

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

Mike Bobak



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 the goal of machine-interpretability [KGs need connected things]

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”



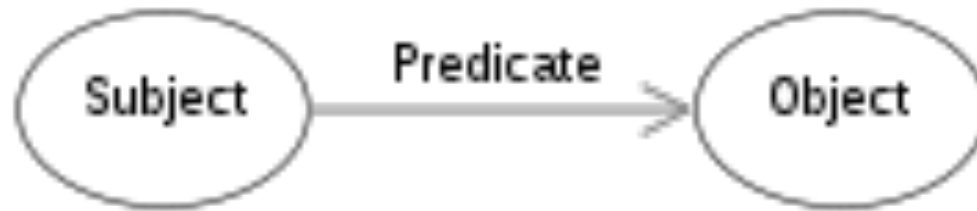
wikipedia.org/wiki/Paris



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

lhce-brat.nlm.nih.gov/index.xhtml#/SKR/Factuality/Rec...
/SKR/Factuality/Reconcile_50/10048494

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⁺ for 9 days and while they were ingesting a diet containing 250 mEq/day of Na⁺ 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 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

←

→

↺

https://earthcube.clowderframework.org

📱

☆

💡

📧81

📄

📺

📶901

🔊

🔍

🔗

✂

⚙

🌐

⋮

Earthcube Clowder

Explore ▾

Help ▾

Search

🔍

Sign up

🔑 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



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 web interface. The browser address bar displays the URL: `earthcube.clowderframework.org/spaces/5f87c52ee4b0a4d76fb2c3ce`. The page title is "resource_registry". A description states: "The EarthCube Resource Registry (ECRR) is intended to provide immediate access to a list of EC capabilities to understand what EC is, and what it isn't. To support this goal, the ECRR project has developed several persistent resources available for wider EarthCube use".

On the right side, there are statistics: "Members: 1", "Collections: 0", and "Datasets: 274". Below these are "External Links" with the URL `https://earthcube.org/resource_registry` and an "Access" section with a "PUBLIC" button.

The main content area is titled "Datasets" and shows "Viewing most recent datasets". A link "View All Datasets" is available. Five dataset cards are displayed:

- 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...

Clowder search results

& a result's metadata(tab) tree listing

← → ↺

earthcube.clowderframework.org/search?query=carbon

☆

61

902

Earthcube Clowder

Explore ▾

Help ▾

Search

Q

Sign up

Login


Search

carbon

Q

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 µatm (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

→ ↺

earthcube.clowderframework.org/datasets/5fa305fee4b097cab4a0021b

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 µatm (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

@id: data.gcoos.org

publisher:

@type: Organization

I

ILLINOIS NCSA

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

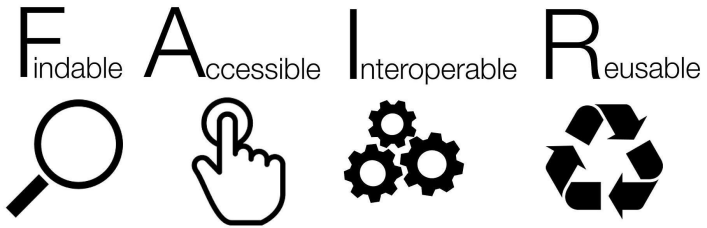
Related Data

- ▲ 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)

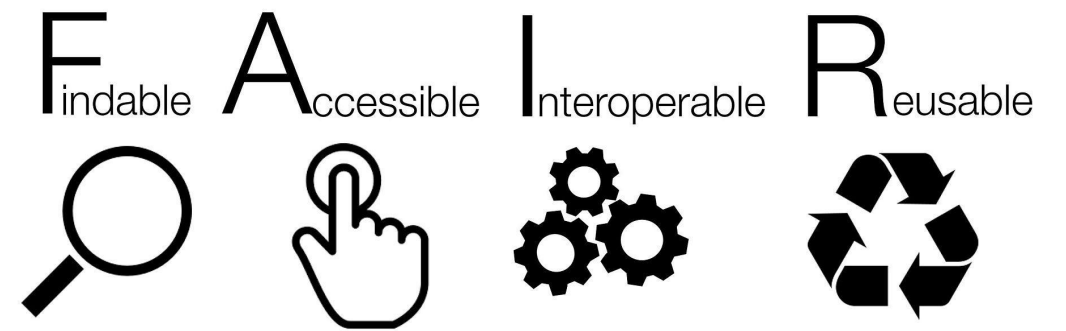
Faster time to science
via metadata use
to get more



