

CODO: an Ontology for collection and analysis of multi-paradigm COviD-19 data

Biswanath Dutta

Associate Professor, Indian Statistical Institute – Bangalore Centre
Bangalore 560059, INDIA

Email: dutta2005@gmail.com, bisu@isibang.ac.in

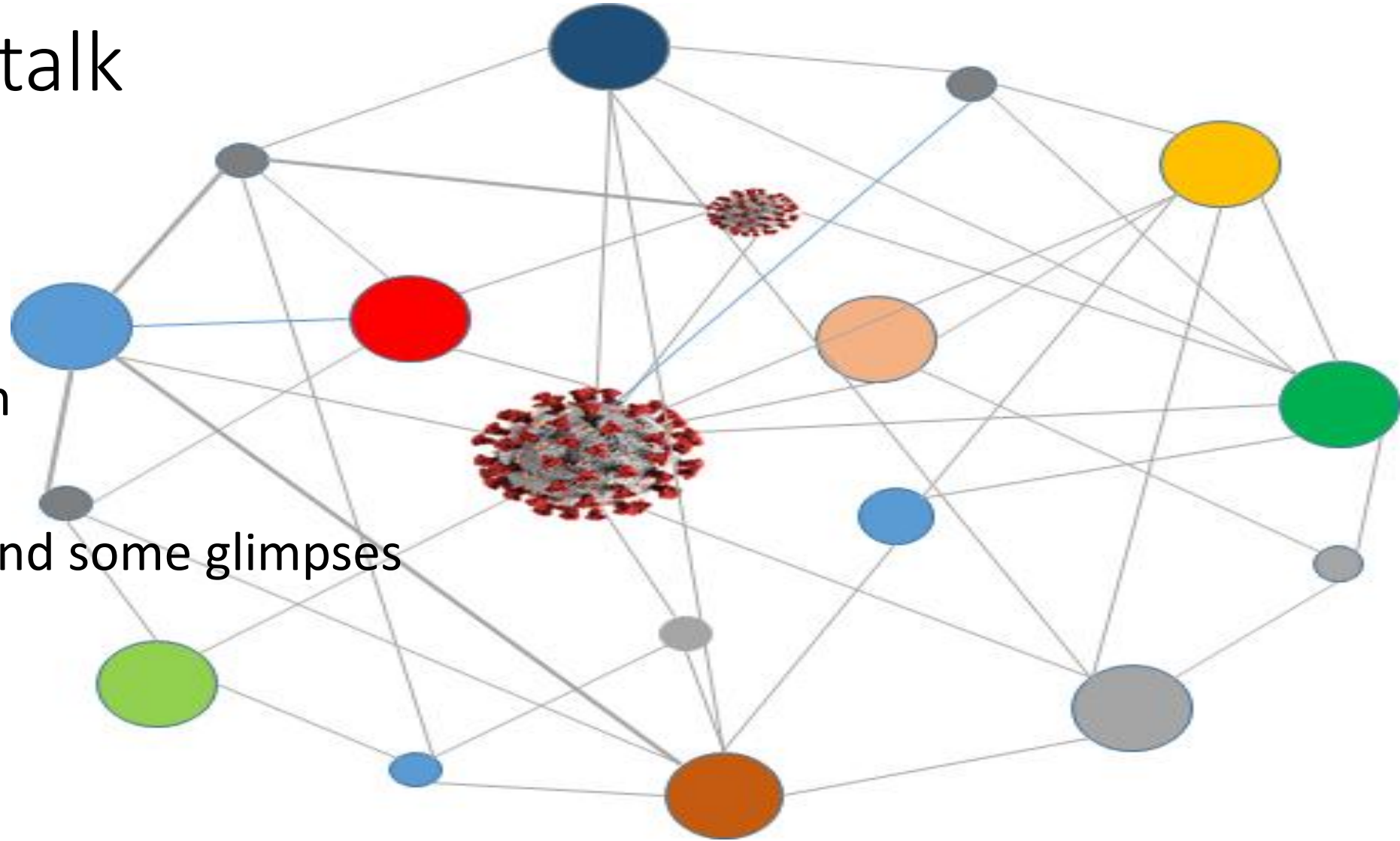
Michael DeBellis

<https://www.michaeldebellis.com/>

Ontology Summit 2022, March 9, 2022 (<https://bit.ly/36Qbu5o>)

Outline of talk

- Background
- Knowledge Graph
- CODO initiative and some glimpses
- Further work



Background

A	B	C	D	E	F	G	H
Case	Date	Age	Sex	City	State	Cluster	Reason
1	2020-03-09T00:00:00	41	Male	Bangalore	Karnataka	From USA	Texas US
2	2020-03-10T00:00:00	0	Female	Bangalore	Karnataka	From USA	Spouse
3	2020-03-10T00:00:00	13	Female	Bangalore	Karnataka	From USA	Daughter
4		0		Bangalore	Karnataka	From Unit	London
5	2020-03-13T00:00:00	26	Male	Bangalore	Karnataka	From the	Greece
6	2020-03-12T00:00:00	76	Male	Kalburgi	Karnataka	From Mid	Saudi Ara
7		0			Karnataka	Unknown	No detail:
8	2020-03-17T00:00:00	32	Male	Bangalore	Karnataka	From Unit	Co passer
9	2020-03-17T00:00:00	63	Male	Kalburgi	Karnataka	From Unit	Co passer
10	2020-03-17T00:00:00	20	Female	Bangalore	Karnataka	From Unit	UK

Situation by Country, Territory & Area

Name	Cases - cumulative total	Cases - newly reported in last 24 hours	Deaths - cumulative total
Global	23,518,343	206,382	810,492
 United States o...	5,649,928 	37,765	175,813
 Brazil	3,605,783 	23,421	114,744
 India	3,167,323 	60,975	58,390
 Russian Feder...	966,189 	4,696	16,568

In-depth: Excess mortality

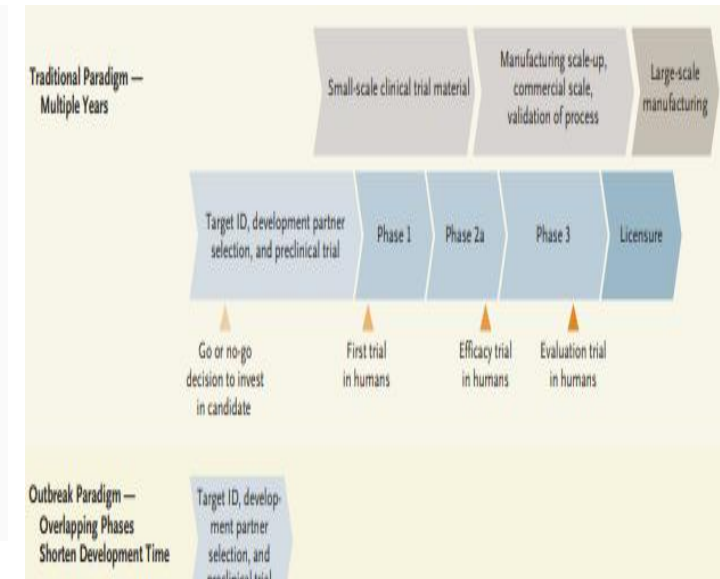
Excess mortality refers to the number of deaths *from all causes* above and beyond what we would have expected to see under 'normal' conditions.¹ In this case, we are interested in how deaths during the COVID-19 pandemic compare to the average number of deaths over the same period in previous years.

