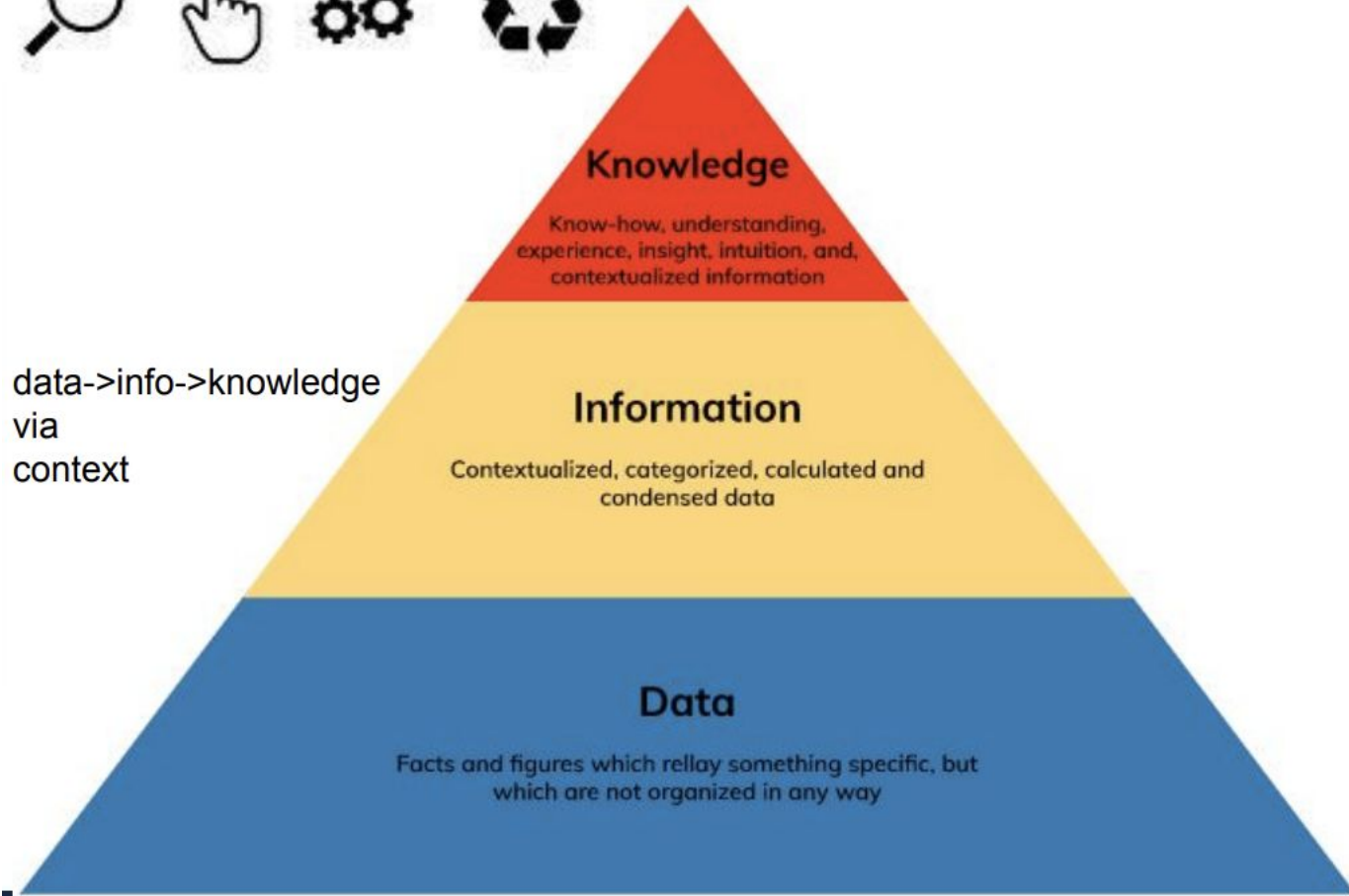
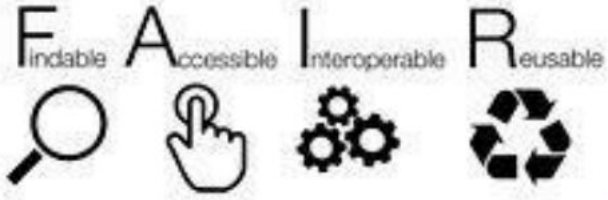


References: <https://www.earthcube.org/eco> | <https://www.go-fair.org/fair-principles/> | <https://github.com/ESIPFed/science-on-schema.org/blob/master/guides/Dataset.md> | <https://github.com/earthcube> | <https://github.com/earthcubearchitecture-ecresourcereg>

