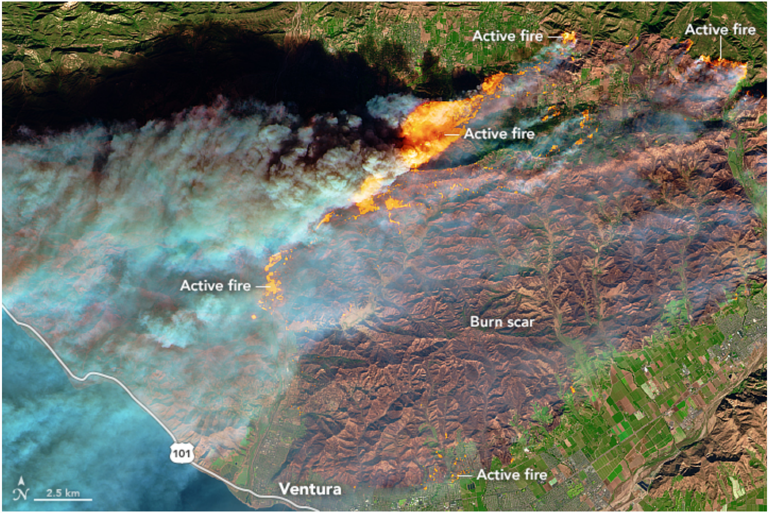
KnowWhereGraph: Enriching and Linking Cross-Domain Knowledge Graphs using Spatially-Explicit AI Technologies to Address Pressing Challenges at the Human-Environment Nexus

Krzysztof Janowicz (for A-6677)

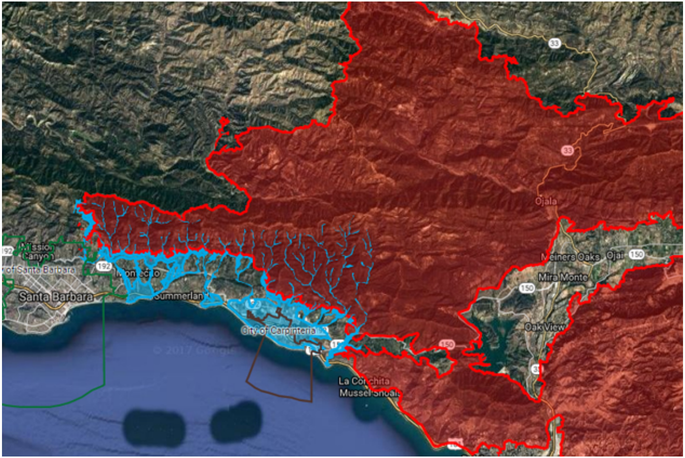
STKO Lab, University of California, Santa Barbara, USA

April 2020

Thomas Fire and Mudslide

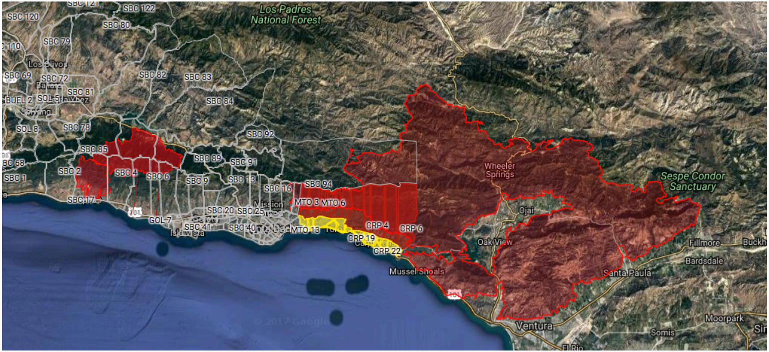


Thomas Fire and Mudslide



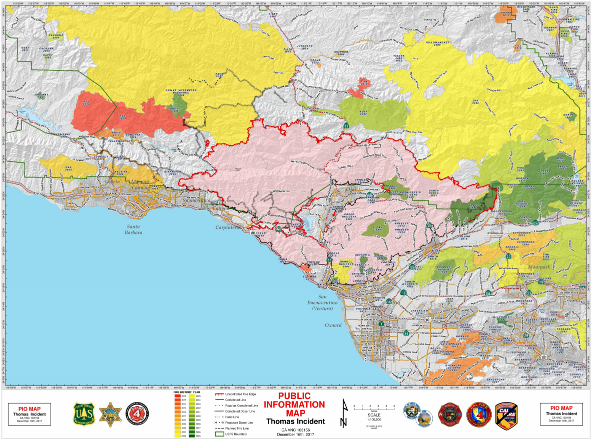
- Screenshot of the Santa Barbara County post-fire hazards map shows the flooding risk areas in blue (shown on 1/5/18; days before the storm/mudslide).

