NCSA faculty fellowship w/iSchool on turning free-text into Knowledge-Graph triples

Mike Bobak

ILLINOIS
NCSA | National Center for Supercomputing Applications

NCSA faculty fellowship w/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:

- 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 braat 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.

- Update 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

Motivation: of machine interpretability of knowledge from free-text

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

blog.google/products/search/introducing-knowledge-graph-things-not

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.

There are several application areas for machine interpretable knowledge

e.g.

Metadata for Machines (M4M)

Domain Experts + **Metadata Experts**
- Registries
- Tools

= **FAIR-ready metadata templates**
- target FAIR Principles
- compliant to community standards
- machine-actionable
- registered/reusable

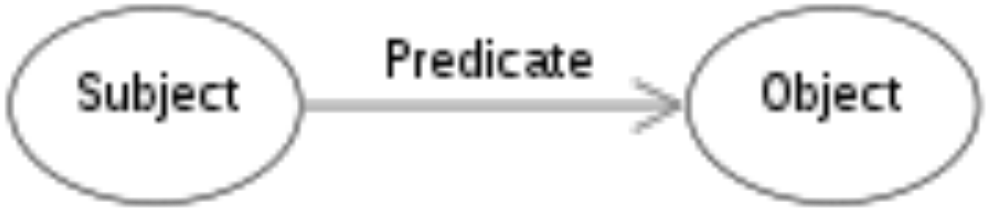
Short workshops that create high-priority machine-actionable metadata for the specific needs of particular communities of practice.

EarthCube

Named-Entity-Recognition & Linking

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

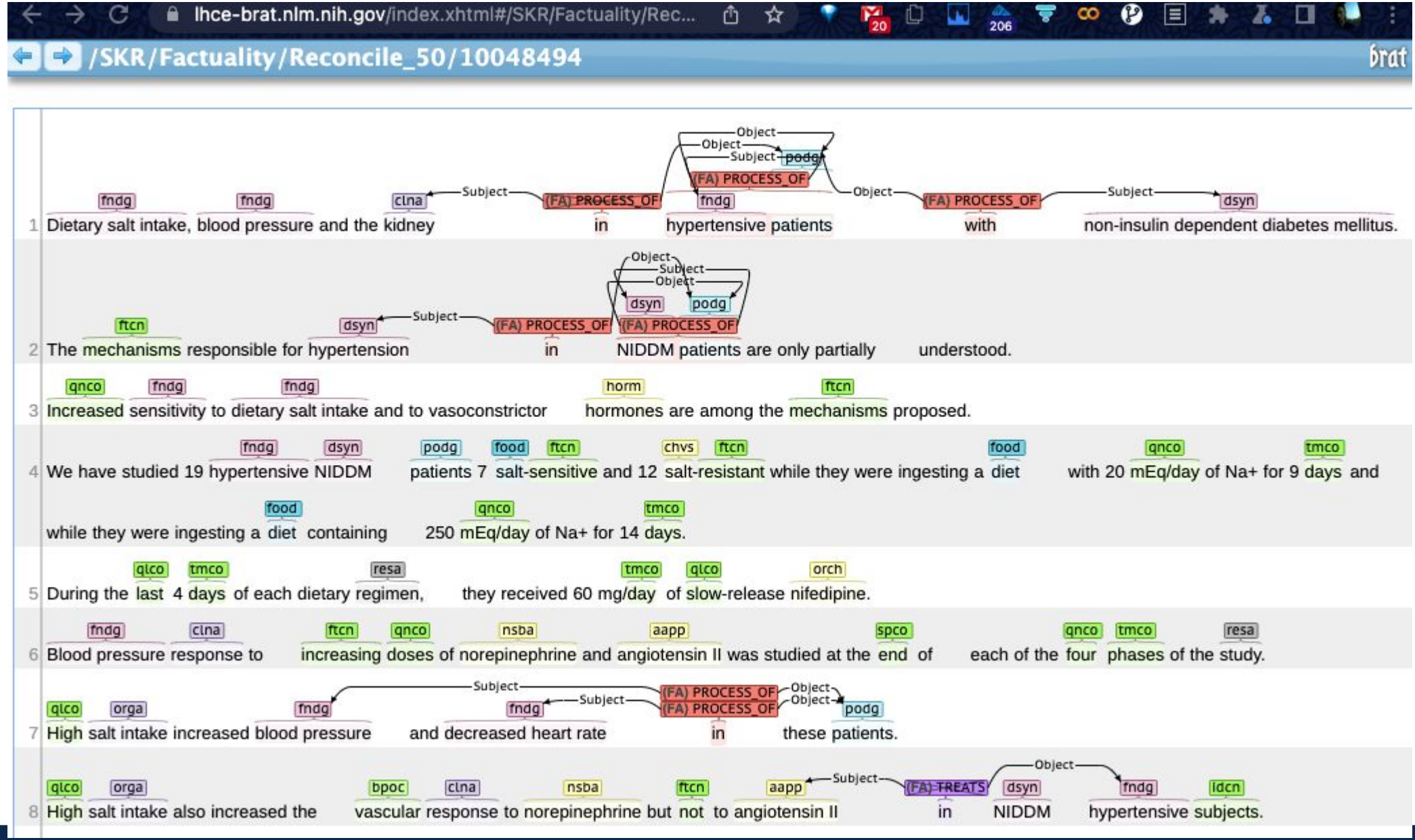
How it works:

- 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



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 braat 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

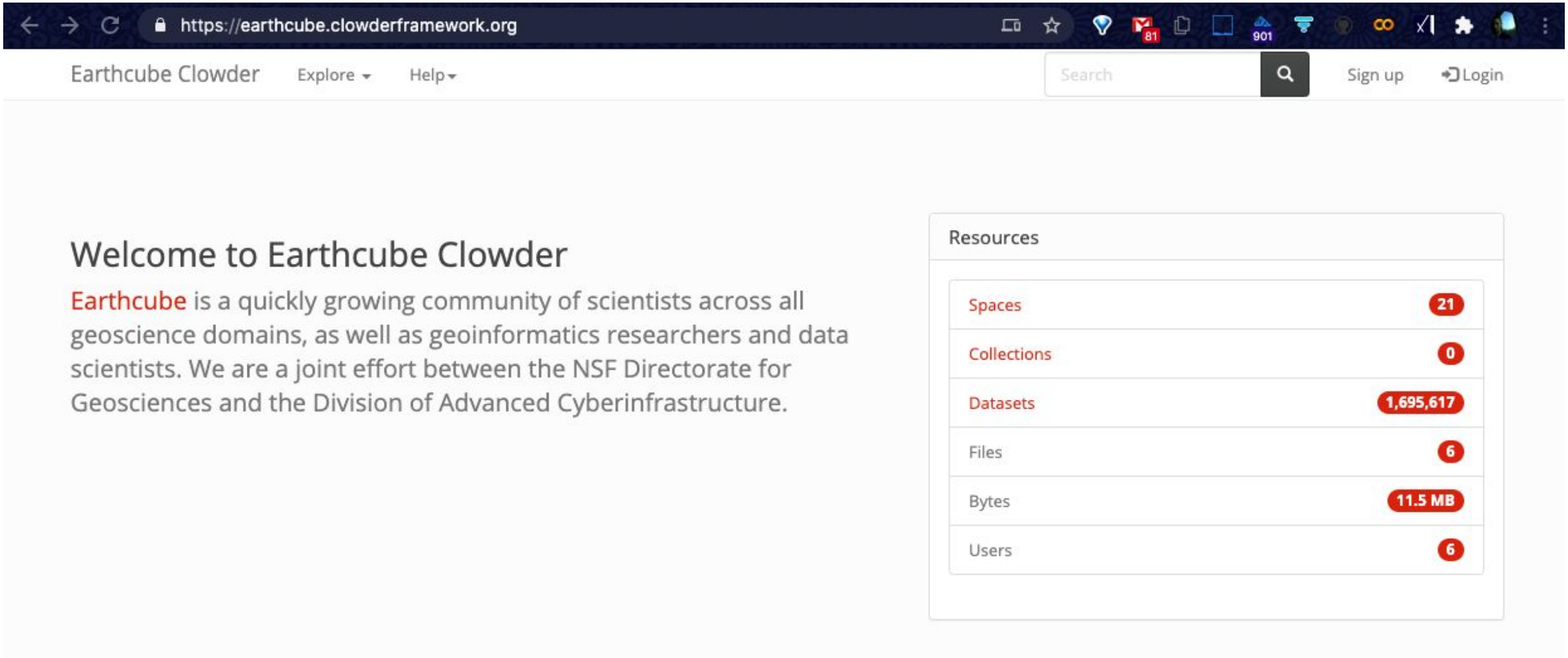
https://isda.ncsa.illinois.edu/~mbobak/

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

Clowder is mentioned in the NIH grant proposal &I will annotate this EC free-text too



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 contintental 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.