Looking at excess mortality is helpful for understanding the total impact of the pandemic on deaths – both direct and indirect. It helps us understand the direct impact by capturing deaths caused by COVID-19 that were not correctly diagnosed and reported, for example because no test for the virus was conducted. It helps us understand the indirect

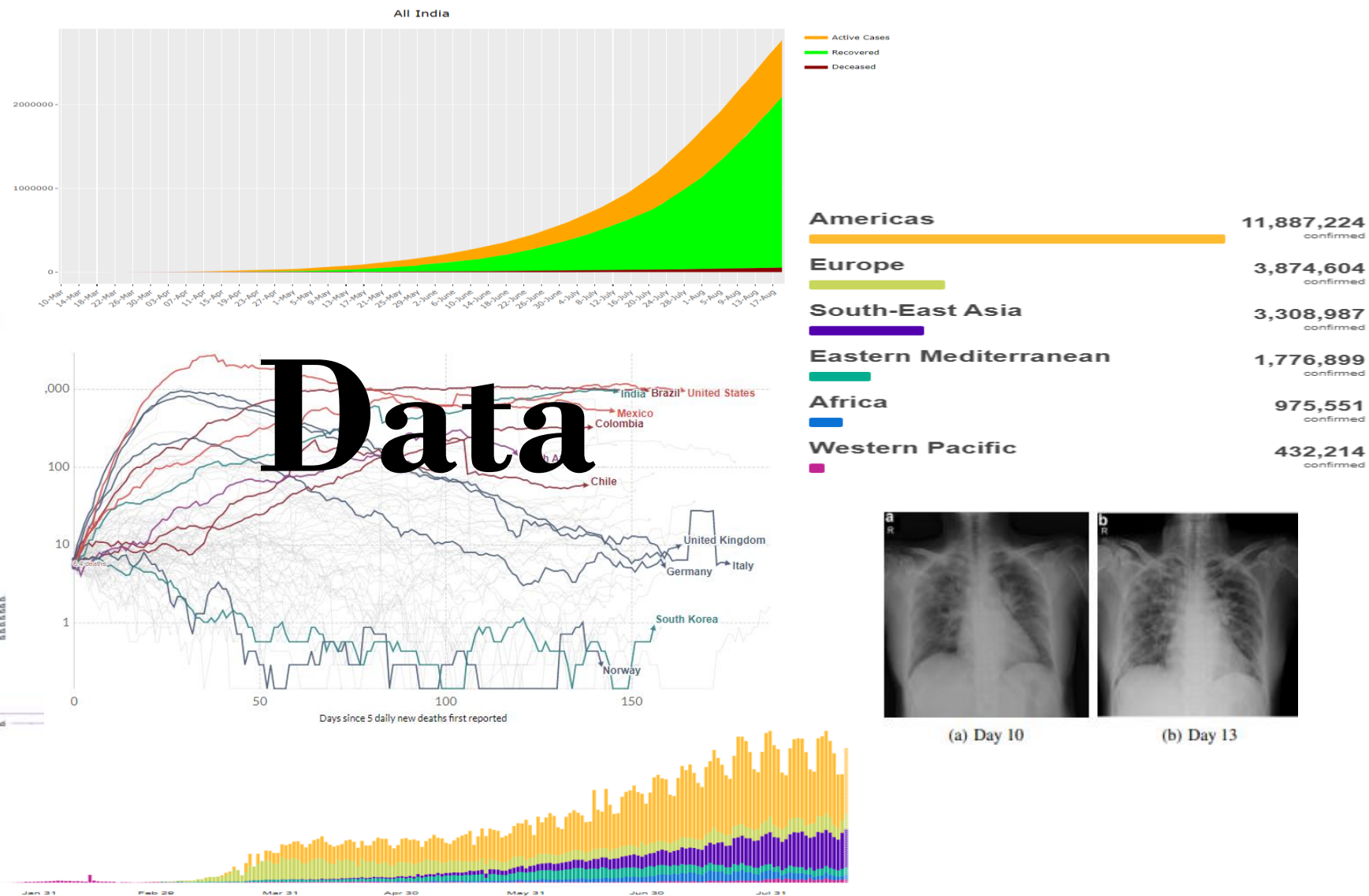
P59595 · NCAP_SARS

Name	Nucleoprotein
Organism	Severe acute respiratory syndrome coronavirus
Gene	Name N  1 Automatic Annotation ORF names 9a
Evidence	1: Evidence at protein level
Annotation score	5/5

Ontology Summit 2022, March 9, 2022
(<https://bit.ly/36Qbu5o>)



Background



Background

BIG Data

Situation by Country, Territory & Area

Name	Cases - cumulative total	Cases - newly reported in last 24 hours	Deaths - cumulative total
Global	23,518,343	206,382	810,492
United States o...	5,649,928	37,765	175,813
Brazil	3,605,783	23,421	114,744
India	3,167,323	60,975	58,390
Russian Feder...	966,189	4,696	17,568

In-depth: Excess mortality

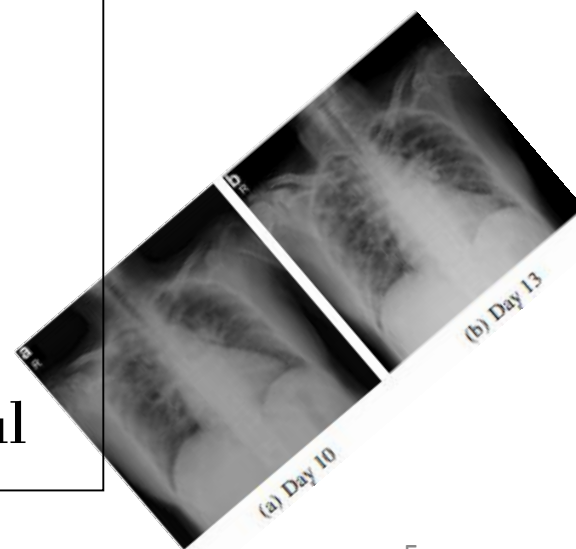
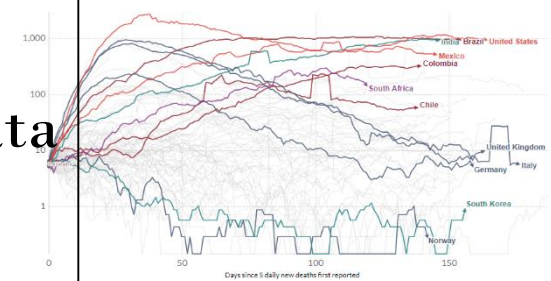
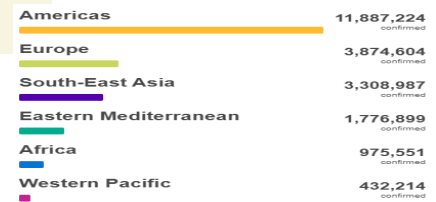
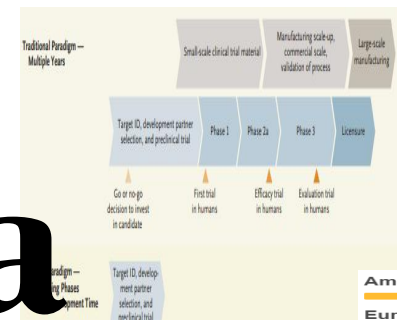
Excess mortality refers to the number of deaths *from all causes* above and beyond what we would have expected to see under 'normal' conditions.⁴ In this case, we're interested in how deaths during the COVID-19 pandemic compare to the average number of deaths over the same period in previous years.

Looking at excess mortality is helpful for understanding the total impact of the pandemic on deaths – both direct and indirect. It helps us understand the direct impact by capturing deaths caused by COVID-19 that were not correctly diagnosed and reported, for example because no test for the virus was conducted. It helps us understand the indirect

P59595 · NCAP_SARS

Name	Nucleoprotein
Organism	Severe acute respiratory syndrome coronavirus 2
Gene	Name N ORF names 9a
Evidence	1: Evidence at protein level
Annotation score	5/5

A	B	C	D	E	F	G	H
Case	Date	Age	Sex	City	State	Cluster	Reason
1	2020-03-09T00:00:00	41	Male	Bangalore	Karnataka	From USA	Texas US
2	2020-03-10T00:00:00	0	Female	Bangalore	Karnataka	From USA	Spouse
3	2020-03-10T00:00:00	13	Female	Bangalore	Karnataka	From USA	Daughter
4		0		Bangalore	Karnataka	From Unit	London
5	2020-03-13T00:00:00	26	Male	Bangalore	Karnataka	From the i	Greece
6	2020-03-12T00:00:00	26	Male	Kalburgi	Karnataka	From Saudi Ara	
7		0		Kalburgi	Karnataka	From Saudi Ara	
8	2020-03-17T00:00:00	32	Male	Bangalore	Karnataka	From USA	Passer
9	2020-03-17T00:00:00	63	Male	Kalburgi	Karnataka	From USA	Passer
10	2020-03-17T00:00:00	20	Female	Bangalore	Karnataka	From USA	Passer

