

SAGE: a Semantic Annotator for knowledge Graph Exploration

Dr. Biswanath Dutta

Associate Professor

DRTC, Indian Statistical Institute, Bangalore

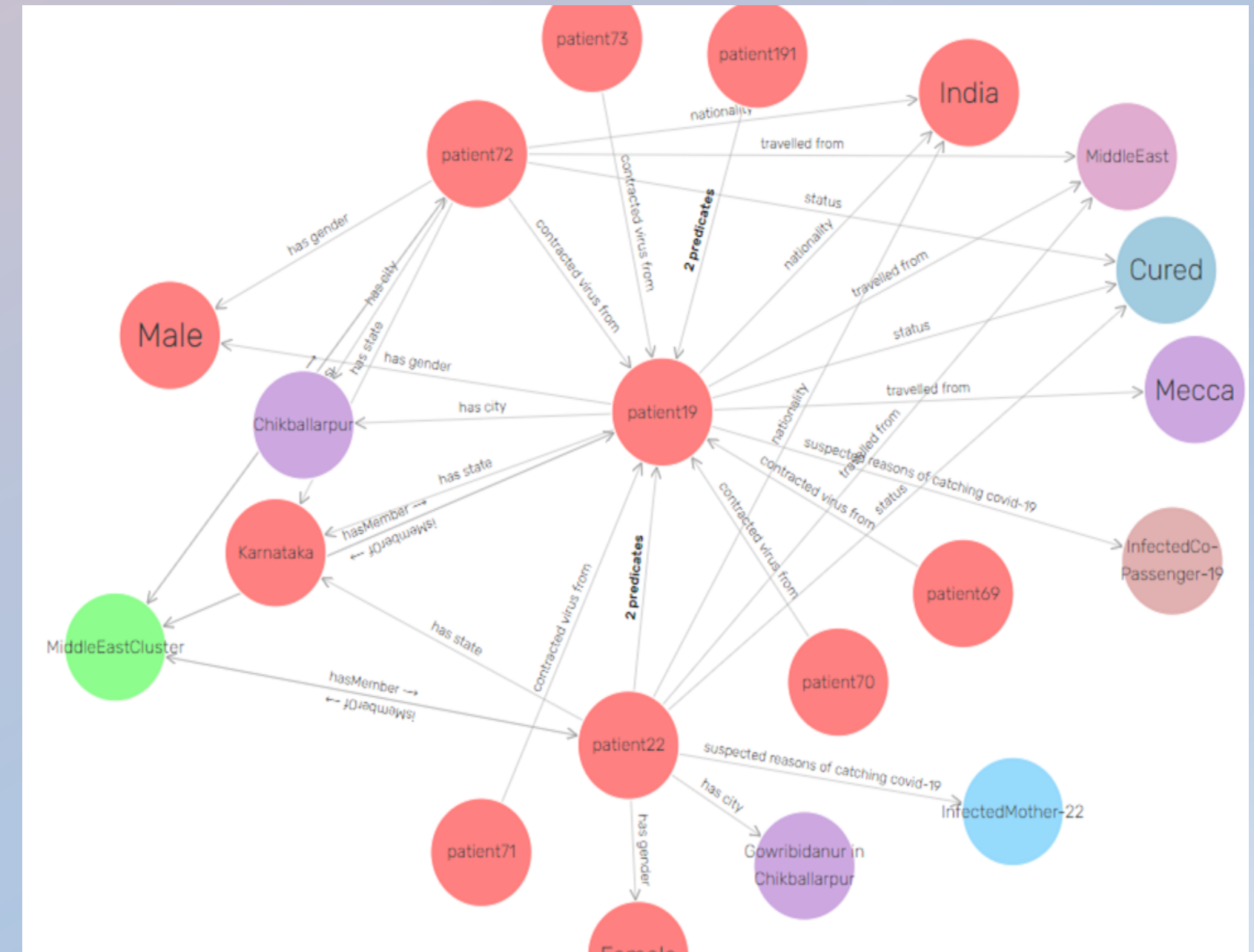
Puranjani Das

Project Associate

DRTC, Indian Statistical Institute, Bangalore

Outline

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



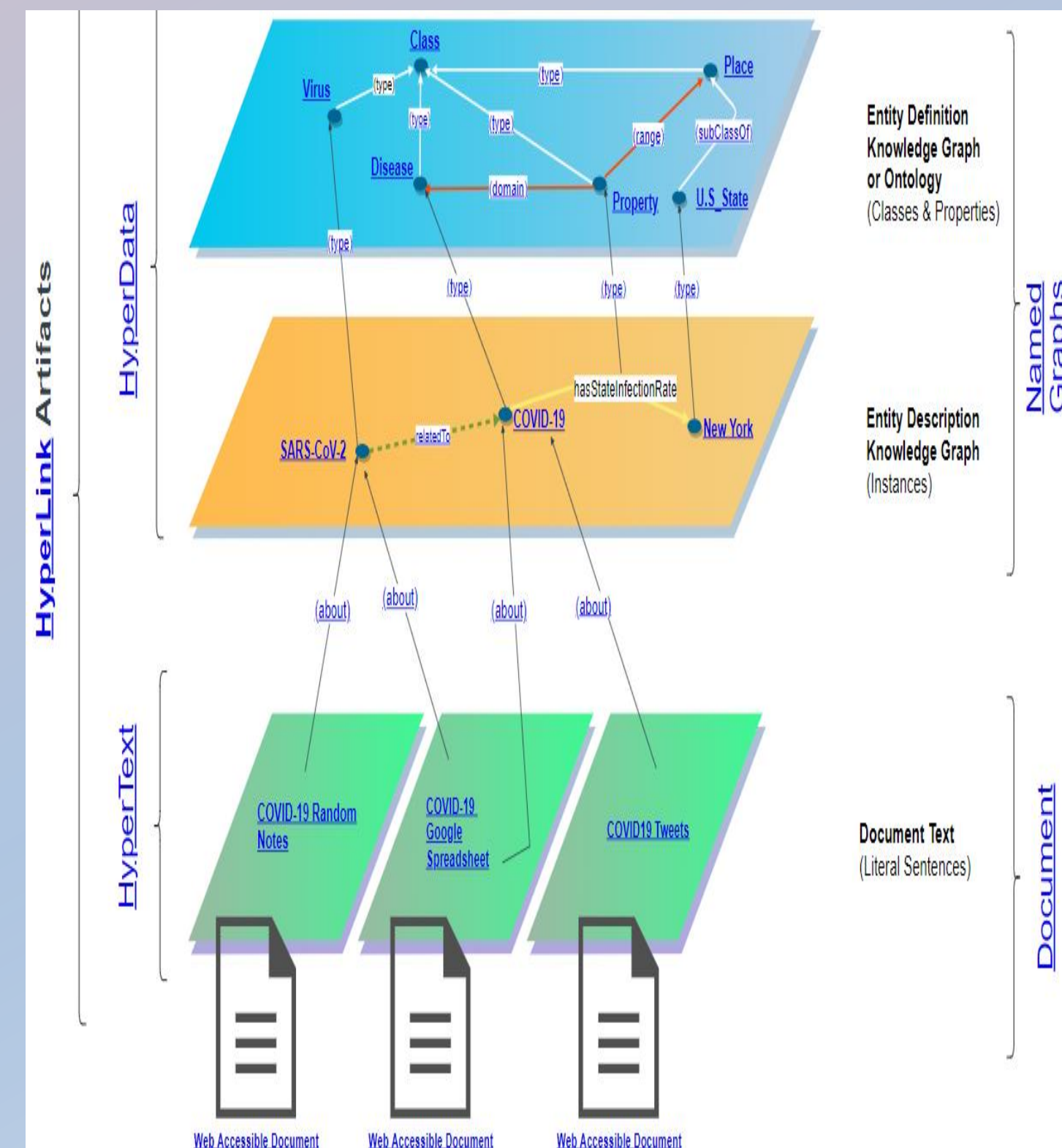
Introduction

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**