Thomas Fire and Mudslide



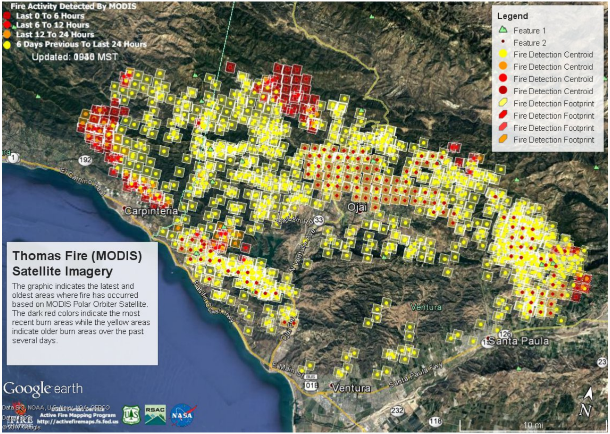
- ‘But your own home, right next to a creek, was in a voluntary evacuation zone. Didn’t that seem odd to you? No. To even be in a voluntary zone seemed significant to me. The problem with evacuating for a flood is that our evacuation zones aren’t broken up by watersheds. They’re broken up by roads. If you look at our zone map, it’s broken up in boxes, and you evacuate people from that box. Part of that is so you can protect it from looting. Once you clear everybody out, you kind of own it, and now you have to keep it safe.’

Thomas Fire and Mudslide as a Vector (Map)



- We know how to **partially represent** such ‘object’ data on a KG, e.g., using GeoSPARQL, SWEET, SSN.

Thomas Fire as a Raster Product

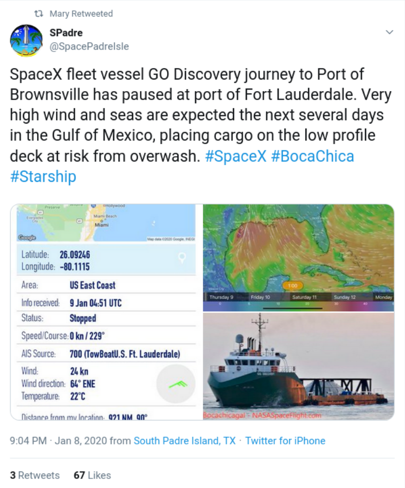


Thomas Fire (MODIS) Satellite Imagery

The graphic indicates the latest and oldest areas where fire has occurred based on MODIS Polar Orbiter Satellite. The dark red colors indicate the most recent burn areas while the yellow areas indicate older burn areas over the past several days.

- **Not so clear how to put such ‘field’ data on the graph.** E.g., annotation ontologies, SSN, object recognition/extraction, semantic image parsing, and so on.

Why SpaceX is so Radically Different?



Mary Retweeted

SPadre
@SpacePadreIsle

SpaceX fleet vessel GO Discovery journey to Port of Brownsville has paused at port of Fort Lauderdale. Very high wind and seas are expected the next several days in the Gulf of Mexico, placing cargo on the low profile deck at risk from overwash. #SpaceX #BocaChica #Starship

Latitude:	26.09246
Longitude:	-80.1115
Area:	US East Coast
Info received:	9 Jan 04:51 UTC
Status:	Stopped
Speed/Course:	0 kn / 229°
AIS Source:	700 (TowBoatU.S. Ft. Lauderdale)
Wind:	24 kn
Wind direction:	64° ENE
Temperature:	22°C
Distance from my location:	921 NM, 90°