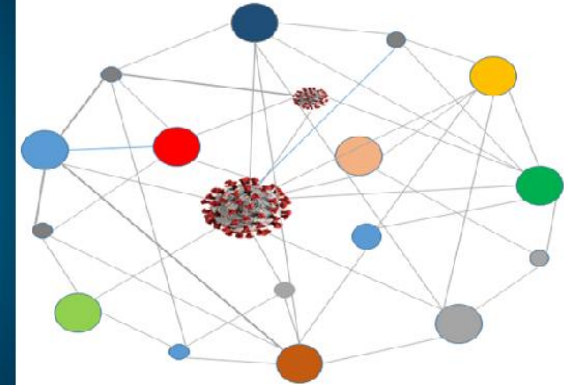


1. **Volume** (the amount of data)
2. **Velocity** (speed at which the data is produced)
3. **Variety** (heterogeneous sources and the nature of data (structured, semi-structured, unstructured))
4. **Veracity** (inconsistencies and uncertainty in data)
-> quality and accuracy of data is a concern
5. **Value** (assessment of data utility)
-> need to turn data into value, something useful

What are the problems?

**large, complex, multi-dimensional,
diverse, disintegrated, and non-
interpretable by machine, data insight
to be drawn manually**

**Result:
obstruct computer assisted reasoning,
a core to AI**



“find all people p who are related to someone r who has been diagnosed with COVID-19 and has not yet been tested.”

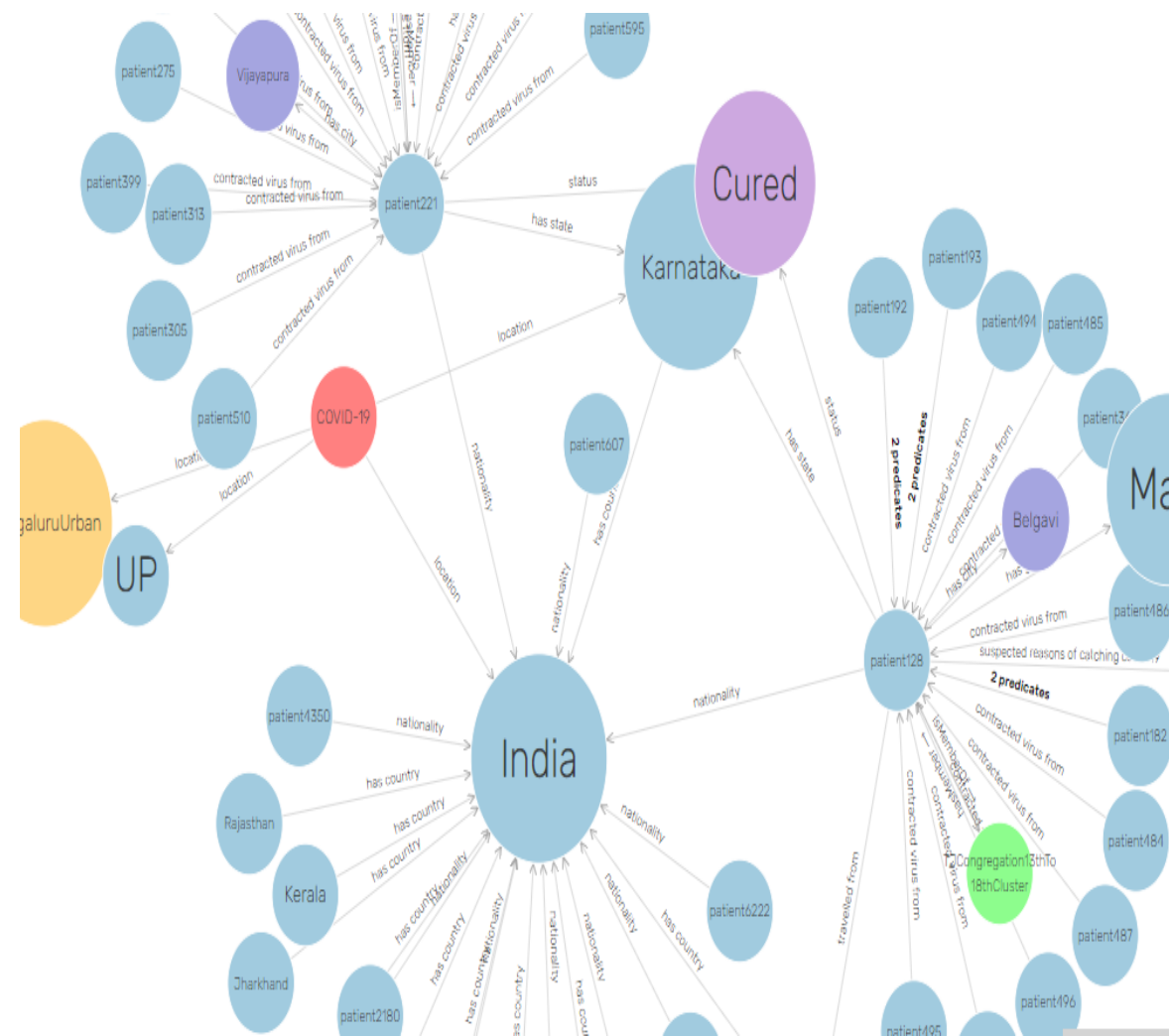
Turning data into knowledge, and essentially
the value is not a trivial problem either for
humans or machines