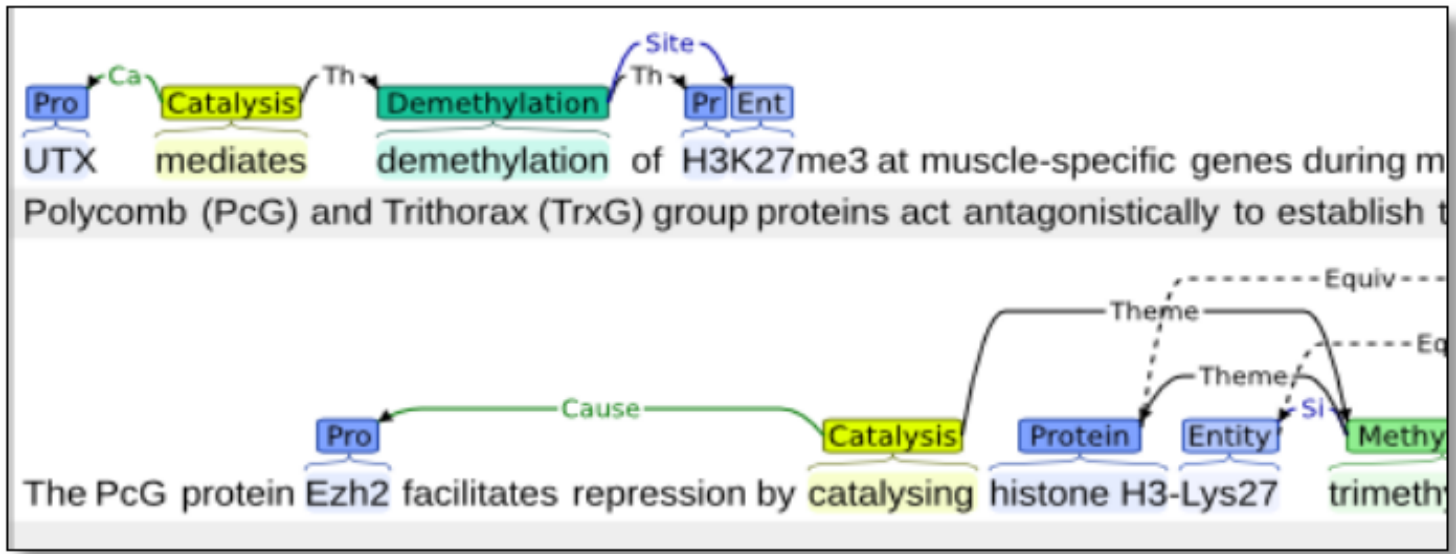
Existing Annotators

Donald John Trump (born June 14, 1946) is the 45th and current president of the United States. Before entering politics, he was a businessman and television personality. Trump was born and raised in the New York City borough of Queens, and received a B.S. degree in economics from the Wharton School at the University of Pennsylvania. He took charge of his family's real-estate business in 1971, renamed it The Trump Organization, and expanded its operations from Queens and Brooklyn into Manhattan. The company built or renovated skyscrapers, hotels, casinos, and golf courses. Trump later started various side ventures, mostly by licensing his name. He owned the Miss Universe and Miss USA beauty pageants from 1996 to 2015, and produced and hosted The Apprentice, a reality television show, from 2003 to 2015. Forbes estimates his net worth to be \$3.1 billion.