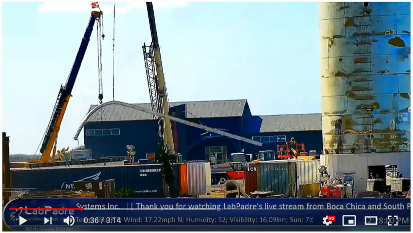
Bocachicagal - NASASpaceFlight.com

9:04 PM · Jan 8, 2020 from South Padre Island, TX · Twitter for iPhone

3 Retweets 67 Likes

Imagine yourself as an analyst at

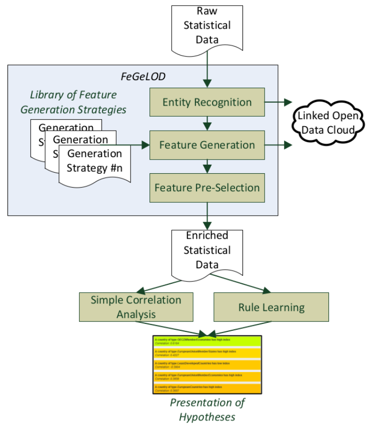
- their **insurance** company
- their **competitors**
- …

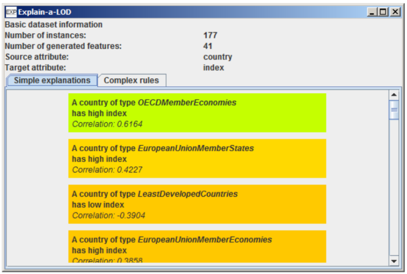


LabPadre
With Maria Pointer
Systems Inc. || Thank you for watching LabPadre's live stream from Boca Chica and South P
Wind: 17.22mph N; Humidity: 52; Visibility: 16.09km; Sun: 7:(
0:36 / 3:14
HD

Can this be **generalized**, e.g., supply chains more broadly?

Explain-a-LOD!




Explain-a-LOD

Basic dataset information

Number of instances:	177
Number of generated features:	41
Source attribute:	country
Target attribute:	index

Simple explanations | Complex rules

A country of type *OECDMemberEconomies*
has high index
Correlation: 0.6164

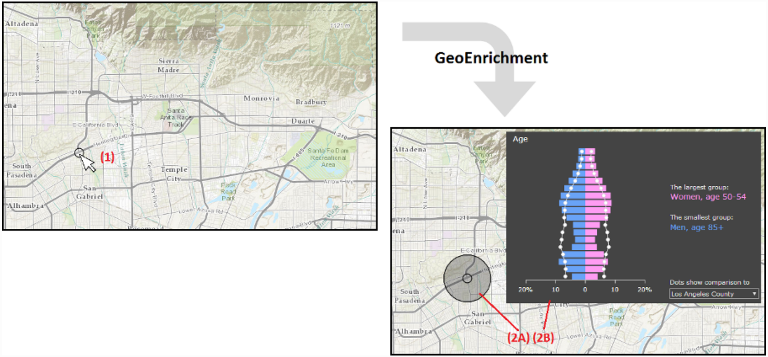
A country of type *EuropeanUnionMemberStates*
has high index
Correlation: 0.4227

A country of type *LeastDevelopedCountries*
has low index
Correlation: -0.3904

A country of type *EuropeanUnionMemberEconomies*
has high index
Correlation: 0.3858

‘Linked Open Data as a means to interpret statistics [and collect] attributes to generate hypotheses for **explaining**[...]’ (Paulheim, 2012)

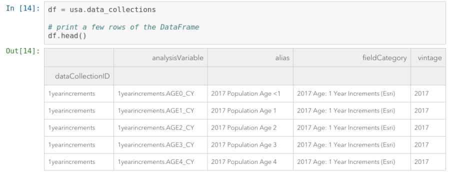
GeoEnrichment



‘Enriches your data by adding demographic and landscape facts about the people and places that surround or are inside your data locations. The output is a duplicate of **your input with new attribute fields** added to the table.’ (Esri)

Esri's GeoEnrichment

Fetch data



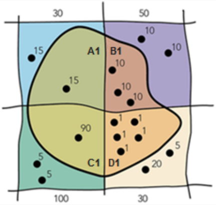
```
In [14]: df = usa.data_collections

         # print a few rows of the DataFrame
         df.head()
```