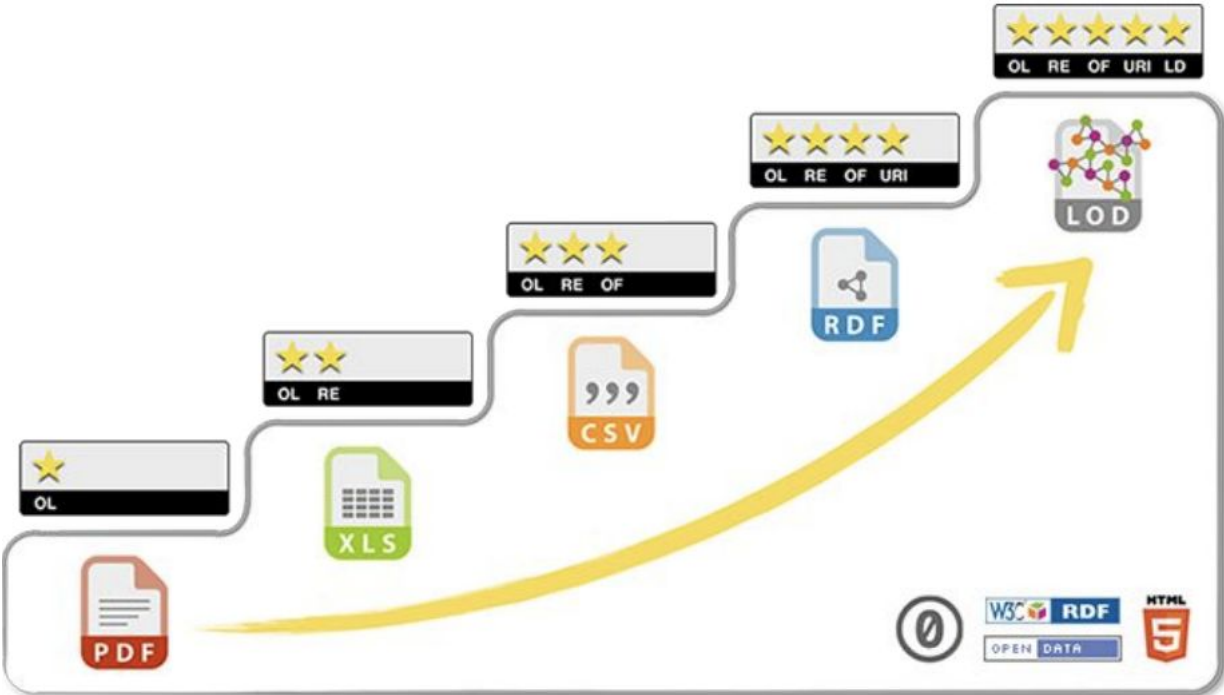
extra slides



Throughput is an EC project that might help us bring in some more of these linkages  
 Linked-Data is what makes these resources

