SAGE: a Semantic Annotator for knowledge Graph Exploration

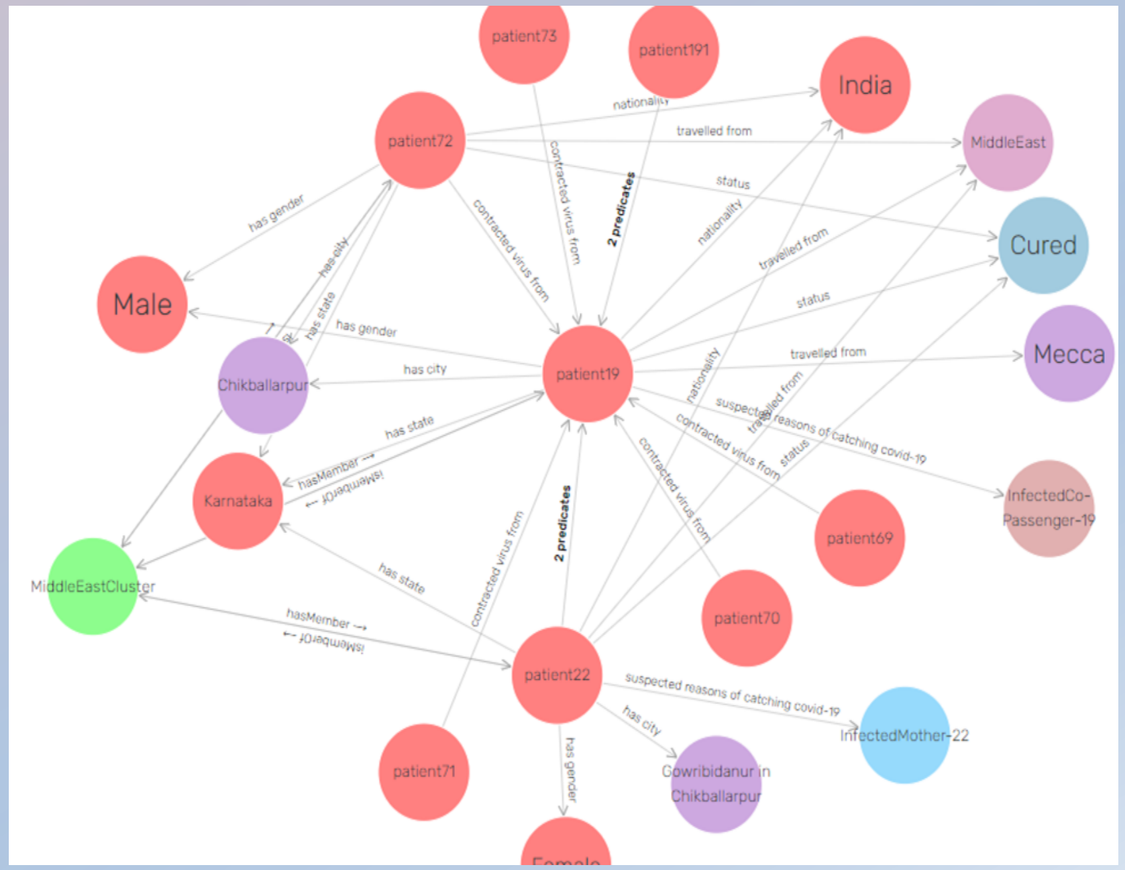
Dr. Biswanath Dutta
Associate Professor
DRTC, Indian Statistical Institute, Bangalore

Puranjani Das
Project Associate
DRTC, Indian Statistical Institute, Bangalore

ASIS&T Mid-Year Conference “Expanding Horizons of Information Science and Technology and Beyond” (virtual, April 11-13, 2023)

Outline

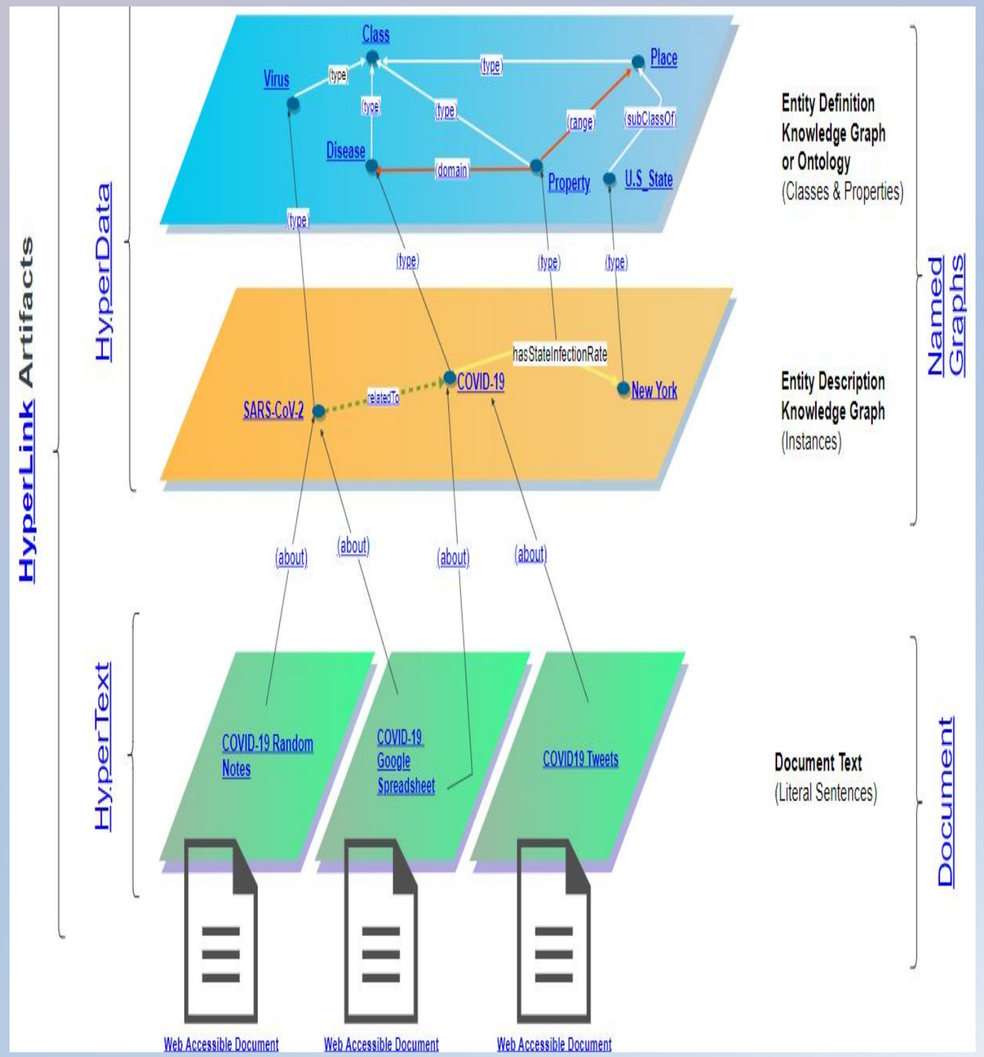
- Introduction
- Related Works
- Motivation
- SAGE
- SAGE Features
- SAGE Applications
- SAGE Architecture & Design Approach
- Evaluation
- Conclusion
- Future Work