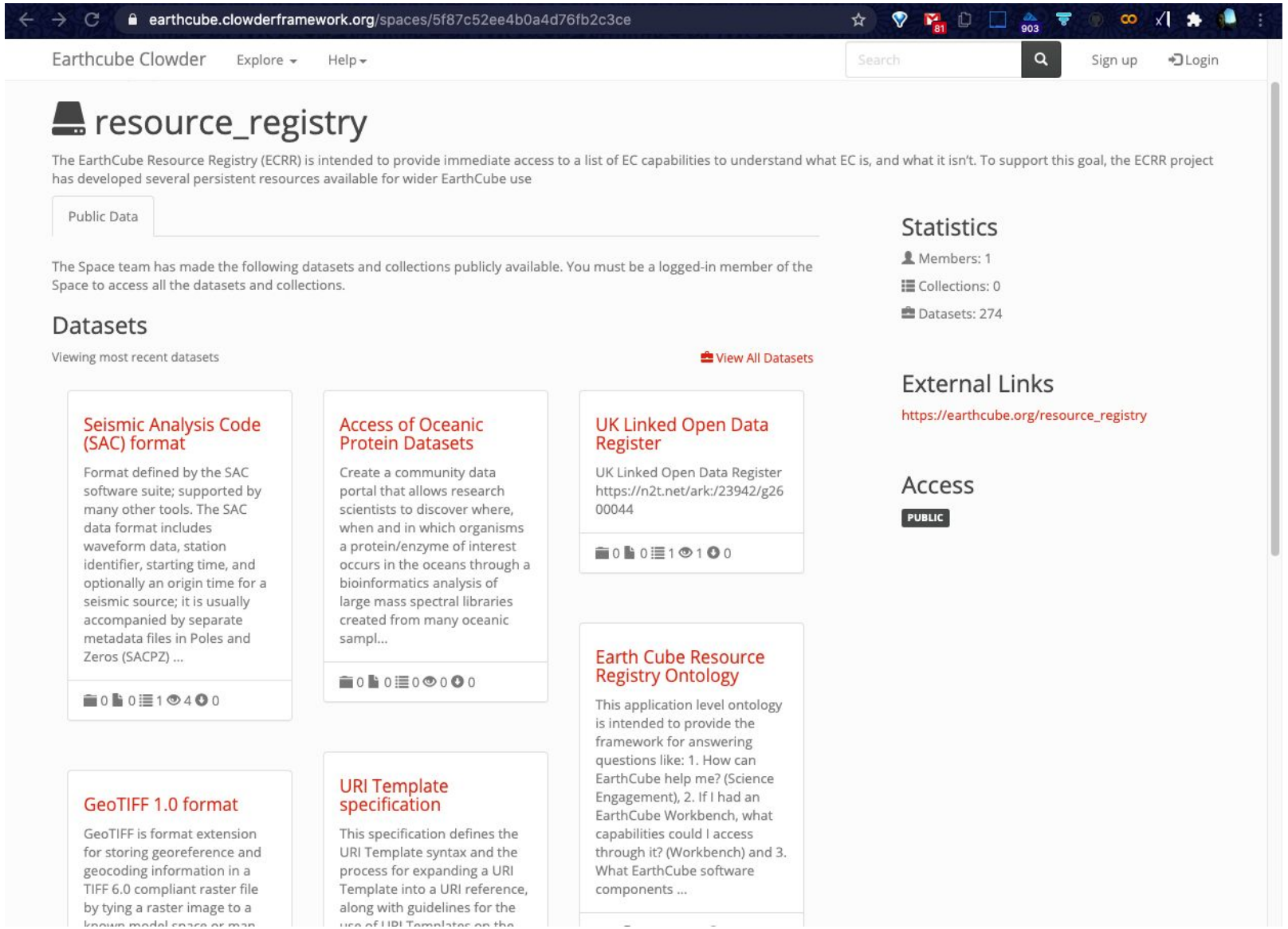
4185 0 1

Clowder organization

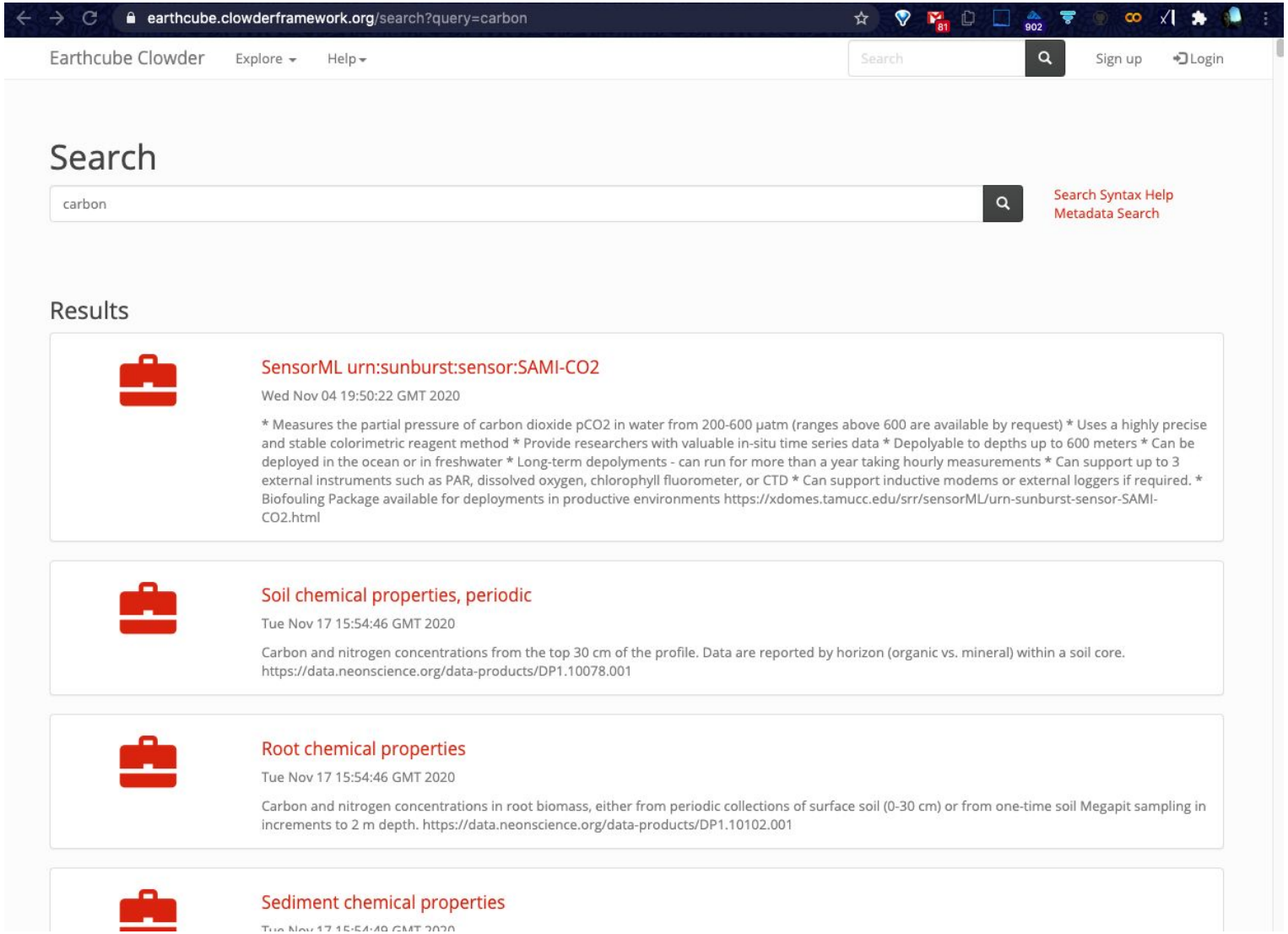
- One *space* per data-facility
- *Datasets* hold metadata
- Also a Resources space:

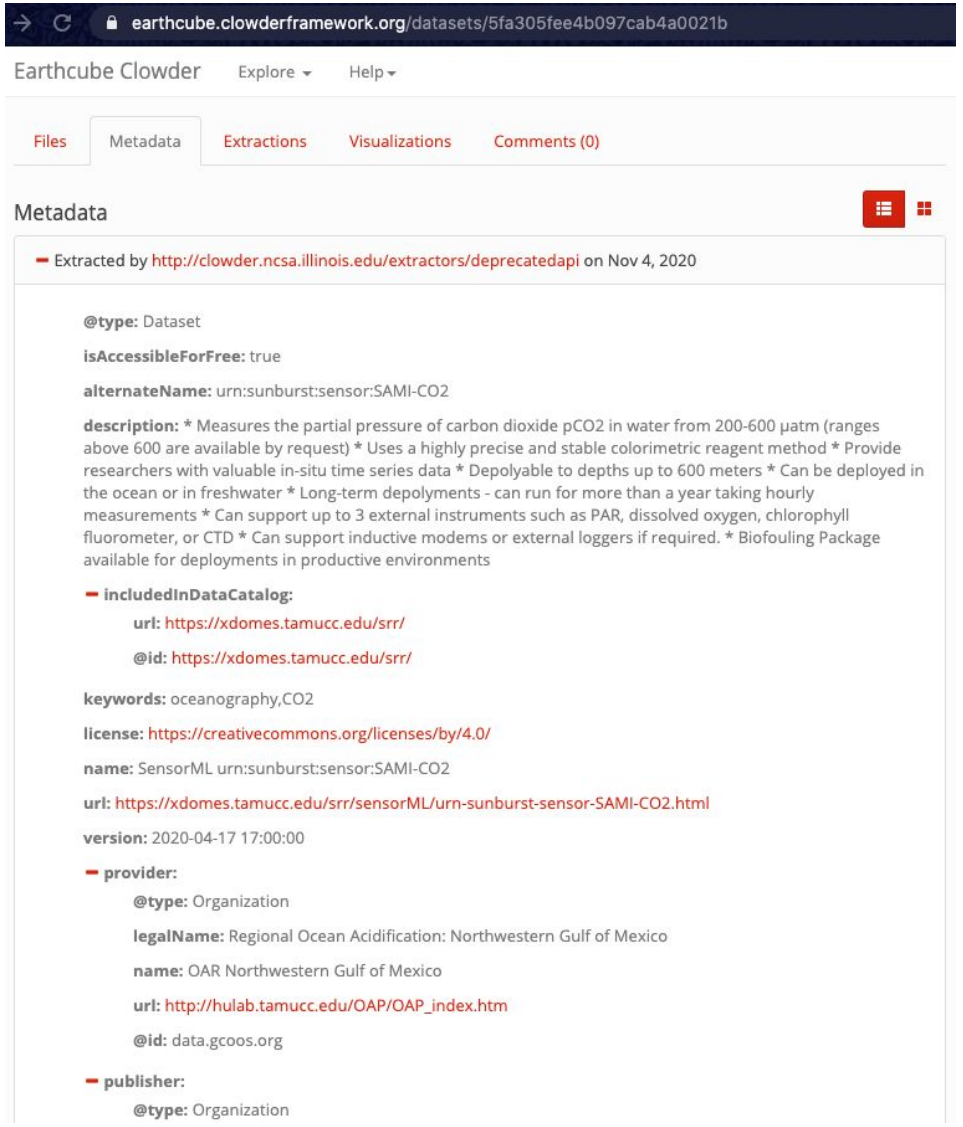
Allows for

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



Clowder search results & a result's metadata(tab) tree listing



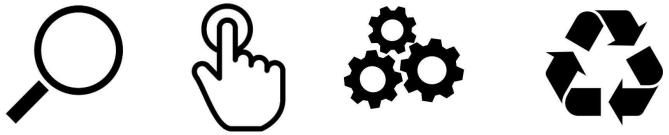


Future work:

- 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

Faster time to science
via metadata use
to get more

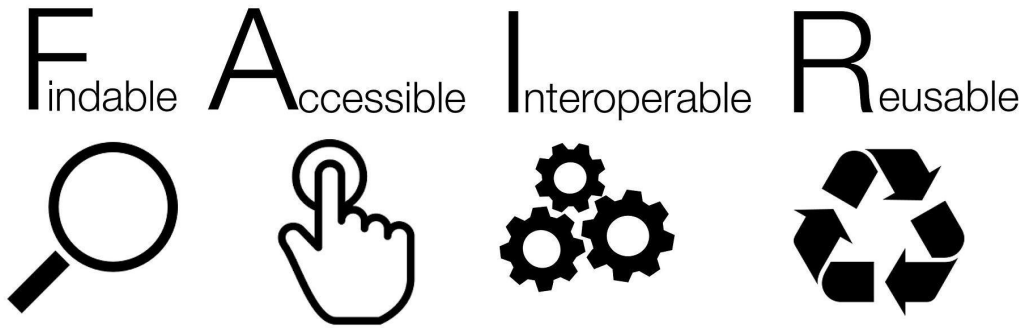
Findable Accessible Interoperable Reusable