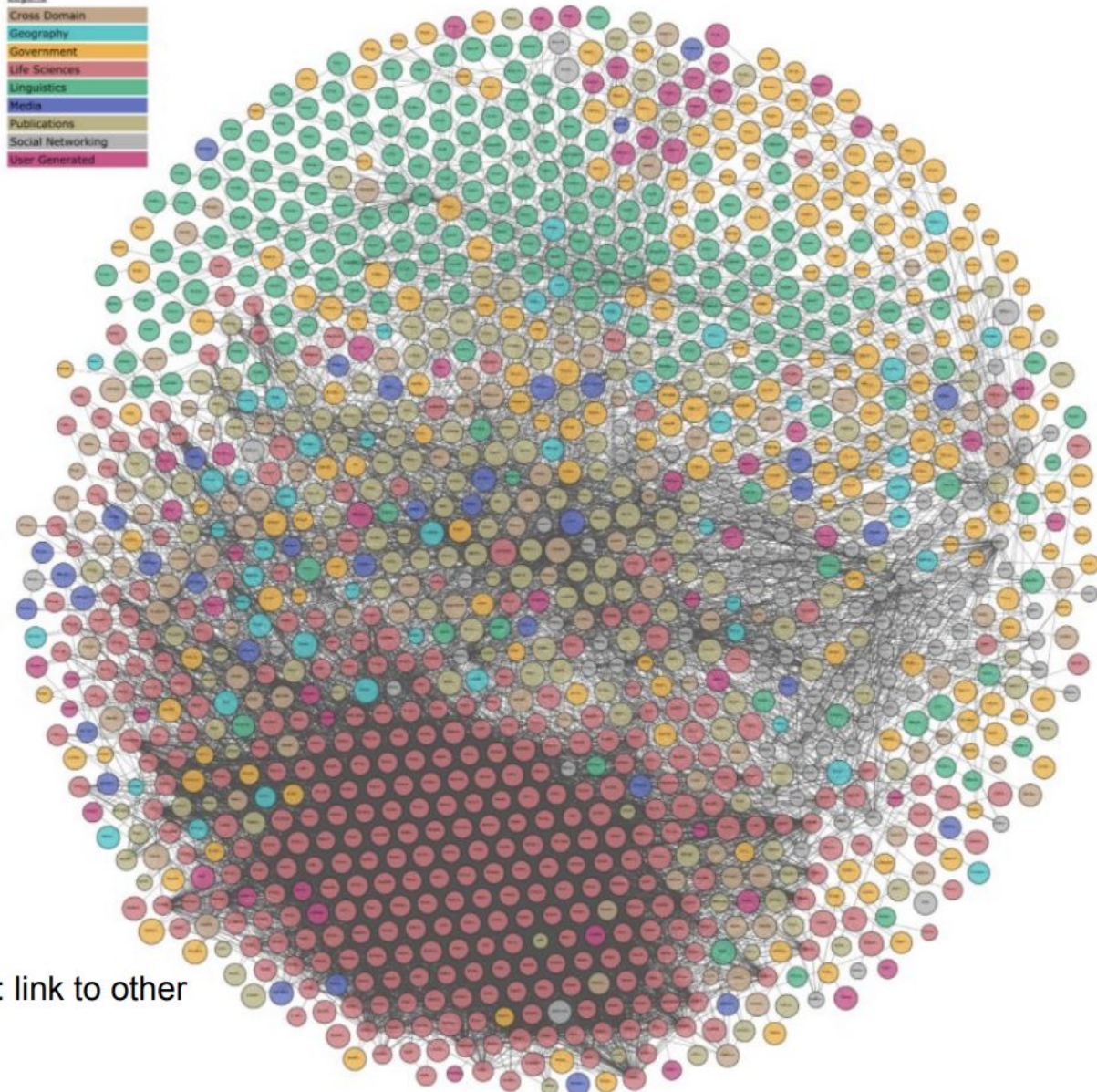


[5stardata.info/en](http://5stardata.info/en) last star is linking to the  
LinkedOpenData cloud [lod-cloud.net](http://lod-cloud.net)



Available as: 1: open online, 2: structured, 3: non-proprietary, 4: ref via URIs, 5: link to other formats

- Legend
- Cross Domain
- Geography
- Government
- Life Sciences
- Linguistics
- Media
- Publications
- Social Networking
- User Generated



Biomed free-text conceptual annotation, applications:

UCSF: to make a conceptual query; AHN adds tagged patients; UIUC relationship tagging

UCSF: Annotated in/ex-clusion criteria, but the logic of the query was not fully automated

While I used MMTx and Metamap while at UCSF, including asking for the source to be posted at NLM;  
So I could more easily alter the algorithm. I did not get to try to make use of the early versions of SemRep;  
As I was already using NLP libs to get the Noun\_Phrases and some of the other modifiers/connection/etc.

Aloha Health: has looser concept connections, but includes patients, and contextual weights

Now also seeing the UMLS (SNOMED/Radlex/..) annotations of the criterion and EMRs, using a private algorithm;  
I would like to get back to extending open NLM and other packages, on a data warehouse that I could get to know better.

UIUC: easier to use SemRep allows for easier text to Knowledge-Graph, and structured queries

Given the pilot grant I worked on with the SemRep author and his grad student, there is some of that  
and the related NER extensions that I looked into, that I would also consider trying to make use of now

The pilot grant did go forward using the clowder-framework; which I extended to make it's datasets more  
discoverable, and could benefit from another FAIR dataset discovery & use application of mine as well

UIC: Hoping to learn more about the potential range of the role today



# AlohaHealth.net

skip this slide

Get a (range of) possible match/es for each criterion at a site

and possible sites to contact for a particular trial

This screenshot shows the 'Criteria' section of the AlohaHealth.net interface. It features a dark background with a hexagonal grid pattern on the left. The 'Criteria' tab is selected, with sub-tabs for 'Inclusion' and 'Exclusion'. Four criteria are listed, each with a slider and a count of matches:

- Age > 21 years old: 18 matches
- Chronic post amputation pain > 6 months: 3 to 9 matches
- No changes to medications or prosthesis for 3 month primary study period: 1 to 6 matches
- Pain episodes lasting > 60 minutes: 60 matches

At the bottom, a fifth criterion 'Stable drug regiment > 6 weeks' is partially visible.

This screenshot shows the right side of the AlohaHealth.net interface. It features a dark background with a hexagonal grid pattern on the left. The 'View sites by category' section is active, showing a legend for site categories:

- University (blue square)
- Hospital (purple square)
- Independent (orange square)
- Site Management Org. (SMD) (red square)

Below the legend, there are two sliders:

- 'Patient geofence radius (miles)' with a slider set to approximately 10 miles.
- 'Number of matching patients' with a slider set to approximately 10.

At the bottom, an 'Active studies' slider is partially visible, and the number '16' is displayed at the bottom center.