Introduction

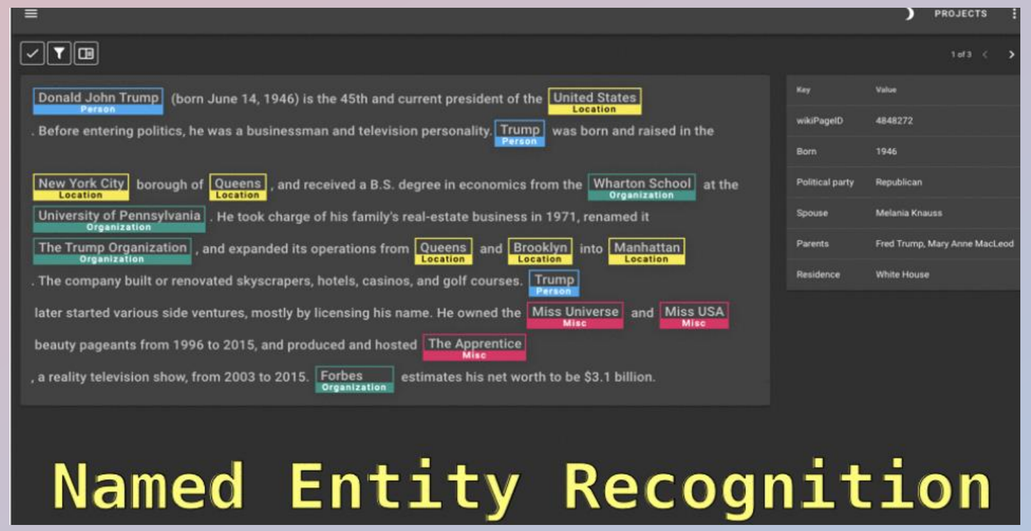
Knowledge graph (KG):
a representation of an intelligent web of data that is informed by an ontology.

Applications:
question-answer systems, semantic search and retrieval, information integration, data visualization and exploration, and automatic **annotation**

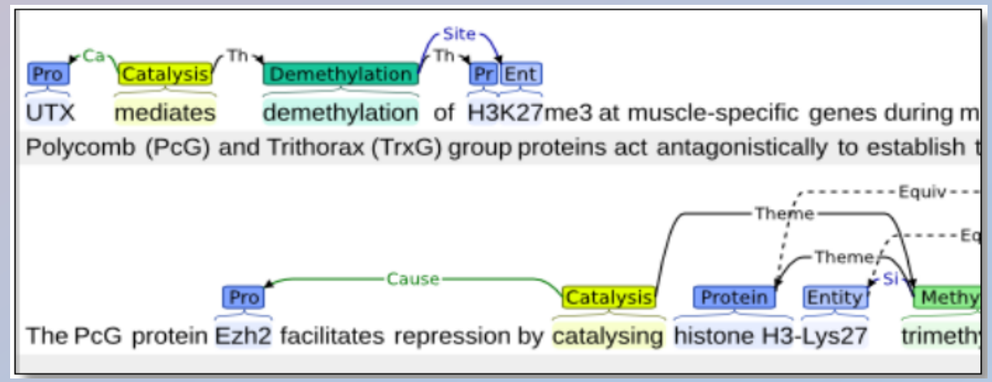


[1]

Existing Annotators



https://github.com/doccano/doccano



https://brat.nlplab.org/examples.html

Existing Annotators (contd…2)

https://www.dbpedia-spotlight.org/

FICLONE [5], MTab4D [6], LinkingPark [7]