Out[14]:

	analysisVariable	alias	fieldCategory	vintage
dataCollectionID				
1yearincrements	1yearincrements.AGE0_CY	2017 Population Age <1	2017 Age: 1 Year Increments (Esri)	2017
1yearincrements	1yearincrements.AGE1_CY	2017 Population Age 1	2017 Age: 1 Year Increments (Esri)	2017
1yearincrements	1yearincrements.AGE2_CY	2017 Population Age 2	2017 Age: 1 Year Increments (Esri)	2017
1yearincrements	1yearincrements.AGE3_CY	2017 Population Age 3	2017 Age: 1 Year Increments (Esri)	2017
1yearincrements	1yearincrements.AGE4_CY	2017 Population Age 4	2017 Age: 1 Year Increments (Esri)	2017

Data apportionment



Weighted Block %

	Pop	
A1	15	(15/30 = 50% x 30)
B1	30	(30/50 = 60% x 50)
C1	90	(90/100 = 90% x 100)
D1	5	(5/30 = 16% x 30)

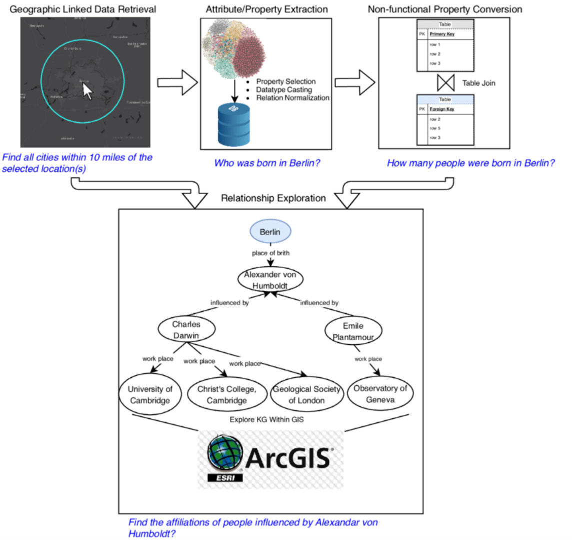
(figures from Esri)

Imagine you could **supercharge GeoEnrichment**

- 'Things not strings' (**semantic** graph data)
- Open-ended, up-to-date **open content** across domains instead of pre-selected themes.
- **Traverse** the graph instead of flat tables

▶ **Reenvision GeoEnrichment by fusing it with Linked Data!**

Regions Studied by Explorers Influenced by von Humboldt



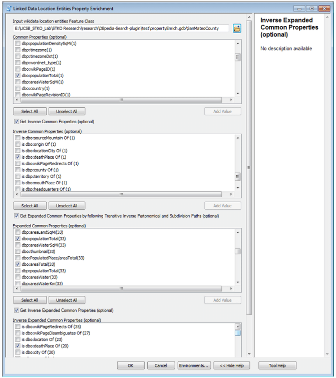
Find all cities within 10 miles of the selected location(s)

Who was born in Berlin?

How many people were born in Berlin?

Find the affiliations of people influenced by Alexandar von Humboldt?

Deeply Integrating Linked Data with GIS



Linked Data Location Entities Property Enrichment

Input wikidata location entities Feature Class

E:\UCSB_STKO_Lab\STKO Research\research\DBpedia-Search-plugin\test\propertyEnrich.gdb\SanMateoCounty

Common Properties (optional)

- □ dbp:populationDensitySqMi(1)
- □ dbp:timezone(1)
- □ dbp:timezoneDst(1)
- □ dbp:wordnet_type(1)
- □ dbo:wikiPageID(1)
- ☑ dbo:populationTotal(1)
- □ dbp:areaWaterSqMi(1)
- □ dbo:country(1)
- □ dbo:wikiPageRevisionID(1)

☑ Get Inverse Common Properties (optional)

Inverse Common Properties (optional)

- □ is dbo:sourceMountain Of (1)
- □ is dbo:origin Of (1)
- □ is dbo:locationCity Of (1)
- ☑ is dbo:deathPlace Of (1)
- □ is dbo:wikiPageRedirects Of (1)
- □ is dbp:county Of (1)
- □ is dbp:territory Of (1)
- □ is dbo:mouthPlace Of (1)
- □ is dbp:headquarters Of (1)

☑ Get Expanded Common Properties by following Transitive Inverse Partonomical and Subdivision Paths (optional)