We advocate for

Machine empowerment

User empowerment

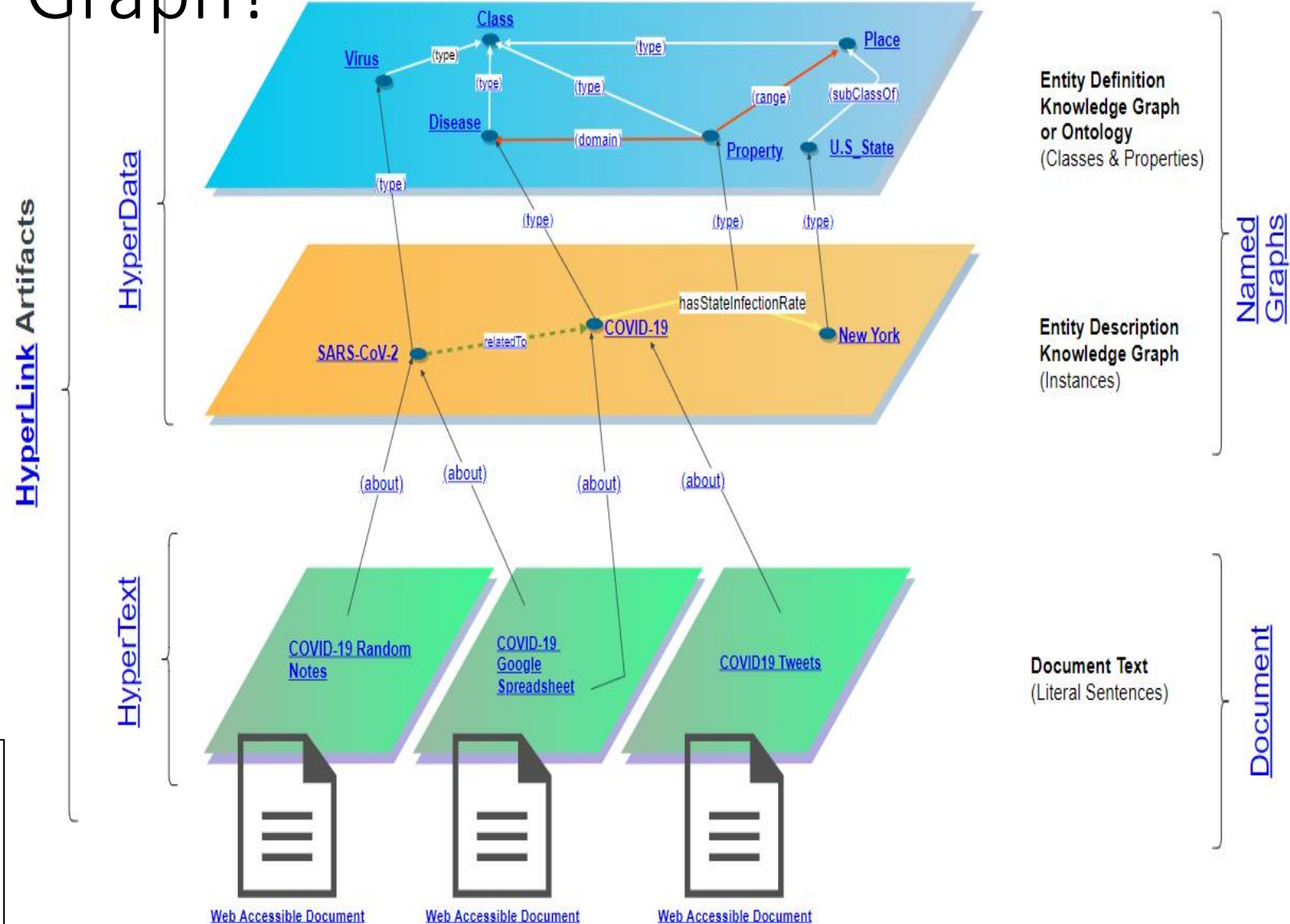
Knowledge Graph



What is Knowledge Graph?

A manifestation of an intelligent **Web of Data** informed by an **ontology***

Ontology is a structured vocabulary consists of human and computer interpretable terms and relations that represent entities and relationships using a formal KR language.



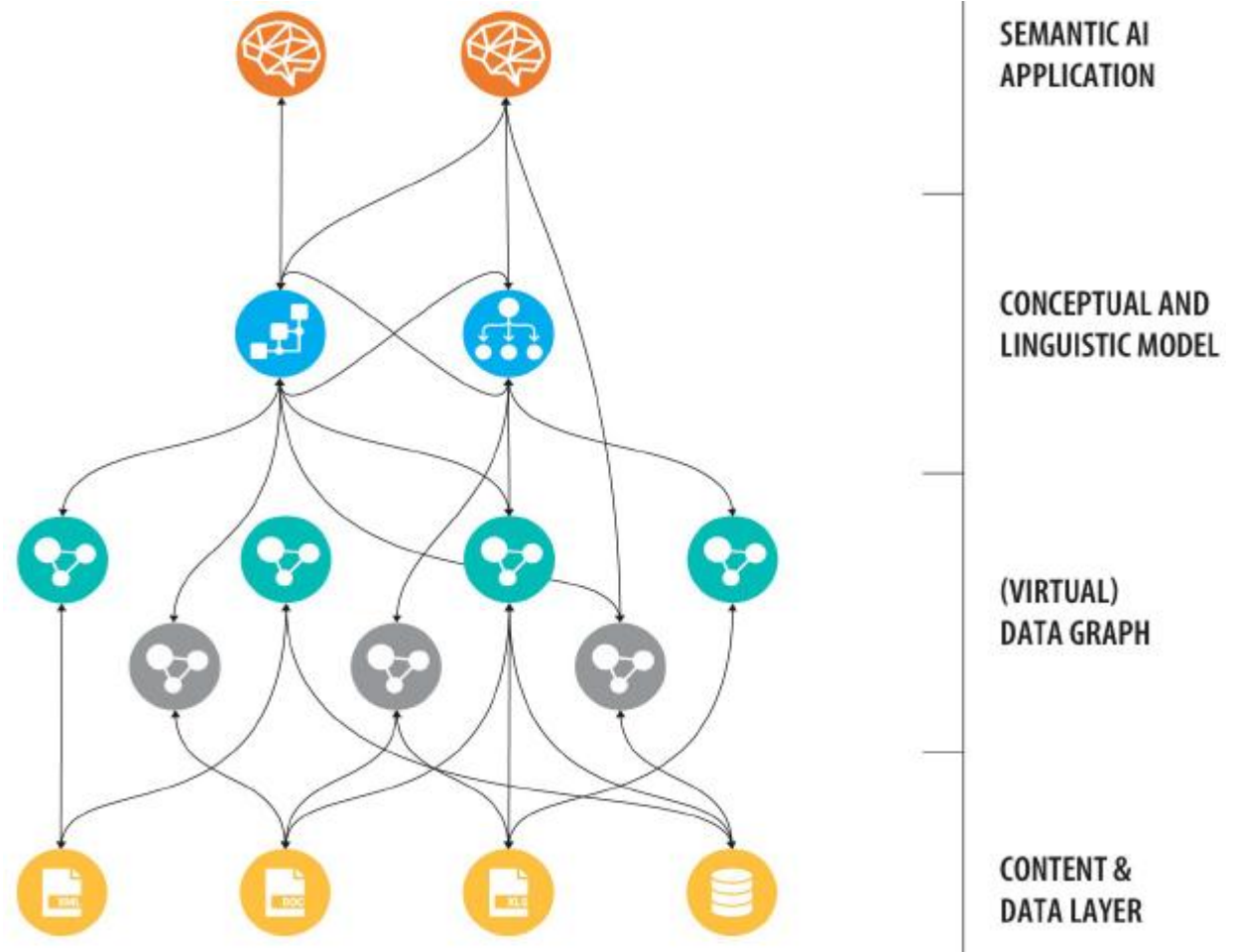
What is Knowledge Graph?

KG can be seen as a

Database: that can be queried

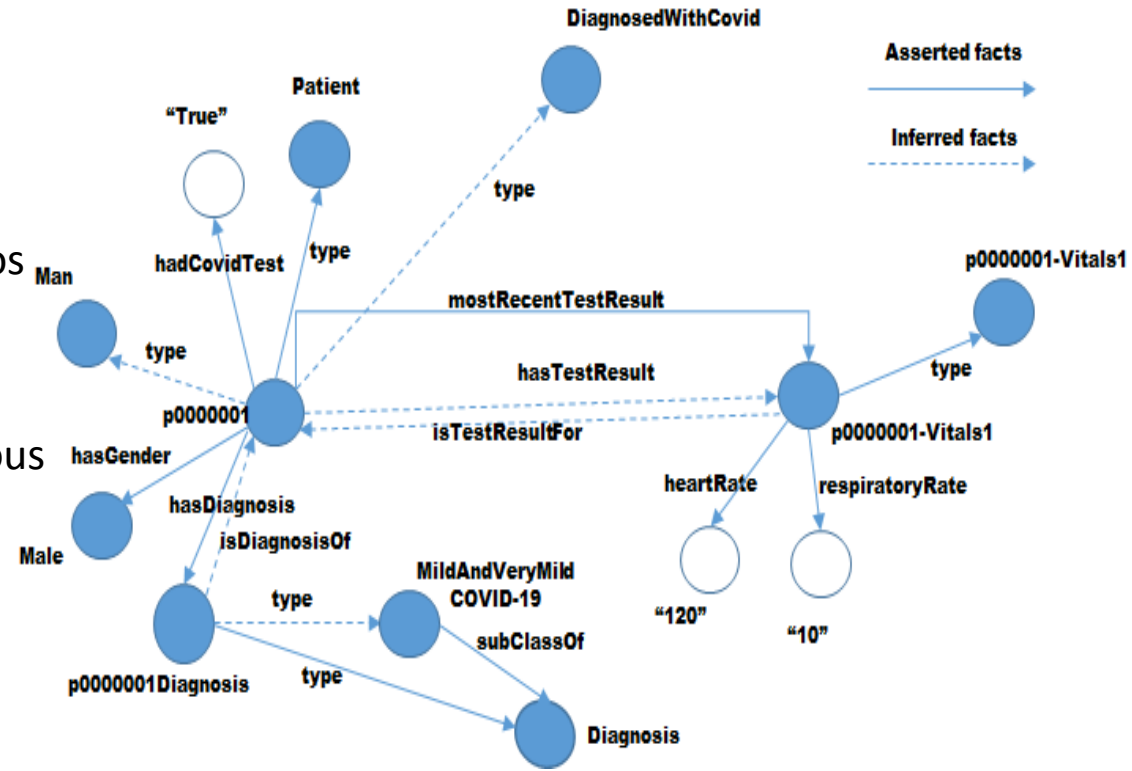
Graph: that can be analyzed as
a network of data

Knowledge base: new facts can
be inferred



Why knowledge graph?