Motivation

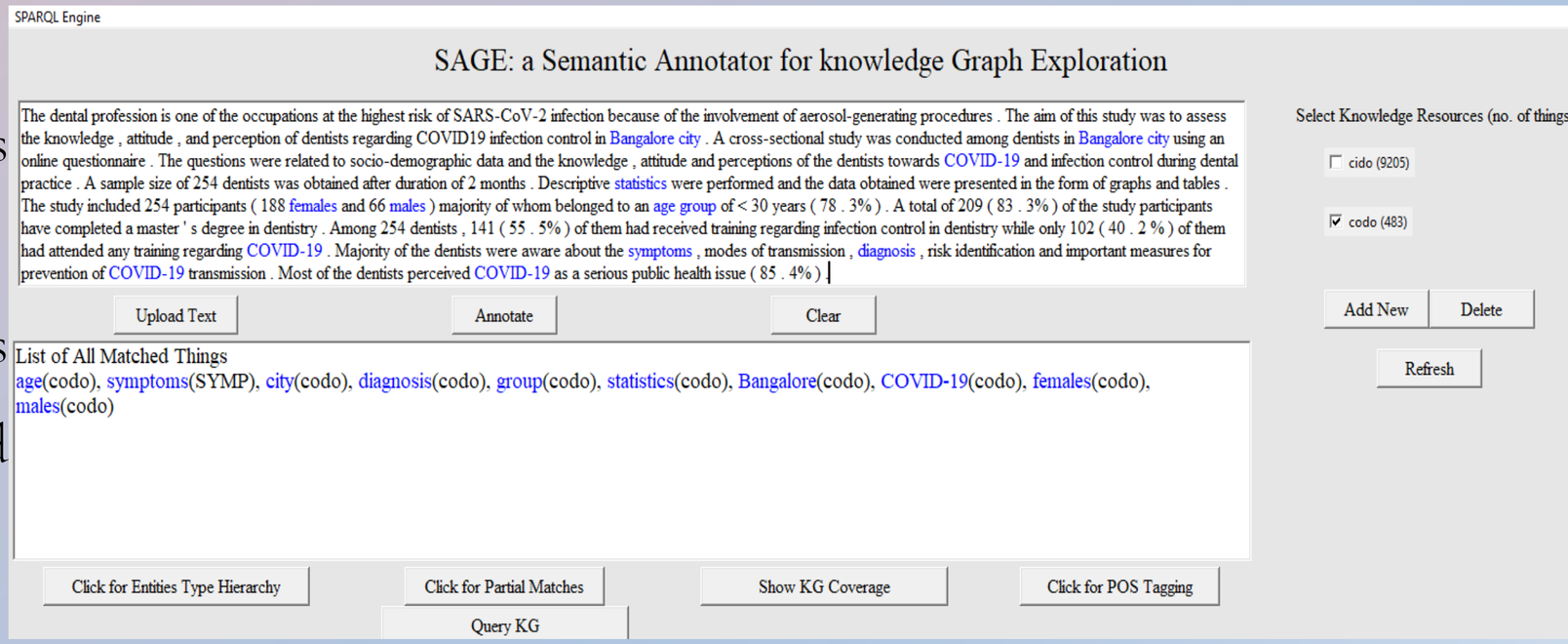
- Annotation of "**Things**"
- A personalized environment for KG exploration

  (priority: domain need(s) and annotation tasks)
- Design of an user-friendly and inclusive system
- A facilitator for enhancement of KG

Annotation refers to a process of spotting, linking, and extracting information about the "things" in the input text from the KG

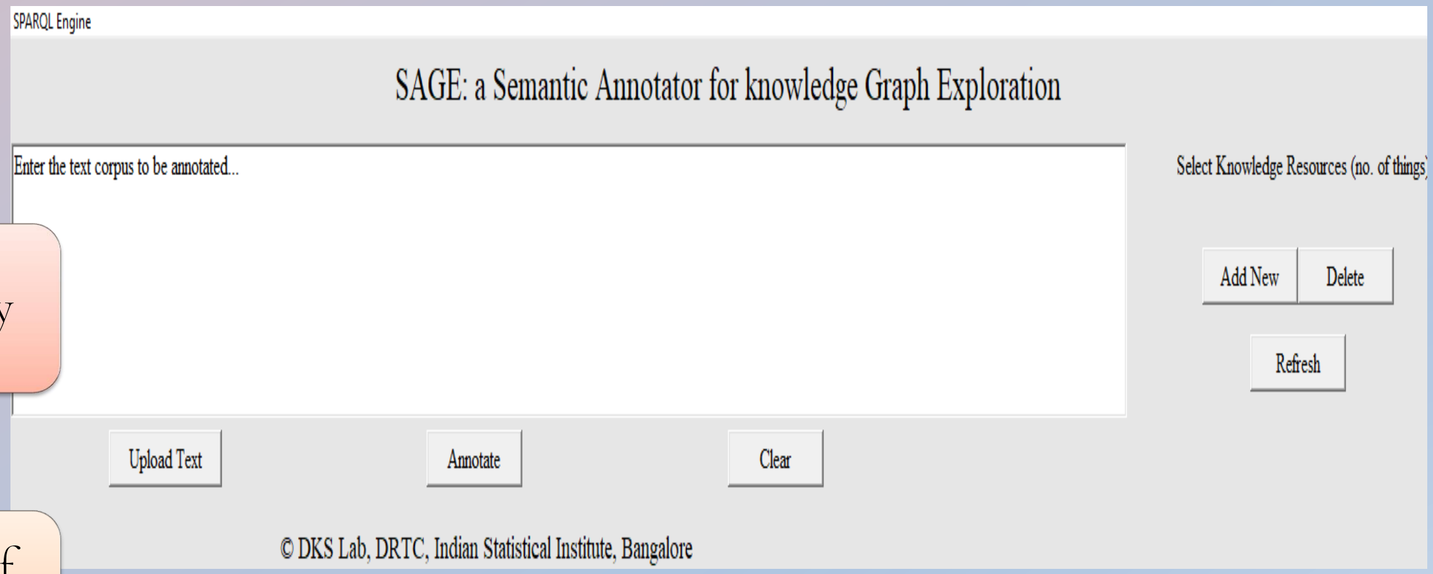
SAGE

- SAGE is a semantic annotator
- Annotates strings in text as things from the KG
- Thing found in the user-given text is annotated, linked, negotiated, and explored locally and/ or on the Web



SAGE Features

- Exploration of KG
- SPARQL Query Interface
- Predefined Query
- Treeview of Entity Types
- Identification of Missing Things
- Estimation of KG Coverage and KG Statistics