resources

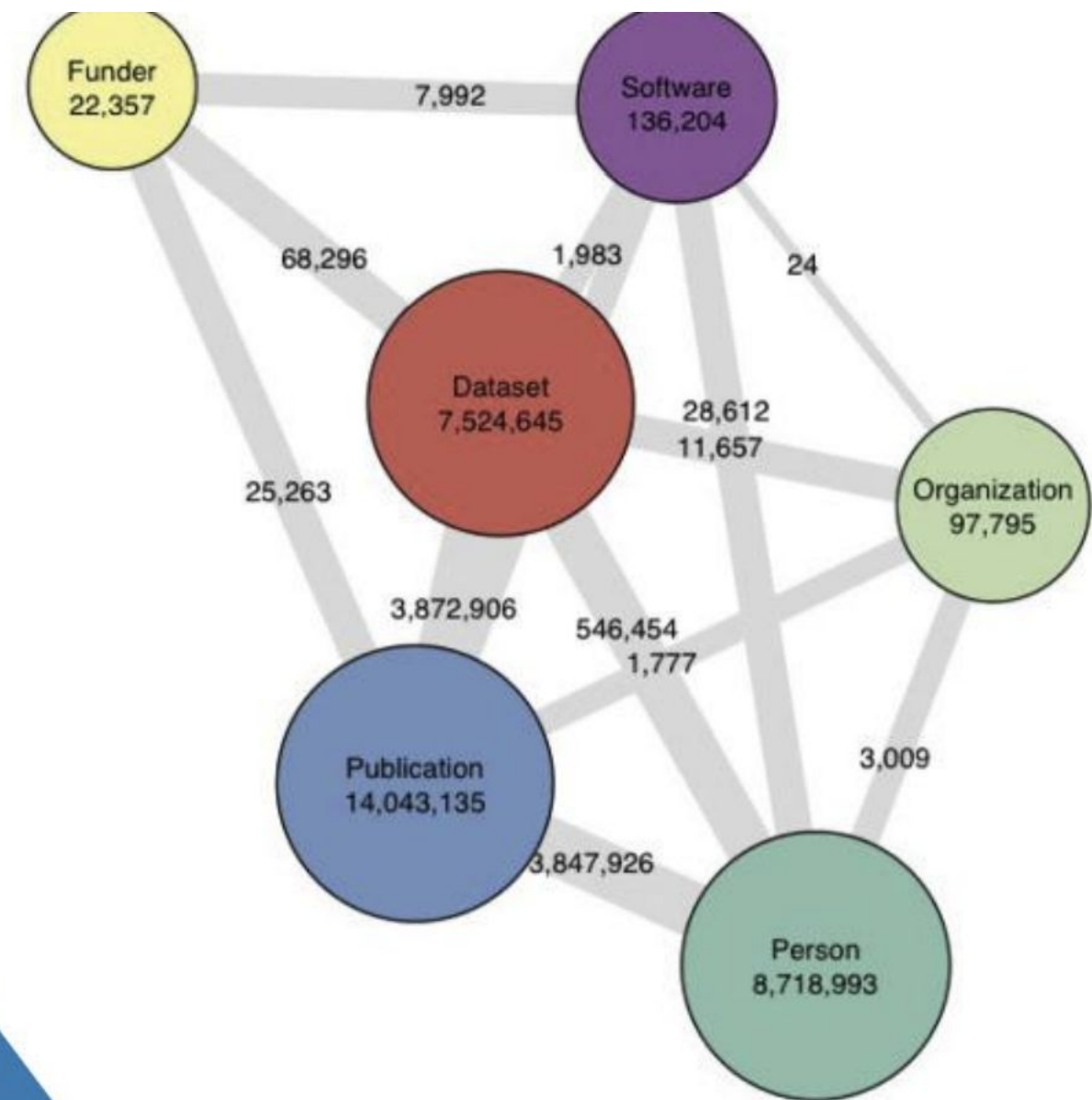
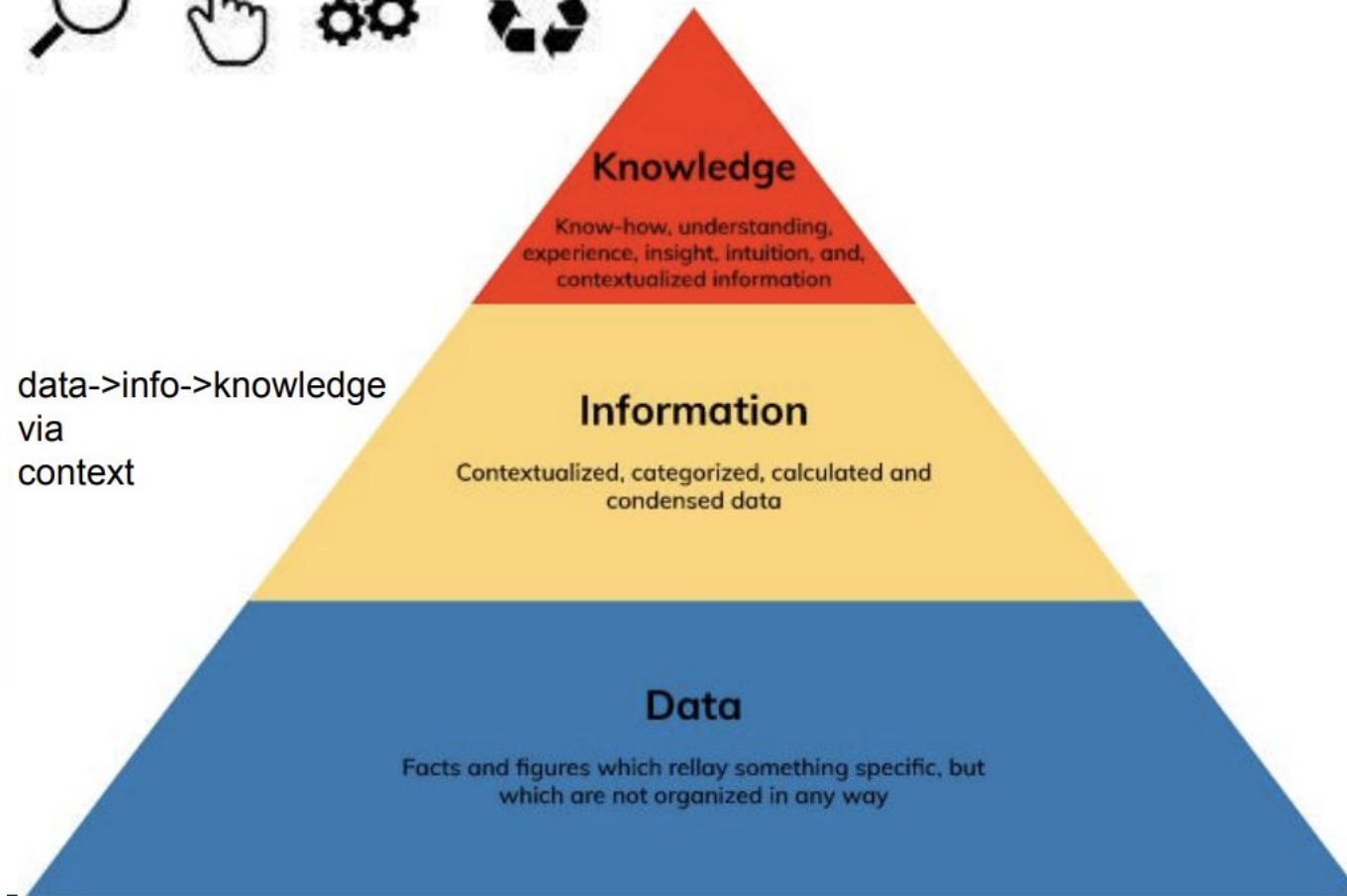
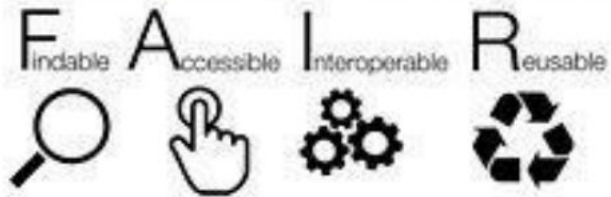
Can take questions later: @Mike Bobak