- Turns "strings" to "things"
- Enables the explicit expression of human knowledge
- Helps in connecting various pieces of data scattered across the data silos
- Facilitates the easy fusion and development of context
- Enables the creation of comprehensive knowledge base by linking various resources
- A tool to extract insight from data by interlinking and analyzing
- A tool to visualize various aspects of a Thing
- A tool to form a backbone for AI and analytics platforms



KG usage: a quick review

- Web search
- Question answering system
- Data integration
- Data collection and analysis
- Data visualization
- Machine learning and advanced analytics

CODO is an open source project

(<https://www.isibang.ac.in/~bisu/covid19project.html>)

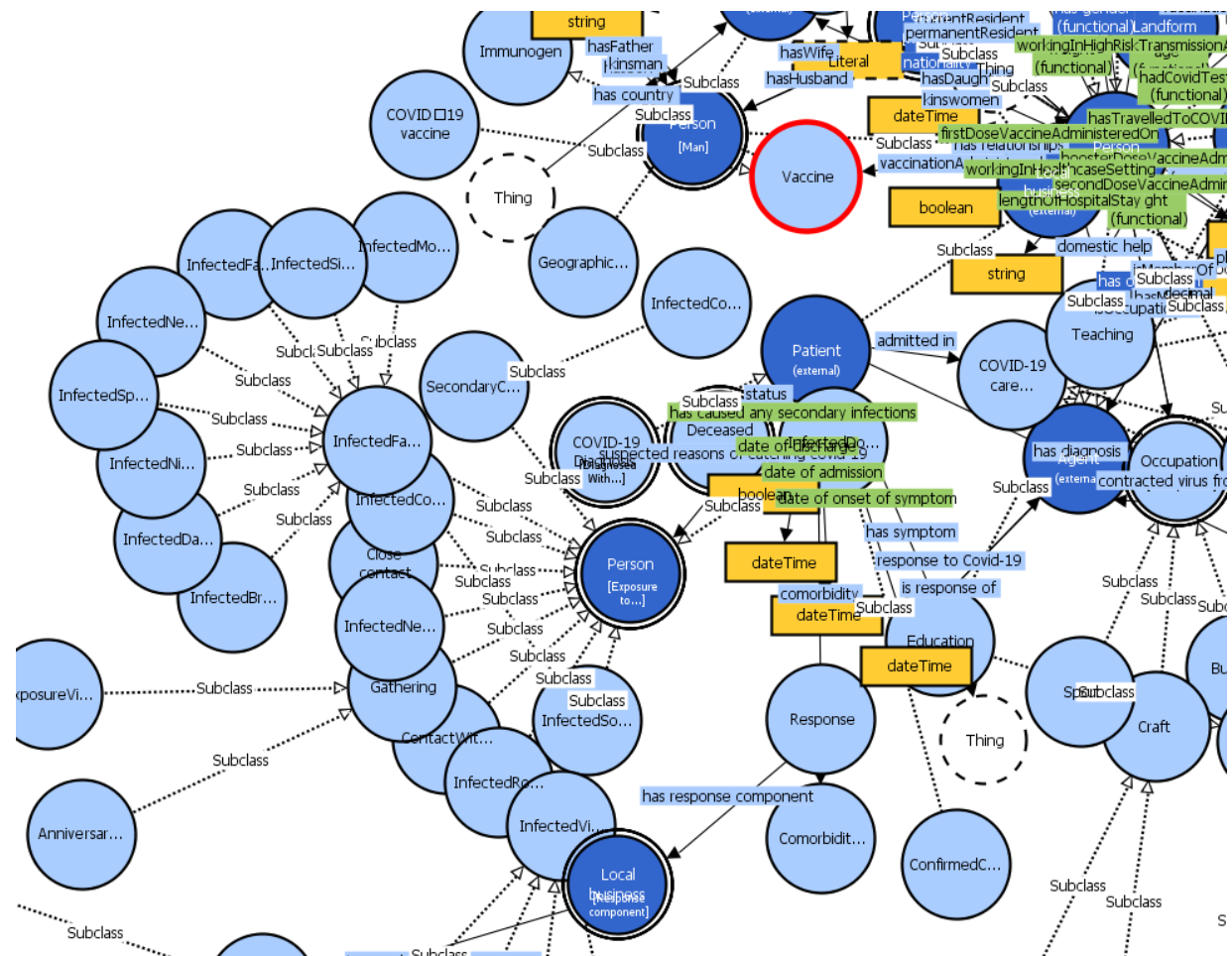
CODO: an Ontology for collection and analysis of multi-paradigm COviD-19 data

CODO Goals

- To serve as an explicit ontology for use by data and service providers to publish COVID-19 data as a FAIR data
- To provide a standards-based reusable vocabulary for the use of various organizations (e.g., government agencies, hospitals) to annotate and describe COVID-19 information
- To facilitate the development and offer of distributed, heterogeneous, semantic services and applications.
 - **E.g.**, Policy and decision support system (Should we continue the lockdown?)
 - Analytics (What are the drugs found to be the most effective in treating COVID-19? Which was the most prevalent comorbidity among the COVID-19 deceased patients?)

CODO Use cases

- Knowledge Graph creation
- Annotation of COVID-19 literature
- Application design
 - Assistive system to medical practitioners, e.g., COVID-19 risk detection system
 - Trace and combat the disease



CODO coverage

- Contact tracing
- Diagnosis
- Disease and Comorbidity
- Symptoms
- Treatment
- Drug
- Vaccination
- Cases, resources, tests
- Preventive measure and responses
- Weather situation
- ...



CODO design methodology

- Followed a hybrid method (utilizing YAMO[11] and Agile methods)
 - Bottom-up and Top-down approach
 - Story driven development
 - Test driven development
 - Rapid interactions
 - Refactoring

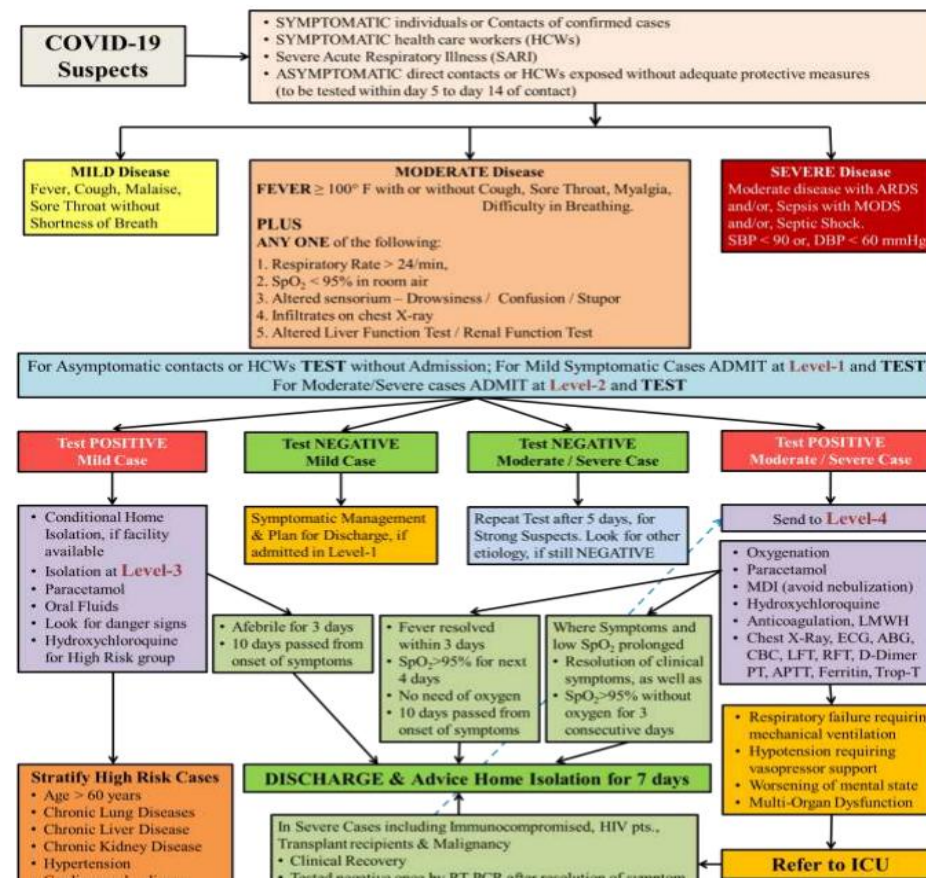
Competency questions

- i. Find all People p who are related to someone r who has been diagnosed with COVID-19 and who has not yet been tested.
- ii. Give me the primary reasons i for the maximum number of COVID-19 patients p .
- iii. Give me the most prevalent symptoms s of Severe COVID-19 d .