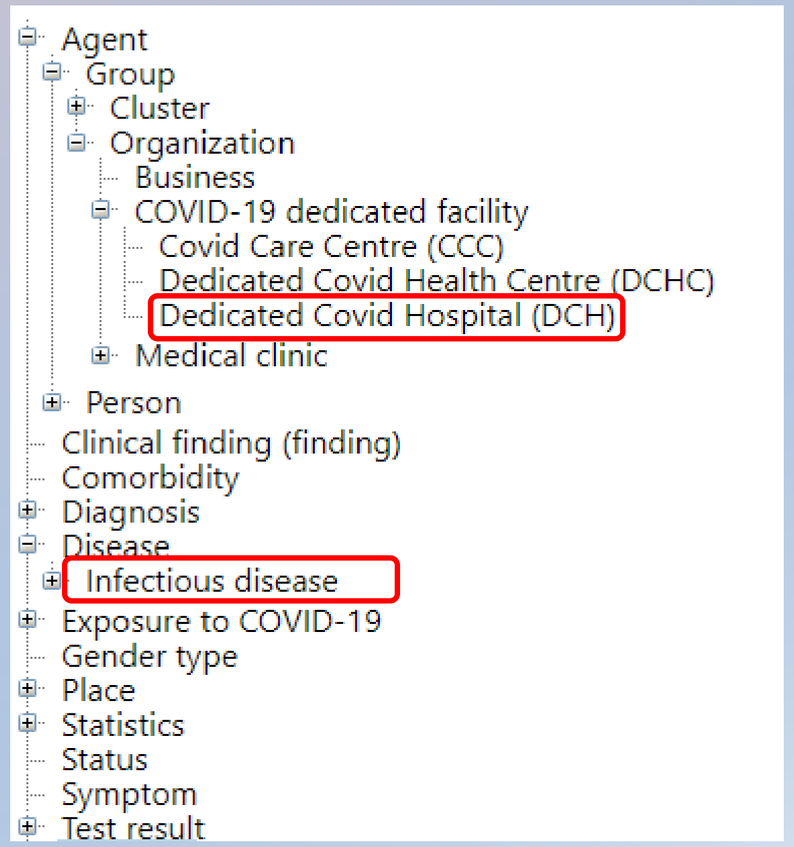


Dutta and Das (2023) [10]

Exploration of KG

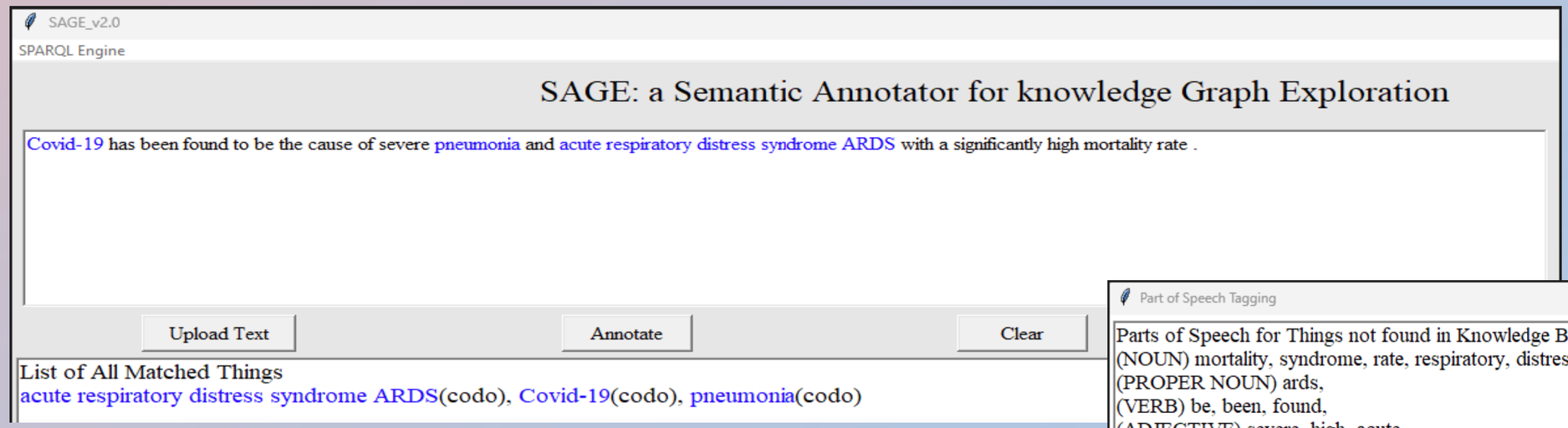
"Diya got affected by an infectious disease while working in a hospital."

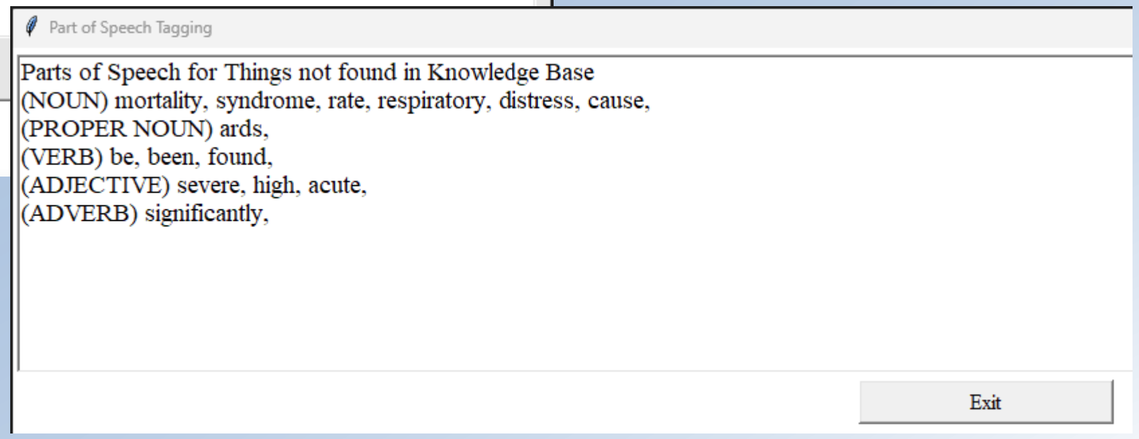
	Terms Identified	Things Retrieved
Exact Match	"infectious disease"	"infectious disease"
Partial Match	"hospital"	"Dedicated Covid Hospital (DCH)"