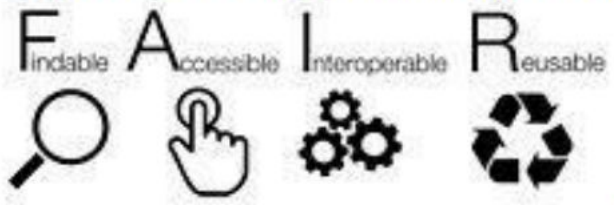
resources

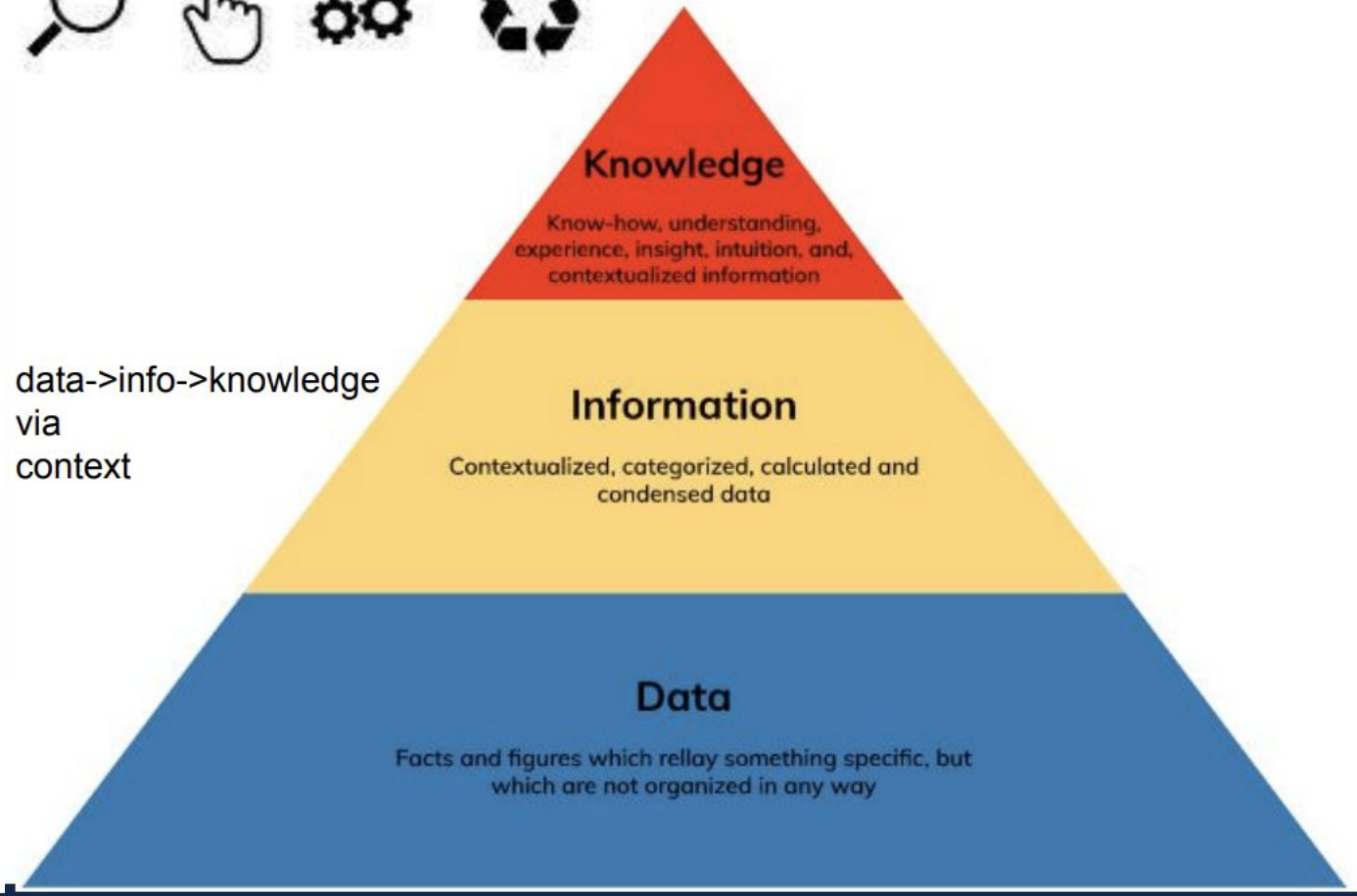
Can take questions later: @Mike Bobak

extra slides