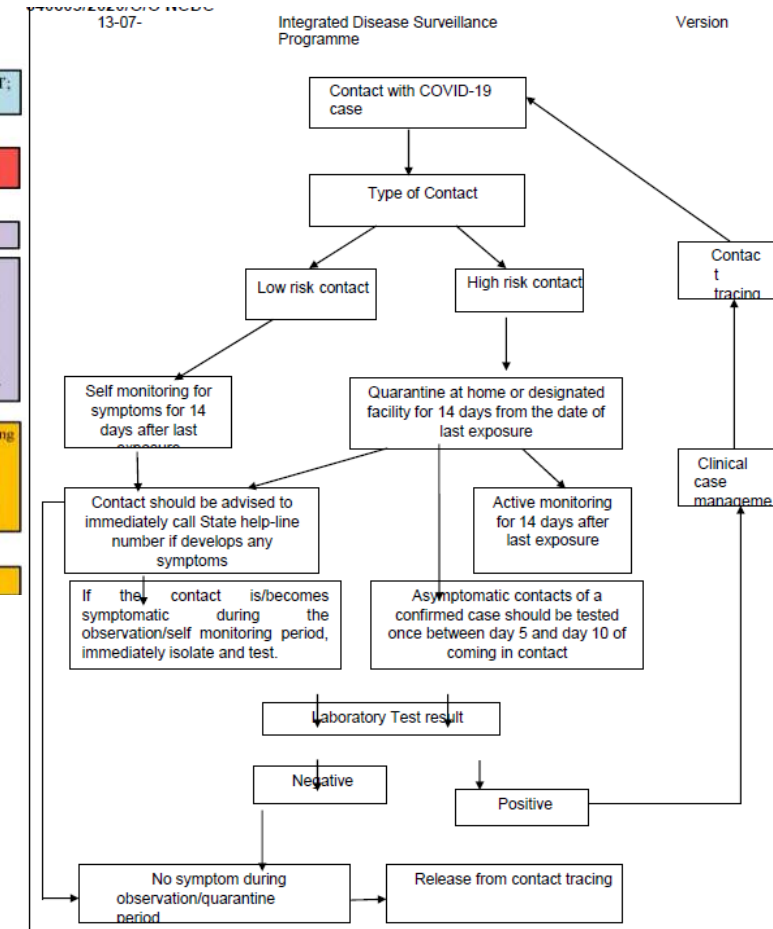
For methodology details, see [references 12 & 13]

Materials consulted

- COVID-19 clinical management protocol, guidelines, etc.
- Government bulletins, circulars, etc.
- Websites:
<https://www.who.int/>,
<https://www.cdc.gov/>, etc.
- Literature
- Vocabularies, like SNOMED CT, schema.org, etc.
- ...
- Datasets



Case	Date	Age	Sex	City	State	Cluster	Reason
15	19-Mar	35	Male	Madikeri in K	Karnataka	From Middle East	Dubai
16	21-Mar	53	Female	Bangalore-Ur	Karnataka	From United Kingdom	Domestic help
17	21-Mar	39	Male	Bangalore-Ur	Karnataka	From the rest of Europe	Amsterdam in Ne
18	21-Mar	21	Male	Bangalore-Ur	Karnataka	From the rest of Europe	Edinburgh in Scot
19	21-Mar	31	Male	Chikballapur	Karnataka	From Middle East	Mecca in Saudi Ar
20	21-Mar	35	Male	Mysore	Karnataka	From Middle East	Dubai
21	22-Mar	35	Male	Dharwad	Karnataka	From Middle East	Dubai
22	22-Mar	64	Female	Gowribidanu	Karnataka	From Middle East	Mother
23	22-Mar	36	Female	Bangalore-Ur	Karnataka	From the rest of Europe	Switzerland and f
24	22-Mar	27	Male	Bangalore-Ur	Karnataka	From the rest of Europe	Germany
25	22-Mar	51	Male	Bangalore-Ur	Karnataka	From United Kingdom	London
26	23-Mar	22	Male	Mangalore-Ur	Karnataka	From Middle East	Dubai
27	23-Mar	46	Male	Bangalore-Ur	Karnataka	From Middle East	Dubai
28	23-Mar	38	Male	Bangalore-Ur	Karnataka	From Middle East	Dubai
29	23-Mar	41	Male	Bangalore-Ur	Karnataka	From United Kingdom	London
30	23-Mar	30	Female	Bangalore-Ur	Karnataka	From the rest of Europe	Spouse



Release 2021-11-01T10:00:00

CODO v1.3 consists of
of classes: 83
of object property: 73
of data property: 48

This version:

<http://w3id.org/codo/1.3>

Latest version:

<http://w3id.org/codo/1.3>

Previous version:

<http://www.isibang.ac.in/ns/codo/1.0>

Authors:

Biswanath Dutta (Indian Statistical Institute, India)
Michael DeBellis (Semantic Web Consultant, USA)

Publisher:

Indian Statistical Institute

Download serialization:

Format [JSON LD](#)

Format [RDF/XML](#)

Format [N Triples](#)

Format [TTL](#)

License:

License <https://creativecommons.org/licenses/by/4.0/>

Visualization:

Visualize with [WebVowl](#)

Cite as:

Dutta, B. and DeBellis, M. (2020). CODO: an ontology for collection and analysis of COVID-19 data. In Proc. of 12th Int. Conf. on Knowledge Engineering and Ontology Development (KEOD), Lisboa, Portugal, 2-4 November 2020, vol. 2, pp. 76-85 (ISBN 978-989-758-474-9, ISSN 2184-3228) (DOI: <https://doi.org/10.5220/0010112500760085>).