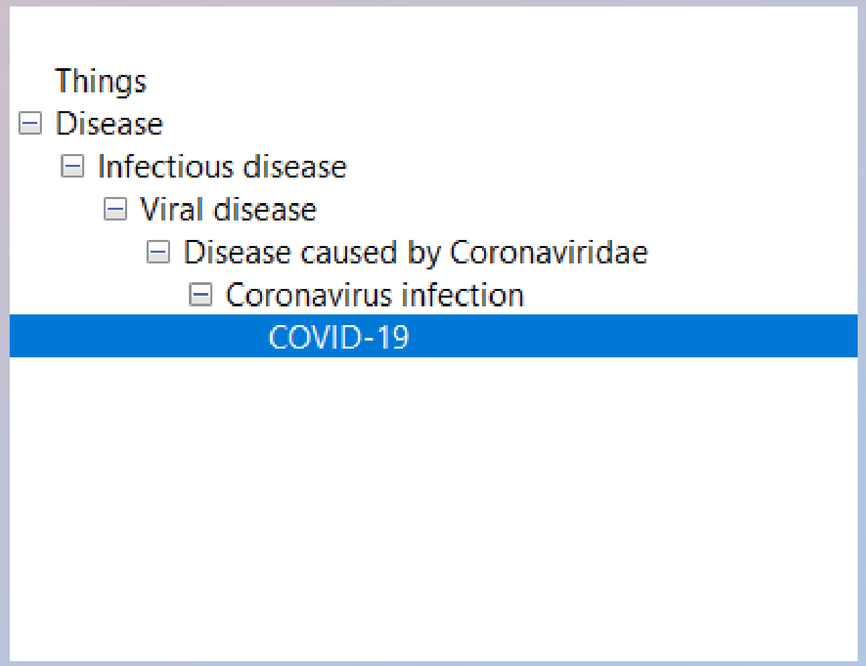
Identification of Missing Things

"COVID-19 has been found to be the cause of severe pneumonia and acute respiratory distress syndrome (ARDS) with a significantly high mortality rate."





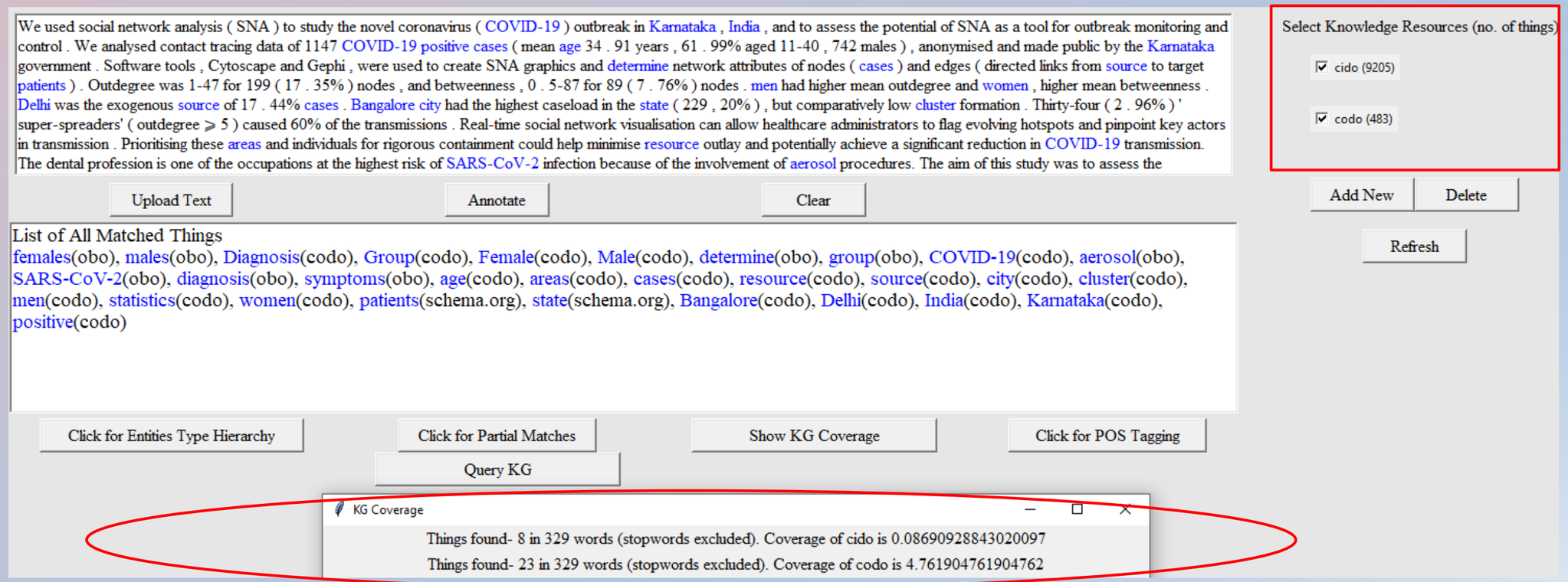
Tree view of Entity Types



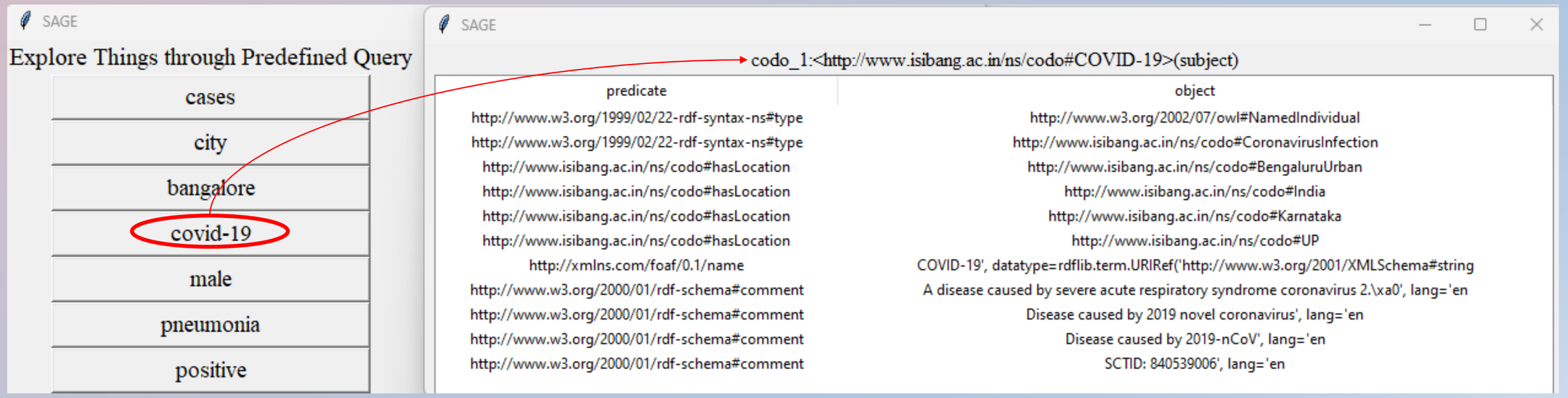
Hierarchy obtained from SAGE for **COVID-19** from CODO ontology