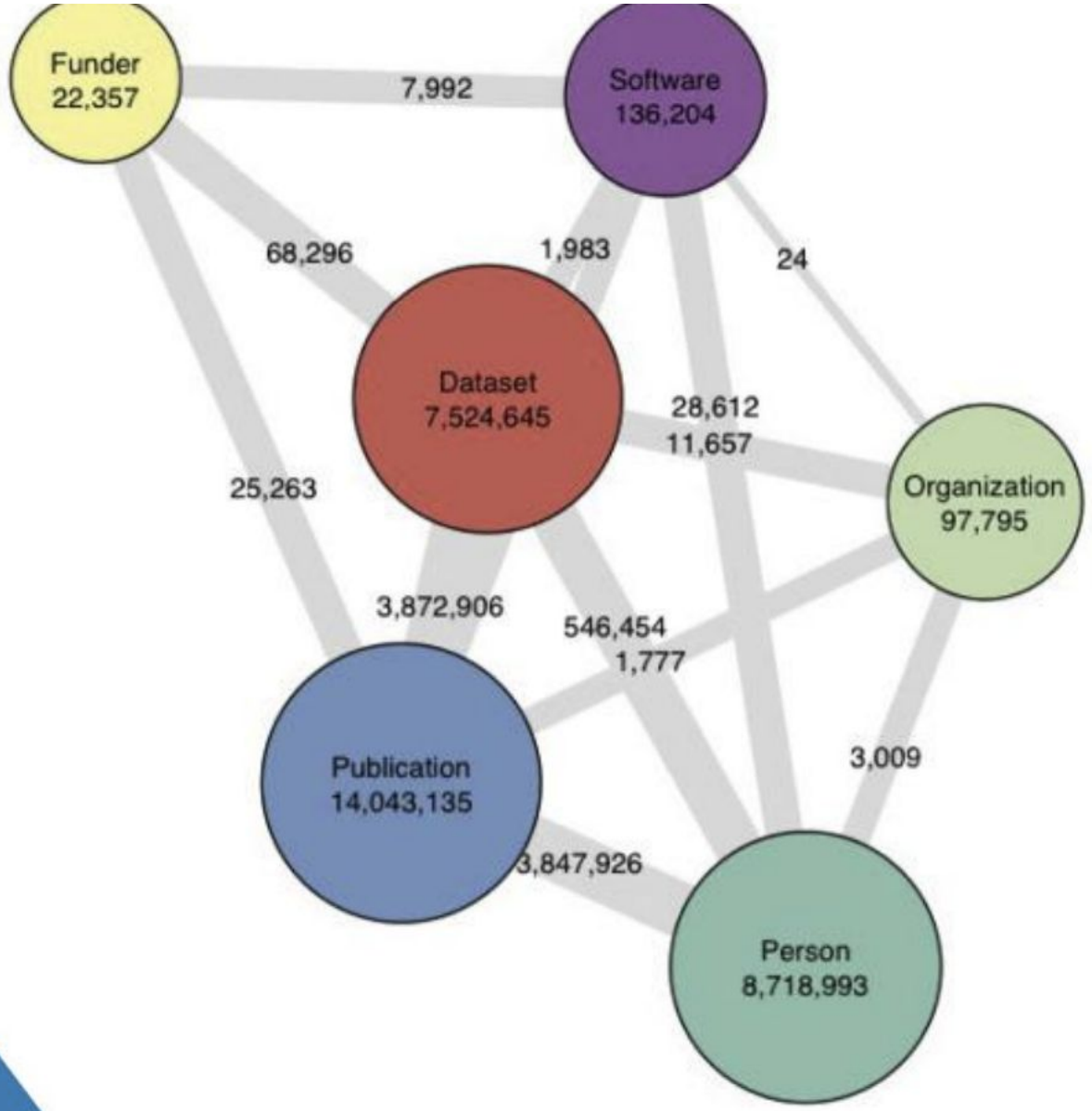


Throughput is an EC project that might help us bring in some more of these linkages
Linked-Data is what makes these resources