Expanded Common Properties (optional)

- □ dbp:areaLandSqMi(33)
- ☑ dbo:populationTotal(33)
- □ dbp:areaWaterSqMi(33)
- □ dbo:thumbnail(33)
- □ dbo:PopulatedPlace/areaTotal(33)
- ☑ dbo:areaTotal(33)
- □ dbp:populationTotal(33)
- □ dbo:areaWater(33)
- □ dbp:areaWaterKm(33)

☑ Get Inverse Expanded Common Properties (optional)

Inverse Expanded Common Properties (optional)

- □ is dbo:wikiPageRedirects Of (35)
- □ is dbo:wikiPageDisambiguates Of (27)
- □ is dbo:location Of (23)
- ☑ is dbo:deathPlace Of (20)
- □ is dbo:city Of (20)

Inverse Expanded Common Properties (optional)

No description available

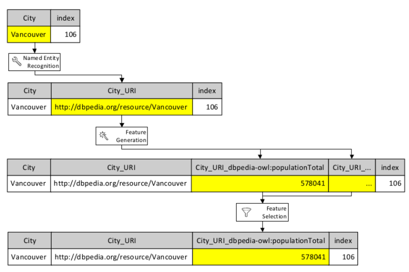
(Mai, G., Janowicz, K., Yan, B., & Scheider, S. (2019). Deeply integrating linked data with geographic information systems. Transactions in GIS, 23(3))

Many well-known challenges:

- Ontologies (ODP) + microtheories vs. learning representations
- Ontology alignment
- Co-reference resolution/ deduplication
- Data fusion / conflation
- KG summarization
- Link prediction
- **Spatial and temporal scopes**
  - What is an ‘essential service’ varies by state (e.g., COVID-19)
  - How long do we have to socially distance

Do not worry, we are not trying to solve the **semantic interoperability** problem ;-)

Named Entity Recognition, Disambiguation, and Linking



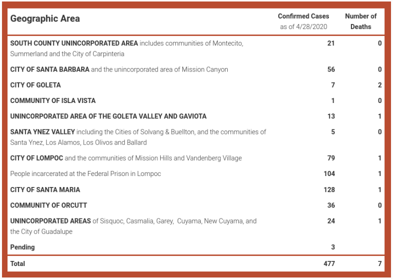
City	index
Vancouver	106

City	City_URI	index
Vancouver	http://dbpedia.org/resource/Vancouver	106

City	City_URI	City_URI_dbpedia-owl:populationTotal	City_URI_...	index
Vancouver	http://dbpedia.org/resource/Vancouver	578041	...	106

City	City_URI	City_URI_dbpedia-owl:populationTotal	index
Vancouver	http://dbpedia.org/resource/Vancouver	578041	106

(Paulheim; ESWC 2012)

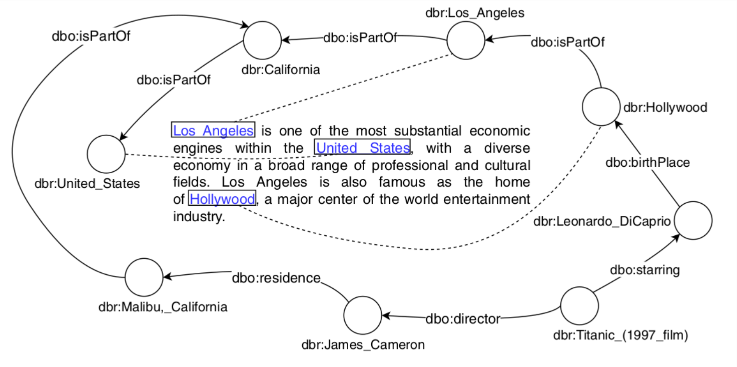


Geographic Area	Confirmed Cases as of 4/28/2020	Number of Deaths
SOUTH COUNTY UNINCORPORATED AREA includes communities of Montecito, Summerland and the City of Carpinteria	21	0
CITY OF SANTA BARBARA and the unincorporated area of Mission Canyon	56	0
CITY OF GOLETA	7	2
COMMUNITY OF ISLA VISTA	1	0
UNINCORPORATED AREA OF THE GOLETA VALLEY AND GAVIOTA	13	1
SANTA YNEZ VALLEY including the Cities of Solvang & Buellton, and the communities of Santa Ynez, Los Alamos, Los Olivos and Ballard	5	0
CITY OF LOMPOC and the communities of Mission Hills and Vandenberg Village	79	1
People incarcerated at the Federal Prison in Lompoc	104	1
CITY OF SANTA MARIA	128	1
COMMUNITY OF ORCUTT	36	0
UNINCORPORATED AREAS of Sisquoc, Casmalia, Garey, Cuyama, New Cuyama, and the City of Guadalupe	24	1
Pending	3	
Total	477	7