Estimation of KG Coverage & Statistics

$$Coverage\ \% = \frac{no.\ of\ exactly\ matched\ things\ in\ text\ from\ KG}{total\ no.\ of\ things\ in\ KG} \times 100$$



Predefined Query



SPARQL Query Interface

Querying the KGs with SPARQL

Enter SPARQL Endpoint URI

KG to be Queried: codo

Upload KG

```
CODO
codo: http://www.isibang.ac.in/ns/codo#
 dc: http://purl.org/dc/elements/1.1/
 foaf: http://xmlns.com/foaf/0.1/
 metadata: http://data.bioontology.org/metadata/
 mod: http://www.isibang.ac.in/ns/mod#
 ontology: http://omv.ontoware.org/2005/05/ontology#
 owl: http://www.w3.org/2002/07/owl#
```

```
select * where {?s ?p ?o} limit 10
```

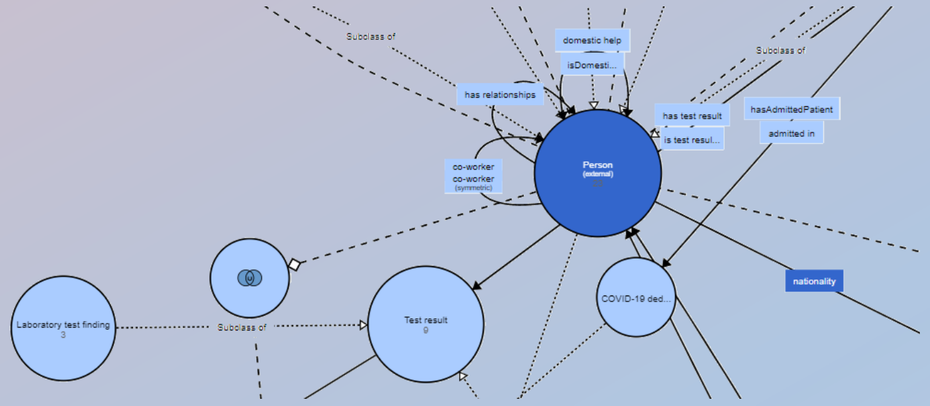
Execute Query

codo

1	2	3
https://schema.org/Country	http://www.isibang.ac.in/ns/codo#Philippines	http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://www.isibang.ac.in/ns/codo#p	N0b405a3e10b049068c02ec01b97980c2	http://www.w3.org/2003/11/swrl#argument1
http://www.w3.org/2002/07/owl#NamedIndividual	http://www.isibang.ac.in/ns/codo#karntStateStat000001	http://www.w3.org/1999/02/22-rdf-syntax-ns#type
A son of your brother or sister.', lang='en	http://www.isibang.ac.in/ns/codo#hasNephew	http://www.w3.org/2000/01/rdf-schema#comment
RestOfEurope', lang='en	http://www.isibang.ac.in/ns/codo#RestOfEurope	http://www.w3.org/2000/01/rdf-schema#label
http://www.w3.org/2002/07/owl#Class	http://www.isibang.ac.in/ns/codo#SecondaryContact	http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://www.isibang.ac.in/ns/codo#UnionTerritory	http://www.isibang.ac.in/ns/codo#DadraAndNagarHaveliAndDama	http://www.w3.org/1999/02/22-rdf-syntax-ns#type
HimachalPradesh', lang='en	http://www.isibang.ac.in/ns/codo#HimachalPradesh	http://www.w3.org/2000/01/rdf-schema#label
http://www.isibang.ac.in/ns/codo#hasAdmittedPatient	http://www.isibang.ac.in/ns/codo#admittedIn	http://www.w3.org/2002/07/owl#inverseOf
http://www.isibang.ac.in/ns/codo#City	http://www.isibang.ac.in/ns/codo#Dammam	http://www.w3.org/1999/02/22-rdf-syntax-ns#type

SAGE Applications

Library Tools
(E.g., Digital
Library System,
Abstracting
Databases, etc.)