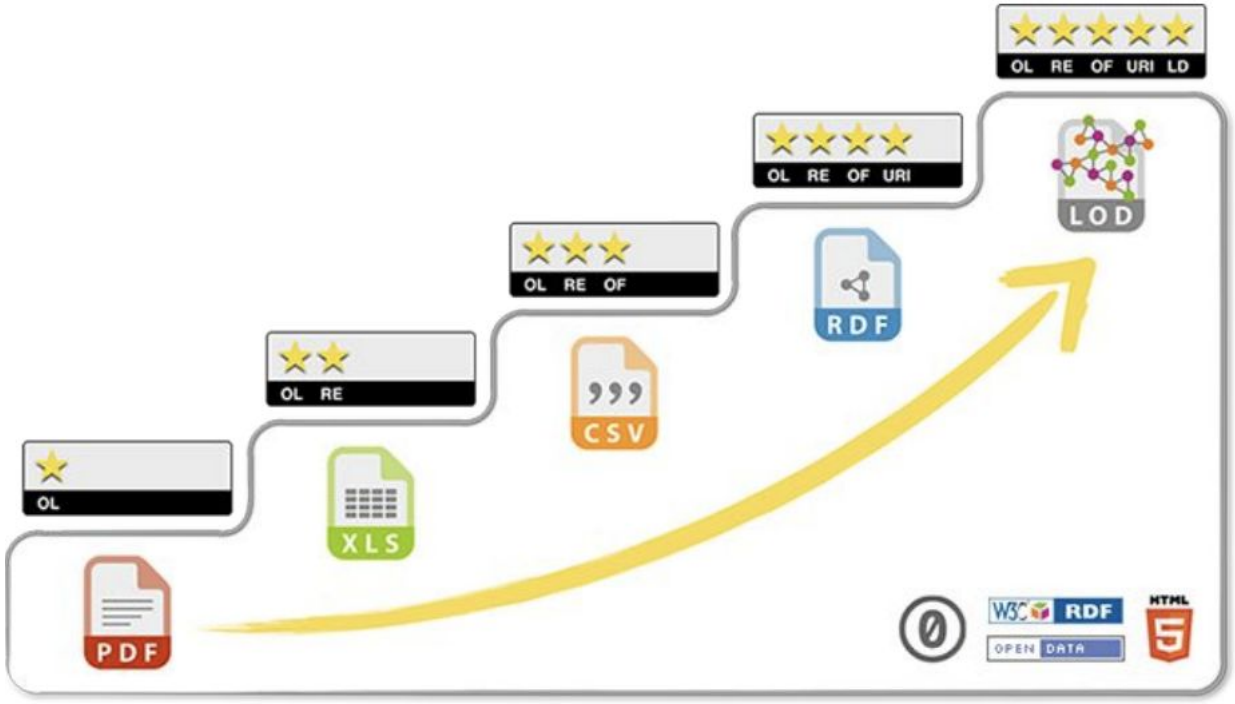


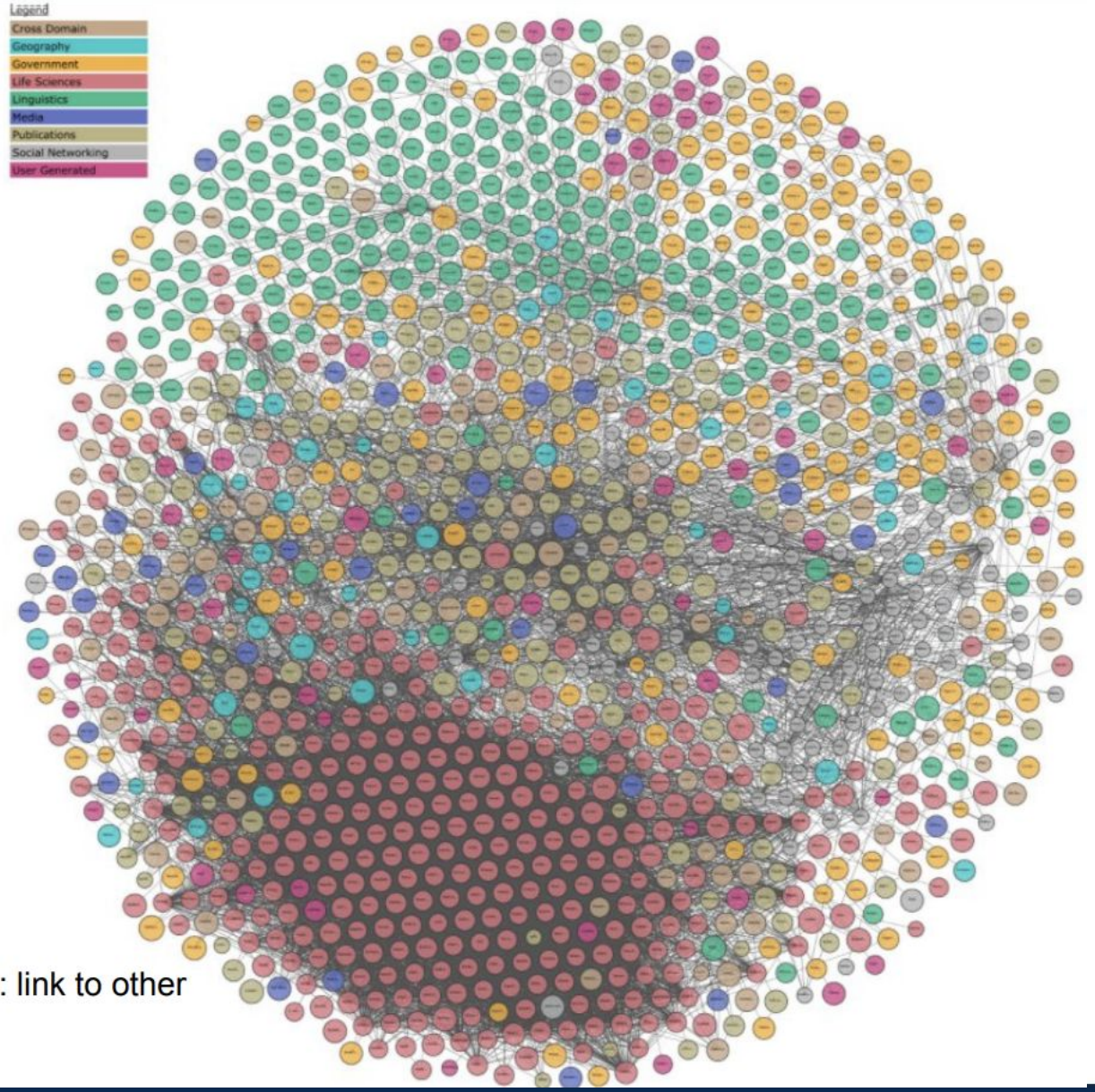


data->info->knowledge
via
context



5stardata.info/en last star is linking to the LinkedOpenData cloud lod-cloud.net





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