(publichealthsbc.org)

Key problem for Explain-a-LOD like tools, e.g., what if there is no 1:1 correspondence for NER? How would data apportionment work for graphed data?

KG Summarization

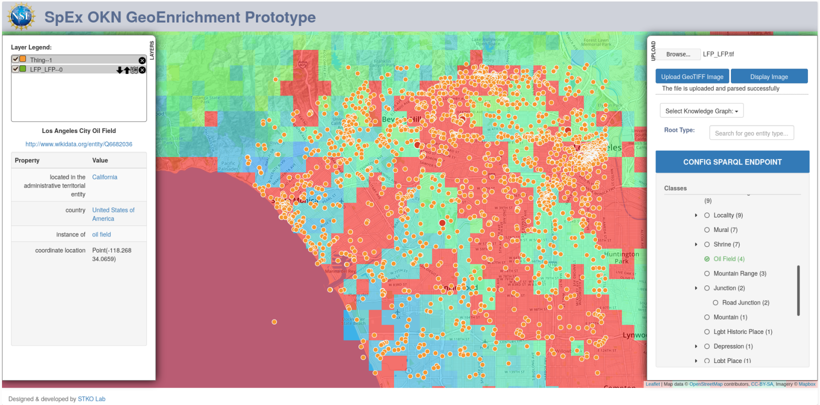


Los Angeles is one of the most substantial economic engines within the United States, with a diverse economy in a broad range of professional and cultural fields. Los Angeles is also famous as the home of Hollywood, a major center of the world entertainment industry.

Many steps required: Data apportionment, relevant features –> **KG summarization**; n-degree path queries –> requires (multi-hop) **link prediction** to handle sparsity

(Yan, B., Janowicz, K., Mai, G., & Zhu, R. (2019). A spatially explicit reinforcement learning model for geographic knowledge graph summarization. Transactions in GIS, 23(3), 620-640.)

Phase I Experiments



SpEx OKN GeoEnrichment Prototype

Layer Legend:

☑ Thing--1

☑ LFP_LFP--0

Los Angeles City Oil Field

http://www.wikidata.org/entity/Q6682036

Property	Value
located in the administrative territorial entity	California
country	United States of America
instance of	oil field
coordinate location	Point(-118.268 34.0659)

LFP_LFP.tif

The file is uploaded and parsed successfully

Root Type: Search for geo entity type...

Classes

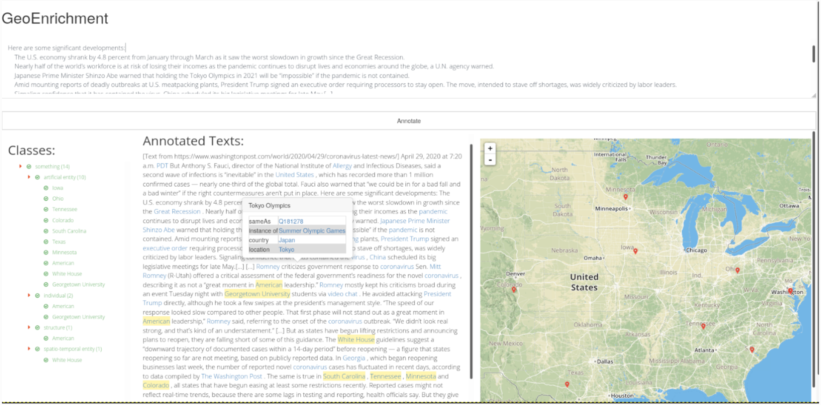
(9)
- ○ Locality (9)
- ○ Mural (7)
- ○ Shrine (7)
- ✓ Oil Field (4)
- ○ Mountain Range (3)
- ○ Junction (2)
  - ○ Road Junction (2)
- ○ Mountain (1)
- ○ Lgbt Historic Place (1)
- ○ Depression (1)
- ○ Lgbt Place (1)

Leaflet | Map data © OpenStreetMap contributors, CC-BY-SA, Imagery © Mapbox

Designed & developed by STKO Lab

Can we **enrich** remotely sensed information or model outputs with auxiliary data (not visible from the sky) such as events that took place in a region?