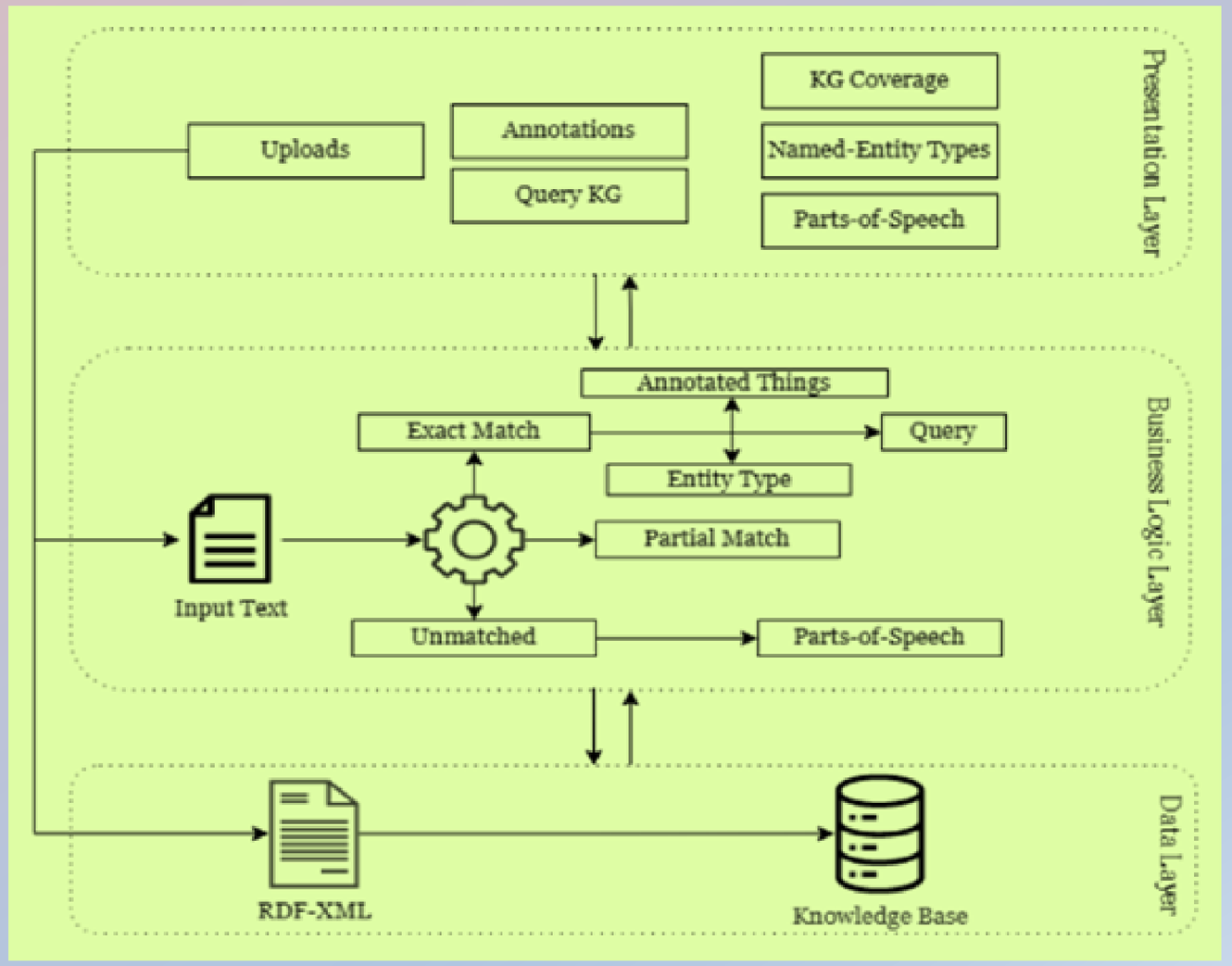


Ontology Development
(E.g., Identification of Gaps)

Scientific Text Exploration
(E.g., Medical Transcription)

Comparison of Coverage of Knowledge Resources

SAGE Architecture



Design Approach

Knowledge Base created from input KG

Algorithm 2 string_found
Require: string1 and string2
Ensure: matched version of string1 in string2 or False

```
S1: DEFINE string_found(string1, string2)
S2:     if string1.lower() in string2.lower() then
S3:             l=string2.lower().index(string1.lower())
S4:             token=string2[l: l+len(string1)]
S5:             return token
S6:     else if plural(string1.lower()) in string2.lower() then
S7:             return plural(string1)
S8:     else if singular(string1.lower()) in string2.lower() then
S9:             return singular(string1)
S10:    return False
```

string found function defined for string matching throughout SAGE (in Exact and Partial matches)

Sl. No.	Thing Category	Query Built
1	Class	SELECT ?s ?p WHERE {?s ?p <URI of thing>}
2	Property	SELECT ?s ?o WHERE {?s <URI of thing> ?o}
3	Instance	SELECT ?p ?o WHERE {<URI of thing> ?p ?o}

Queries defined for the exact matches found

Dutta and Das (2023) [11]

Evaluation

1. Precision and Recall of Partial Matches

Terms from text	Terms extracted from Knowledge Graph
virus	contracted virus from
acute	Acute Respiratory Distress Syndrome (ARDS)
repiratory	repiratory rate
coronavirus	Coronavirus infection
health	Dedicated Covid Health Centre (DCHC)
Recall	5
Correct recall	5
Precision	100%

Partial Match results **without** using WordNet

Average Precision: 100%

Terms from Text	Terms extracted from Knowledge Graph										
example	cases	daily increased cases	case id	covid-19 case on							
virus	contracted virus from										
know	no. of bed shortage	no. of beds needed	no. of icu beds needed	no. of icu bed shortage							
naming	name										
facilitate	domestic help										
human	Man										
acute	Acute Respiratory Distress Syndrome (ARDS)										
repiratory	repiratory rate										
coronavirus	Coronavirus infection										
following	next test result										
world	Man										
health	Dedicated Covid Health Centre (DCHC)										
nation	country code	state wise statistics	country wise statistics	state patient ID	Country	State wise statistics	State	has country	Country wise	has state	state code	
Recall	13	3	3	3	1	1	1	1	1	1	1	29
Correct Recall	11	2	2	2	1	1	1	1	1	1	1	24
Precision	82.76%											

Partial Match results **after** using WordNet

Average Precision: ~88%

Resources (datasets) used:
CODO (https://w3id.org/codo)
Input Text: [12-14]

Dutta and Das (2023) [11]

Evaluation (contd…2)

2. Features comparison

Features	DBpedia Spotlight	SAGE
Annotation (of exact things)	✓	✓
Retrieval (of similar 'things', not complete matches)	X	✓
Things Retrieved		
Object Properties	X	✓
Datatype Properties	X	✓
Classes	✓	✓
Named Individuals	✓	✓
Other Features		
Type information	✓	✓
Spotting Missing Terms	X	✓
Upload Text Corpus	✓	✓
Customized KB supported	X	✓
Predefined Query	X	✓
Coverage of KB	X	✓

Dutta and Das (2023) [10 & 11]

3. Response Time for primary annotation

Ontology Name	Total Things	Length of text before removing stopwords	Annotation Time (seconds)
COVID-19 (CIDO)	11598	200	10.448144674301147
		500	17.92897057533264
		1000	29.03405261039734
		2000	51.816343784332275

Relationship between the time taken and length of input text taken from Ciotti et al. (2020) [9]

Summary

- SAGE is a semantic annotator and use it for
  - Mapping/ spotting/ discovering things in the text
  - Retrieval of facts about things in text
  - Querying and exploring things from multiple KGs from a single platform
  - Comparing the coverage of KGs and ontologies
  - Identifying missing things/terms in an ontology, knowledge graph

SAGE: Semantic Annotator for knowledge Graph Exploration

https://tinyurl.com/yc8p5nm3

SAGE (V1.1) is a desktop application for "thing" annotation. Here, "thing" refers to any concept (aka class), named individuals (aka entities), entity relations (aka object properties), and attributes (aka data properties). The system utilizes existing knowledge graphs (KGs) to convert any given input text into annotated things. The annotated "thing" can then either be browsed on the Web or retrieved along with its associated facts (axioms) from the knowledge base without writing a SPARQL query.