Key	Value
wikiPageID	4848272
Born	1946
Political party	Republican
Spouse	Melania Knauss
Parents	Fred Trump, Mary Anne MacLeod
Residence	White House

Named Entity Recognition

<https://github.com/doccano/doccano>



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



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

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

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

SPARQL Engine

SAGE: a Semantic Annotator for knowledge Graph Exploration

The dental profession is one of the occupations at the highest risk of SARS-CoV-2 infection because of the involvement of aerosol-generating procedures . The aim of this study was to assess the knowledge , attitude , and perception of dentists regarding COVID19 infection control in [Bangalore city](#) . A cross-sectional study was conducted among dentists in [Bangalore city](#) using an online questionnaire . The questions were related to socio-demographic data and the knowledge , attitude and perceptions of the dentists towards [COVID-19](#) and infection control during dental practice . A sample size of 254 dentists was obtained after duration of 2 months . Descriptive [statistics](#) were performed and the data obtained were presented in the form of graphs and tables . The study included 254 participants (188 [females](#) and 66 [males](#)) majority of whom belonged to an [age group](#) of < 30 years (78 . 3%) . A total of 209 (83 . 3%) of the study participants have completed a master ' s degree in dentistry . Among 254 dentists , 141 (55 . 5%) of them had received training regarding infection control in dentistry while only 102 (40 . 2 %) of them had attended any training regarding [COVID-19](#) . Majority of the dentists were aware about the [symptoms](#) , modes of transmission , [diagnosis](#) , risk identification and important measures for prevention of [COVID-19](#) transmission . Most of the dentists perceived [COVID-19](#) as a serious public health issue (85 . 4%) |

[Upload Text](#) [Annotate](#) [Clear](#)

Select Knowledge Resources (no. of things)

☐ cido (9205)

☒ codo (483)

[Add New](#) [Delete](#)

[Refresh](#)

List of All Matched Things

[age\(codo\)](#), [symptoms\(SYMP\)](#), [city\(codo\)](#), [diagnosis\(codo\)](#), [group\(codo\)](#), [statistics\(codo\)](#), [Bangalore\(codo\)](#), [COVID-19\(codo\)](#), [females\(codo\)](#), [males\(codo\)](#)

[Click for Entities Type Hierarchy](#) [Click for Partial Matches](#) [Show KG Coverage](#) [Click for POS Tagging](#)

[Query KG](#)

Exploration of KG

SPARQL Query
Interface

Predefined Query

SAGE Features

Treeview of Entity
Types

Identification of
Missing Things

Estimation of KG
Coverage and KG
Statistics

SPARQL Engine

SAGE: a Semantic Annotator for knowledge Graph Exploration

Enter the text corpus to be annotated...

Select Knowledge Resources (no. of things)