Phase I Experiments



GeoEnrichment

Here are some significant developments:
The U.S. economy shrank by 4.8 percent from January through March as it saw the worst slowdown in growth since the Great Recession.
Nearly half of the world's workforce is at risk of losing their incomes as the pandemic continues to disrupt lives and economies around the globe, a U.N. agency warned.
Japanese Prime Minister Shinzo Abe warned that holding the Tokyo Olympics in 2021 will be "impossible" if the pandemic is not contained.
Amid mounting reports of deadly outbreaks at U.S. meatpacking plants, President Trump signed an executive order requiring processors to stay open. The move, intended to stave off shortages, was widely criticized by labor leaders.
Signaling confidence that it has contained the virus, China scheduled its big legislative meetings for late May [...]

Annotate

Classes:

- something (14)
  - artificial entity (10)
    - Iowa
    - Ohio
    - Tennessee
    - Colorado
    - South Carolina
    - Texas
    - Minnesota
    - American
    - White House
    - Georgetown University
  - individual (2)
    - American
    - Georgetown University
  - structure (1)
    - American
  - spatio-temporal entity (1)
    - White House

Annotated Texts:

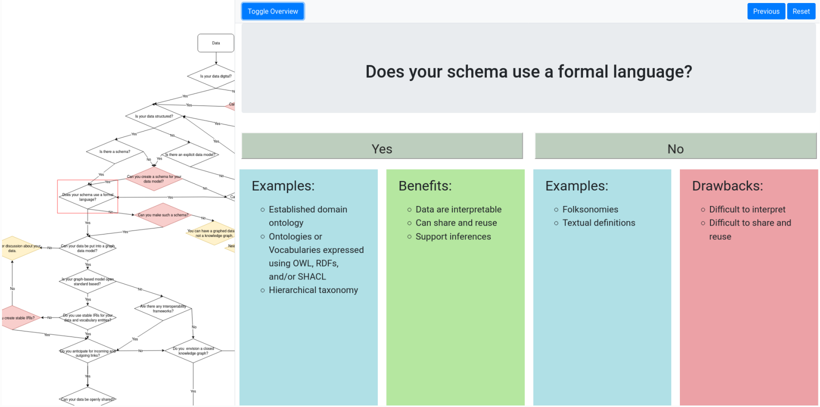
[Text from https://www.washingtonpost.com/world/2020/04/29/coronavirus-latest-news/] April 29, 2020 at 7:20 a.m. PDT But Anthony S. Fauci, director of the National Institute of Allergy and Infectious Diseases, said a second wave of infections is "inevitable" in the United States , which has recorded more than 1 million confirmed cases — nearly one-third of the global total. Fauci also warned that "we could be in for a bad fall and a bad winter" if the right countermeasures aren't put in place. Here are some significant developments: The
U.S. economy shrank by 4.8 percer … v the worst slowdown in growth since
the Great Recession . Nearly half o … g their incomes as the pandemic
continues to disrupt lives and econ … warned. Japanese Prime Minister
Shinzo Abe warned that holding th … ssible" if the pandemic is not
contained. Amid mounting reports … ng plants, President Trump signed an
executive order requiring processo … o stave off shortages, was widely
criticized by labor leaders. Signaling … virus , China scheduled its big
legislative meetings for late May.[...] [...] Romney criticizes government response to coronavirus Sen. Mitt Romney (R-Utah) offered a critical assessment of the federal government's readiness for the novel coronavirus , describing it as not a "great moment in American leadership." Romney mostly kept his criticisms broad during an event Tuesday night with Georgetown University students via video chat . He avoided attacking President Trump directly, although he took a few swipes at the president's management style. "The speed of our response looked slow compared to other people. That first phase will not stand out as a great moment in American leadership," Romney said, referring to the onset of the coronavirus outbreak. "We didn't look real strong, and that's kind of an understatement." [...] But as states have begun lifting restrictions and announcing plans to reopen, they are falling short of some of this guidance. The White House guidelines suggest a "downward trajectory of documented cases within a 14-day period" before reopening — a figure that states reopening so far are not meeting, based on publicly reported data. In Georgia , which began reopening businesses last week, the number of reported novel coronavirus cases has fluctuated in recent days, according to data compiled by The Washington Post . The same is true in South Carolina , Tennessee , Minnesota and Colorado , all states that have begun easing at least some restrictions recently. Reported cases might not reflect real-time trends, because there are some lags in testing and reporting, health officials say. But they give