Browse

<https://w3id.org/codo>

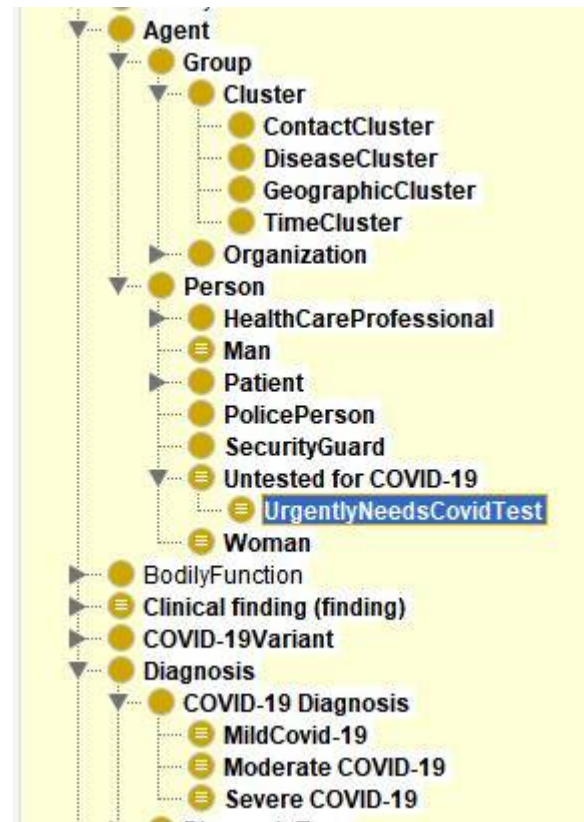
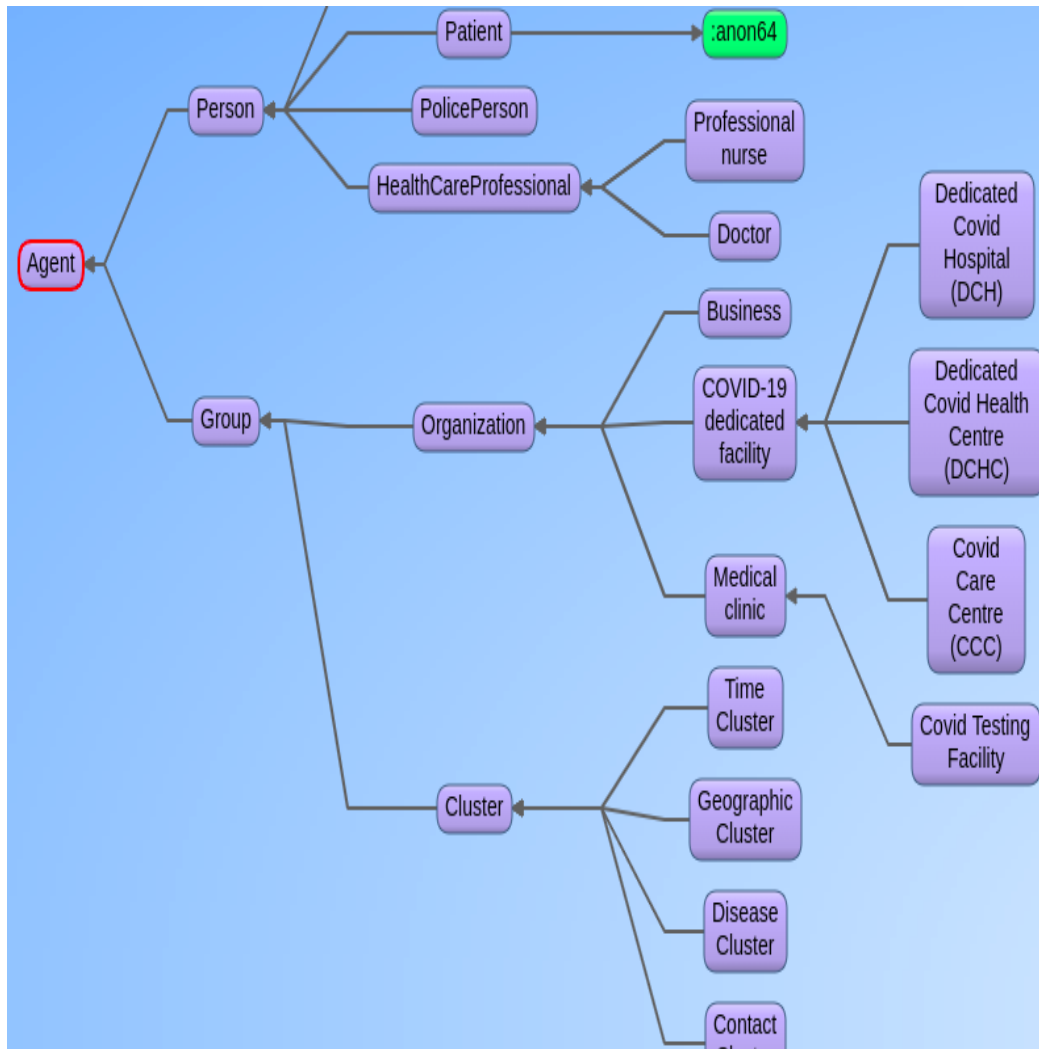
Also, available from

<https://github.com/biswanathdutta/CODO>

Abstract

Ontology Summit 2022, March 9, 2022
(<https://bit.ly/36Qbu5o>)

CODO v1.3 [some glimpses (1 of 3)]



Equivalent To +

● **Person**
and (hasCloseRelationship some 'Diagnosed With COVID-19')
and (hadCovidTest value false)

SubClass Of +

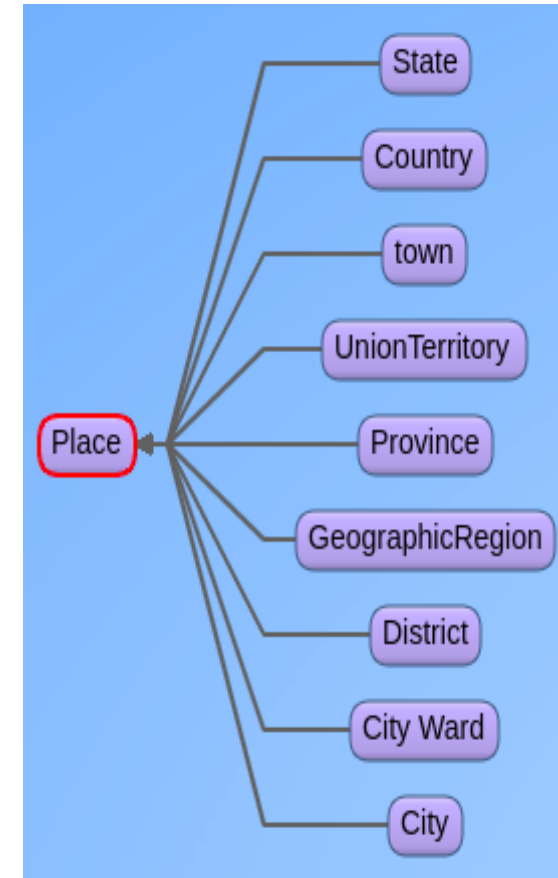
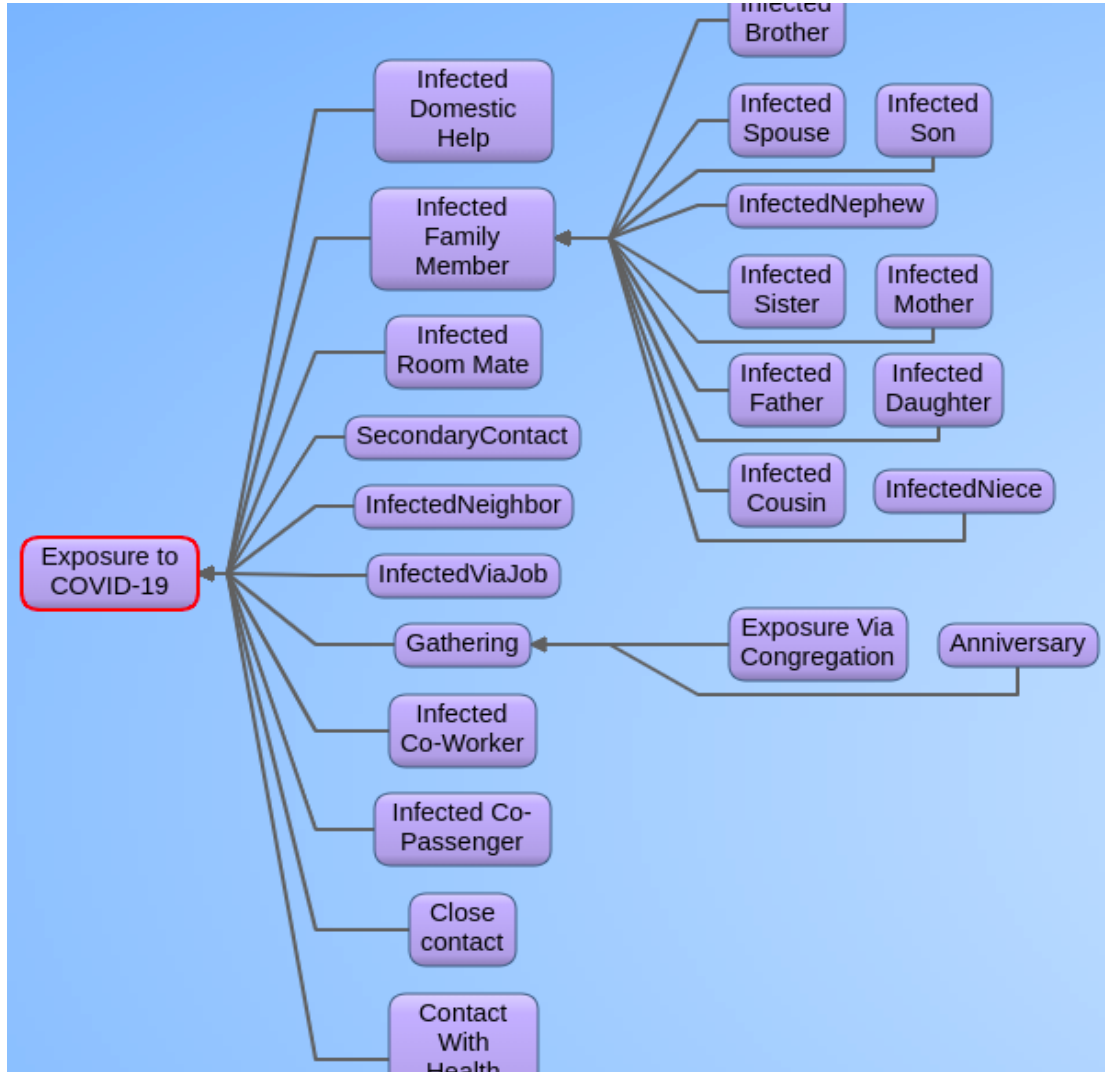
● 'Untested for COVID-19'