Add New Delete

Refresh

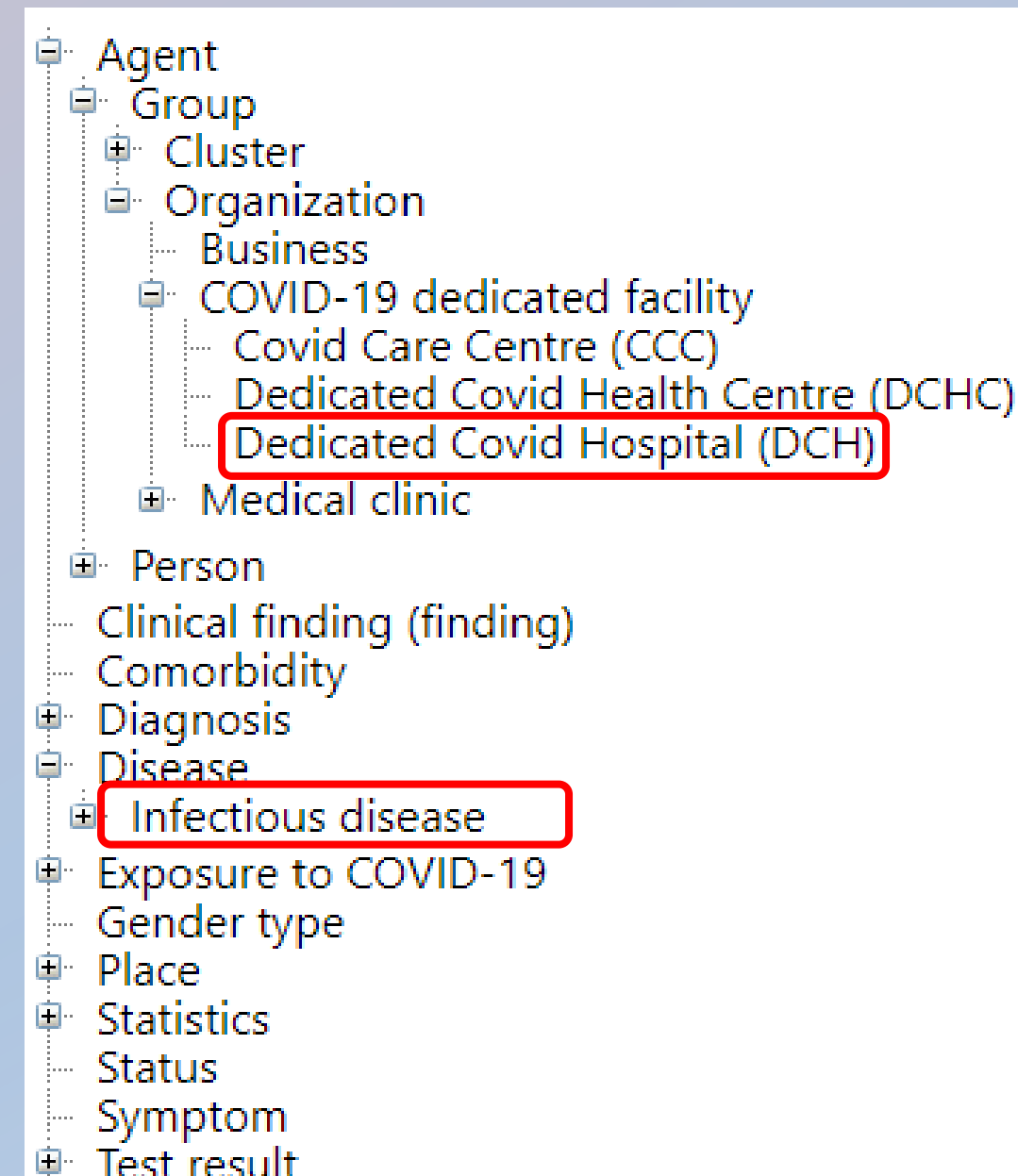
Upload Text Annotate Clear

© DKS Lab, DRTC, Indian Statistical Institute, Bangalore

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.”

SAGE_v2.0
SPARQL Engine

SAGE: a Semantic Annotator for knowledge Graph Exploration

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

Upload Text Annotate Clear

List of All Matched Things
acute respiratory distress syndrome ARDS(codo), Covid-19(codo), pneumonia(codo)

Part of Speech Tagging

Parts of Speech for Things not found in Knowledge Base
(NOUN) mortality, syndrome, rate, respiratory, distress, cause,
(PROPER NOUN) ards,
(VERB) be, been, found,
(ADJECTIVE) severe, high, acute,
(ADVERB) significantly,

Exit

Tree view of Entity Types

Things
[-] Disease
[-] Infectious disease
[-] Viral disease
[-] Disease caused by Coronaviridae
[-] Coronavirus infection
COVID-19