SAGE's GUI makes it easy for users with less technological expertise to upload, annotate, and explore things from the KGs. The system displays the entity type hierarchy. It also shows the number of things available in the KGs and calculates their coverage against the input text.

Download:

SAGE v2.0 (March 6, 2023)

SAGE v1.0

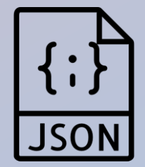
SAGE primary features

(*features written in red color are the new additions to the SAGE v2.0*)

- Personalization- SAGE GUI provides an easy way to select and upload knowledge graphs on the system based on individual domain needs and annotation tasks.
- Exact Match- SAGE annotates the exact matched things within the text from the KGs. By clicking the annotated things, we can then browse them on the Web.
- Partial Match- SAGE provides an annotated list of partially matched things, in order to provide the user with contextually relevant resources (when no exact match is found).
  - SAGE v2.0 Partial match is further enhanced by WordNet.
- Treeview of entity types.
- Predefined Query- Apart from annotating and exploring things on the Web as described above, SPARQL SELECT queries are defined on exact matches, which return all the tuples associated with the matched things. These queries are defined for each matched entity and the user does not have to write them.

Future Works

Multi-lingual Support




```
<http://www.w3.org/People/Berners-Lee/card#i>
        <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
        <http://xmlns.com/foaf/0.1/Person> .
<http://www.w3.org/People/Berners-Lee/card#i>
        <http://xmlns.com/foaf/0.1/name>
        "Tim Berners-Lee"@en .
<http://www.w3.org/People/Berners-Lee/card#i>
        <http://xmlns.com/foaf/0.1/name>
        "Τιμ Μπέρνερς Λι"@gr .
```

Multiple ontology format support

SAGE as a web service

References

1. Idehen, Kingsley U. (2020). Linked Data, Ontologies, and Knowledge Graphs. https://www.linkedin.com/pulse/linked-data-ontologies-knowledge-graphs-kingsley-uyi-idehen/
2. Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011). DBpedia spotlight: Shedding light on the web of documents. ACM International Conference Proceeding Series, 1–8. https://doi.org/10.1145/2063518.2063519
3. BRAT rapid annotation tool. (n.d.). Retrieved December 13, 2022, from https://brat.nlplab.org/
4. Doccano, GitHub. (n.d.). Retrieved December 13, 2022, from https://github.com/doccano
5. Chabchoub, M., Gagnon, M. & Web, A. Z (2018). FICLONE: improving DBpedia spotlight using named entity recognition and collective disambiguation. Open Journal Semantic Web, 5(1): 12–28.
6. Nguyen, Phuc et al. (2022). MTab4D: Semantic Annotation of Tabular Data with DBpedia'. 1 Jan. 2022 : 1 – 25.
7. Shuang Chen, Alperen Karaoglu, Carina Negreanu, Tingting Ma, Jin-Ge Yao, Jack Williams, Feng Jiang, Andy Gordon, Chin-Yew Lin, (2022). LinkingPark: An automatic semantic table interpretation system. Journal of Web Semantics, 74. https://doi.org/10.1016/j.websem.2022.100733.
8. Hogenboom, F., Frasincar, F., & Kaymak, U. (2010). An overview of approaches to extract information from natural language corpora. Information Foraging Lab 69.
9. Ciotti, M., Ciccozzi, M., Terrinoni, A., Jiang, Wen-Can, Wang, Cheng-Bin, & Bernardini, Sergio (2020). The COVID-19 pandemic. Critical Reviews in Clinical Laboratory Sciences, 57(6): 365-388. https://doi.org/10.1080/10408363.2020.1783198
10. Dutta, B. and Das, Puranjani (2023). SAGE: A Semantic Annotator for knowledge Graph Exploration. In ASIS&T Mid-Year Conference "Expanding Horizons of Information Science and Technology and Beyond" (virtual, April 11-13, 2023) DOI: https://doi.org/10.5281/zenodo.7597207
11. Dutta, Biswanath. and Das, Puranjani. (2023). Semantic Annotator for Knowledge Graph Exploration: Pattern-Based NLP Technique. Journal of Information and Knowledge (Formerly SRELS Journal of Information Management), 60(1), 49-62. https://doi.org/10.17821/srels/2023/v60i1/170889
12. WHO. Naming the coronavirus disease. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it
13. Lotfi, M., Hamblin, M. R., & Rezaei, N. (2020). COVID-19: Transmission, prevention, and potential therapeutic opportunities. Clinica chimica acta; international journal of clinical chemistry, 508, 254–266. https://doi.org/10.1016/j.cca.2020.05.044
14. Andersen, J. P., Nielsen, M. W., Simone, N. L., Lewiss, R. E., & Jagsi, R. (2020). COVID-19 medical papers have fewer women first authors than expected. eLife, 9, e58807. https://doi.org/10.7554/eLife.58807

Thank you

Acknowledgement

This work is executed under the research project entitled "Integrated and Unified Data Model for Publication and Sharing of prolonged pandemic data as FAIR Semantic Data: COVID-19 as a case study", funded by Indian Statistical Institute Kolkata.

Contact

Biswanath Dutta, Ph.D.

Email: dutta2005@gmail.com

bisu@isibang.ac.in

bisu@isibang.ac.in

Twitter: @biswanath_dutta