Tokyo Olympics

sameAs	Q181278
Instance of	Summer Olympic Games
country	Japan
location	Tokyo

Can we improve the **geographic** components of successful commercial systems such as **OpenCalais**, e.g., by introducing the concept of a **region**?

Phase I Experiments


Data
Is your data digital?
Yes
No
Is your data structured?
Is there a schema?
Is there an explicit data model?
Can you create a schema for your data model?
Does your schema use a formal language?
Can you make such a schema?
You can have a graphed data not a knowledge graph.
er discussion about your data.
Can your data be put into a graph data model?
Is your graph-based model open standard based?
Are there any interoperability frameworks?
u create stable IRIs?
Do you use stable IRIs for your data and vocabulary entities?
Do you anticipate for incoming and outgoing links?
Do you envision a closed knowledge graph?
Can your data be openly shared?
Toggle Overview
Previous
Reset

Does your schema use a formal language?

Yes

Examples:

- Established domain ontology
- Ontologies or Vocabularies expressed using OWL, RDFs, and/or SHACL
- Hierarchical taxonomy

Benefits:

- Data are interpretable
- Can share and reuse
- Support inferences

No

Examples:

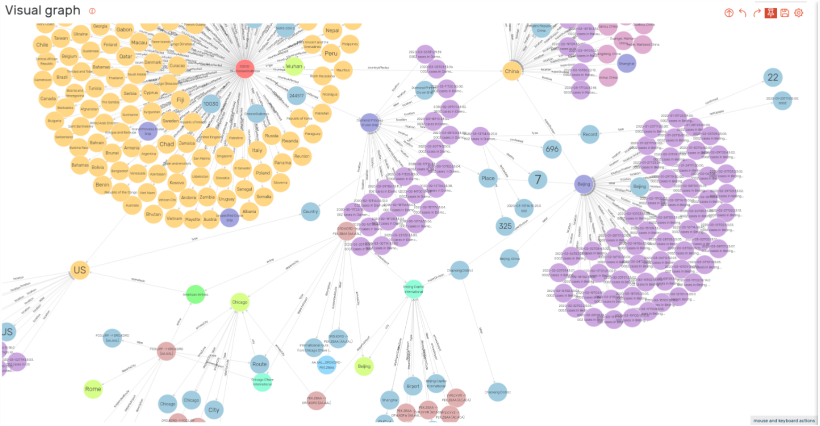
- Folksonomies
- Textual definitions

Drawbacks:

- Difficult to interpret
- Difficult to share and reuse

Knowledge graphs are about direct access to AI-ready, data-level statements (not dataset search). How can we guide industry in understanding whether their data is **graph-ready**?

Our COVID-19 Knowledge Graph



Geographically integrating data across domains with a focus on social distancing measures, transportation, supply chain disruptions, canceled local events, geographic regions...

KnowWhereGraph Vision

- (Geographic) space and time matter not only for the obvious reason that everything happens somewhere and at some time, but because knowing **where** and **when** things happen is critical to understanding **why** and how they happened or will happen.
- With 'KnowWhere' Graph we hope to take **GeoEnrichment** to the next level, by providing **open** graph-based linking and semantic enrichment technologies far beyond pre-defined data themes and silos.
- The ultimate goal of our project is to understand how to **engineer** meaningful **features** (independent variables) via knowledge graph-based GeoEnrichment for **downstream models** such as supply chain forecasting or soil health mapping
- This remains a **key roadblock** for our partners in industry, governments, and non-profits.