Hierarchy obtained from SAGE for
COVID-19 from CODO ontology

Estimation of KG Coverage & Statistics

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

We used social network analysis (SNA) to study the novel coronavirus (COVID-19) outbreak in Karnataka , India , and to assess the potential of SNA as a tool for outbreak monitoring and control . We analysed contact tracing data of 1147 COVID-19 positive cases (mean age 34 . 91 years , 61 . 99% aged 11-40 , 742 males) , anonymised and made public by the Karnataka government . Software tools , Cytoscape and Gephi , were used to create SNA graphics and determine network attributes of nodes (cases) and edges (directed links from source to target patients) . Outdegree was 1-47 for 199 (17 . 35%) nodes , and betweenness , 0 . 5-87 for 89 (7 . 76%) nodes . men had higher mean outdegree and women , higher mean betweenness . Delhi was the exogenous source of 17 . 44% cases . Bangalore city had the highest caseload in the state (229 , 20%) , but comparatively low cluster formation . Thirty-four (2 . 96%) ' super-spreaders' (outdegree ≥ 5) caused 60% of the transmissions . Real-time social network visualisation can allow healthcare administrators to flag evolving hotspots and pinpoint key actors in transmission . Prioritising these areas and individuals for rigorous containment could help minimise resource outlay and potentially achieve a significant reduction in COVID-19 transmission. The dental profession is one of the occupations at the highest risk of SARS-CoV-2 infection because of the involvement of aerosol procedures. The aim of this study was to assess the

Upload Text Annotate Clear

List of All Matched Things
females(obo), males(obo), Diagnosis(codo), Group(codo), Female(codo), Male(codo), determine(obo), group(obo), COVID-19(codo), aerosol(obo), SARS-CoV-2(obo), diagnosis(obo), symptoms(obo), age(codo), areas(codo), cases(codo), resource(codo), source(codo), city(codo), cluster(codo), men(codo), statistics(codo), women(codo), patients(schema.org), state(schema.org), Bangalore(codo), Delhi(codo), India(codo), Karnataka(codo), positive(codo)

Click for Entities Type Hierarchy Click for Partial Matches Show KG Coverage Click for POS Tagging

Query KG

Select Knowledge Resources (no. of things)

☒ cido (9205)

☒ codo (483)

Add New Delete

Refresh

KG Coverage

Things found- 8 in 329 words (stopwords excluded). Coverage of cido is 0.08690928843020097

Things found- 23 in 329 words (stopwords excluded). Coverage of codo is 4.761904761904762

Predefined Query

SAGE

Explore Things through Predefined Query

cases
city
bangalore
covid-19
male
pneumonia
positive

SAGE

codo_1:<http://www.isibang.ac.in/ns/codo#COVID-19>(subject)

predicate	object
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#NamedIndividual
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.isibang.ac.in/ns/codo#CoronavirusInfection
http://www.isibang.ac.in/ns/codo#hasLocation	http://www.isibang.ac.in/ns/codo#BengaluruUrban
http://www.isibang.ac.in/ns/codo#hasLocation	http://www.isibang.ac.in/ns/codo#India
http://www.isibang.ac.in/ns/codo#hasLocation	http://www.isibang.ac.in/ns/codo#Karnataka
http://www.isibang.ac.in/ns/codo#hasLocation	http://www.isibang.ac.in/ns/codo#UP
http://xmlns.com/foaf/0.1/name	COVID-19', datatype=rdflib.term.URIRef('http://www.w3.org/2001/XMLSchema#string
http://www.w3.org/2000/01/rdf-schema#comment	A disease caused by severe acute respiratory syndrome coronavirus 2.\xa0', lang='en
http://www.w3.org/2000/01/rdf-schema#comment	Disease caused by 2019 novel coronavirus', lang='en
http://www.w3.org/2000/01/rdf-schema#comment	Disease caused by 2019-nCoV', lang='en
http://www.w3.org/2000/01/rdf-schema#comment	SCTID: 840539006', lang='en