extra slides

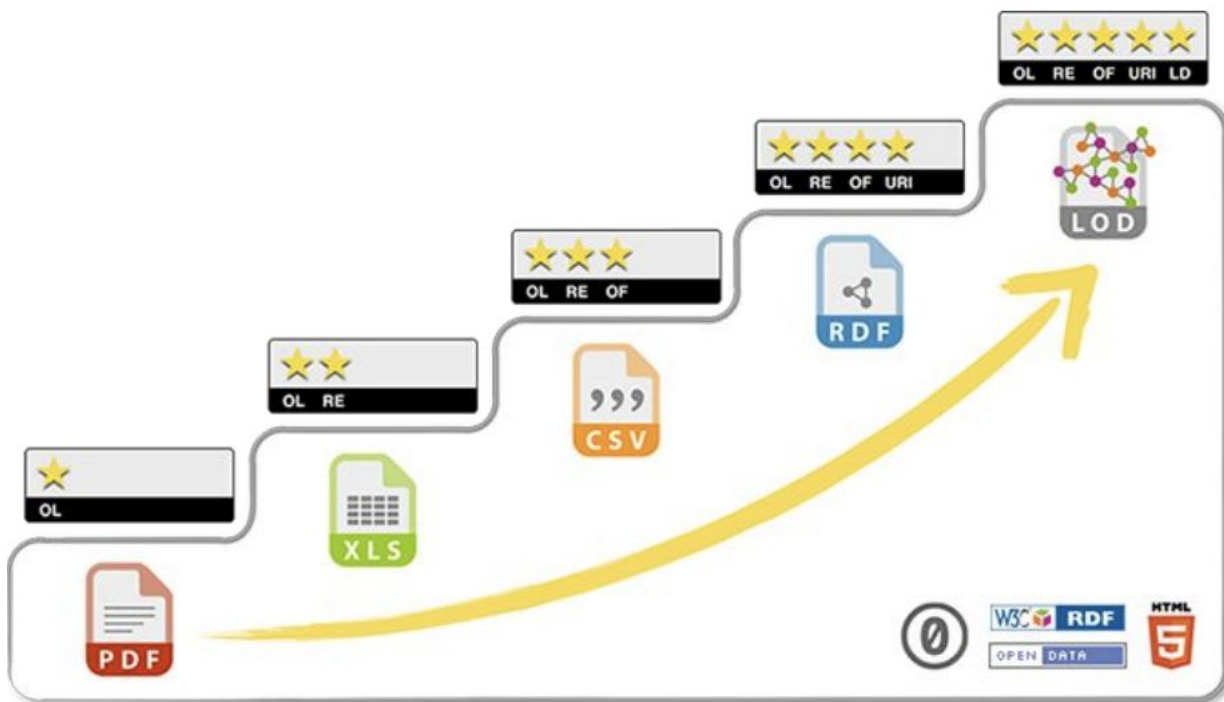


Throughput is an EC project that might help us bring in some more of these linkages

Linked-Data is what makes these resources



5stardata.info/en 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