SPARQL Query Interface

Querying the KGs with SPARQL

Enter SPARQL Endpoint URI

KG to be Queried

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#>

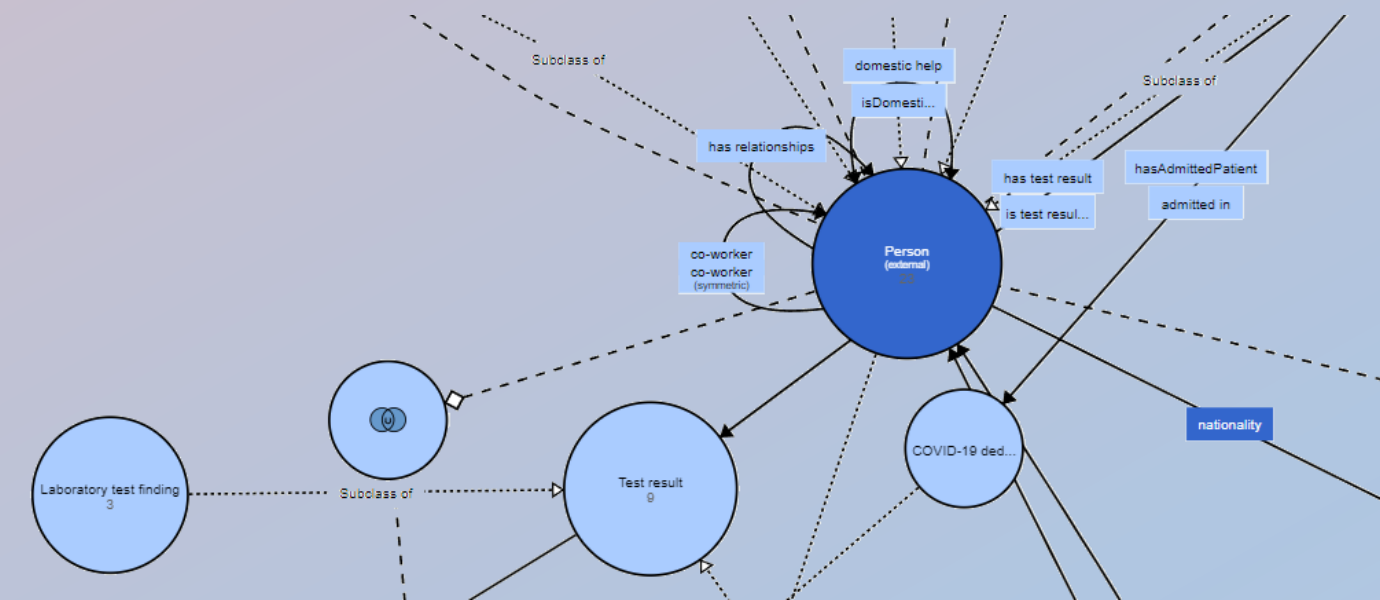
select * where {?s ?p ?o} limit 10

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)

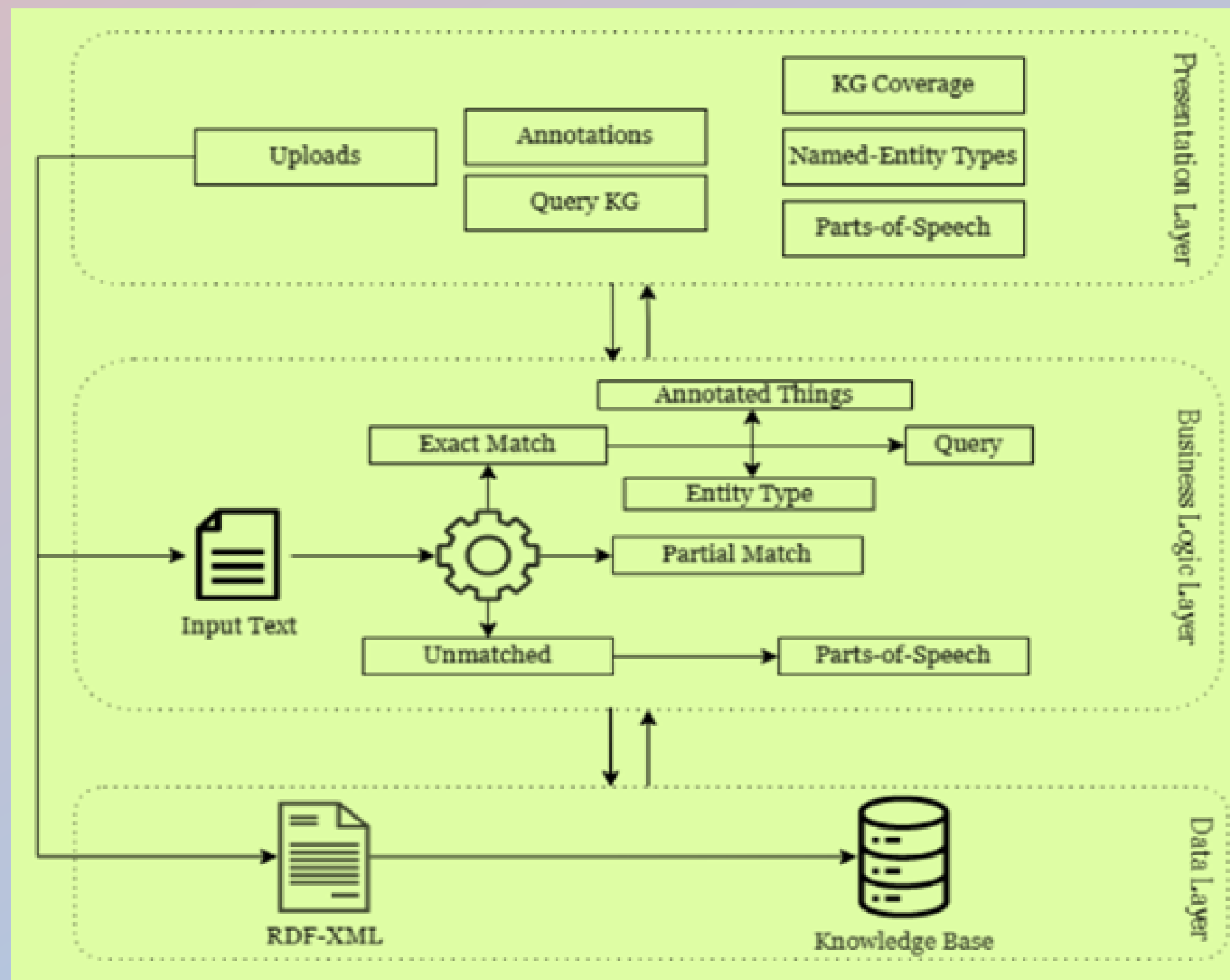


Scientific Text Exploration
(E.g., Medical Transcription)



Comparison of Coverage of Knowledge Resources

SAGE Architecture



Design Approach

```

{
  "https://w3id.org/codo#bedShortage": {
    "words": [
      "bed shortage",
      "bedShortage"
    ],
    "types": [],
    "subclassof": [],
    "subpropertyof": [
      "https://w3id.org/codo#resource"
    ],
    "inverseof": [],
    "comment": [
      "number of bed shortage."
    ],
    "domain": [
      "https://w3id.org/codo#Statistics"
    ],
    "range": [
      "http://www.w3.org/2001/XMLSchema#integer"
    ]
  },
  ...
}

```

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

Evaluation

1. Precision and Recall of Partial Matches

Terms from text	Terms extracted from Knowledge Graph
virus	contracted virus from
	Acute Respiratory Distress Syndrome (ARDS)
acute	
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%

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

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%

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	✓

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

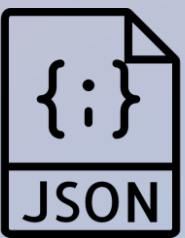
(features written in **red color** are the new additions to the SAGE v2.0)

- 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., Frasnica, 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](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](https://twitter.com/biswanath_dutta)