General class axioms +

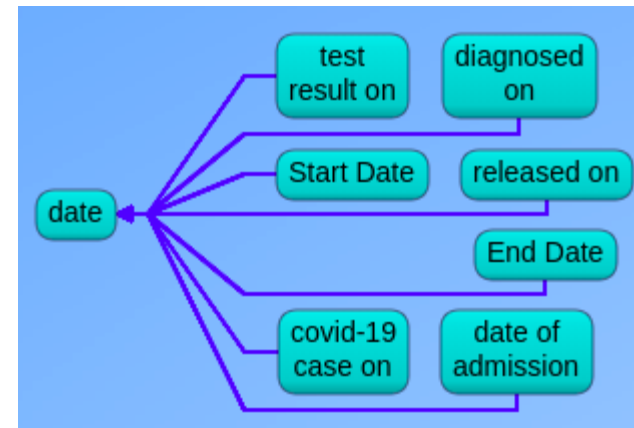
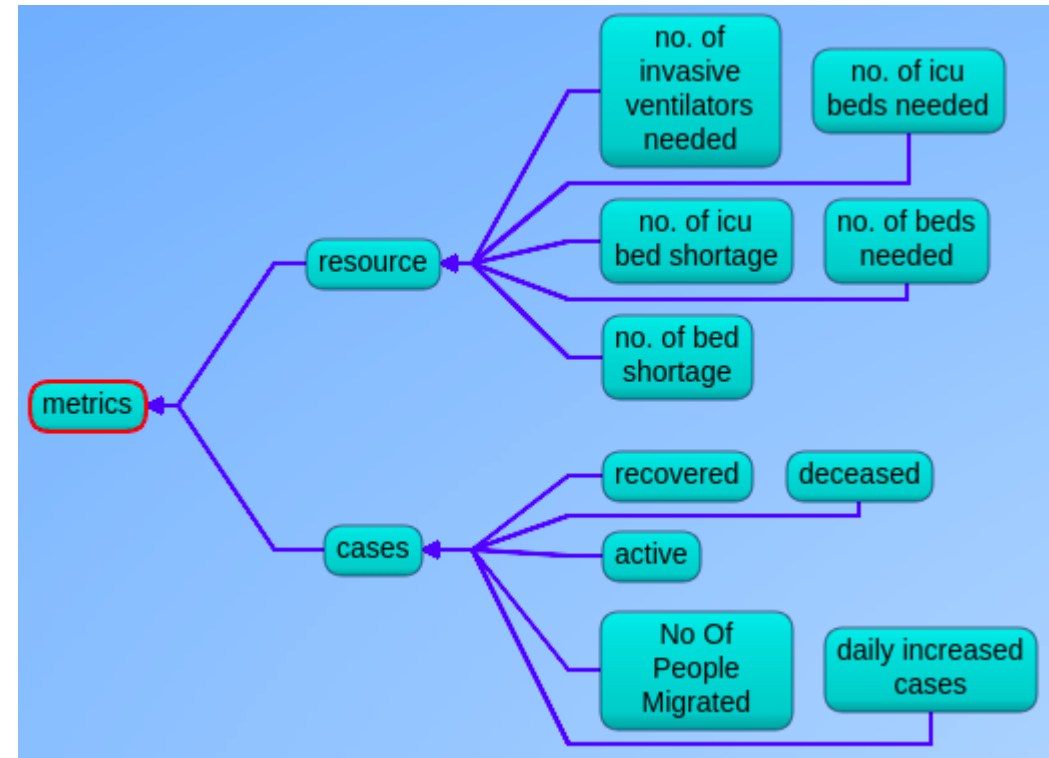
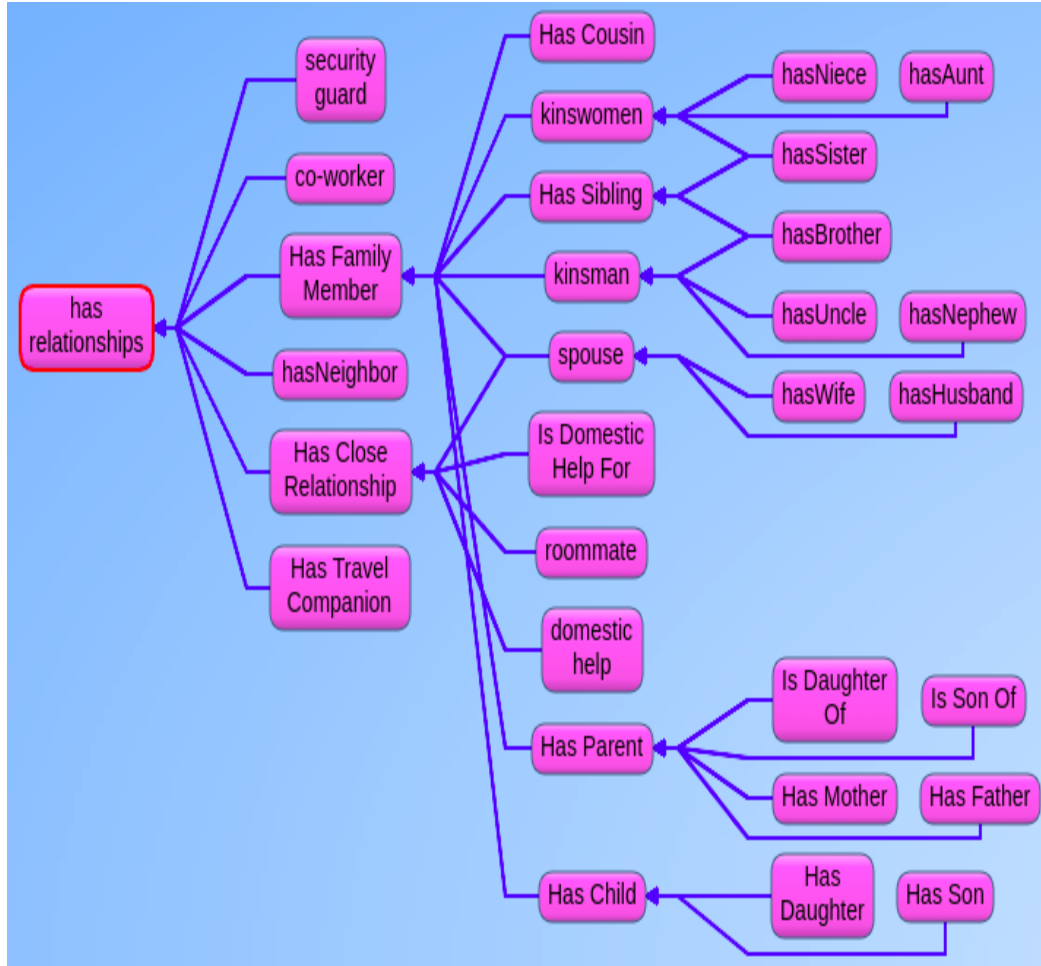
SubClass Of (Anonymous Ancestor)

● 'diagnosed with variant' some 'Coronavirus variant'
● **Person**
and (hadCovidTest value false)

CODO v1.3 [some glimpses (2 of 3)]



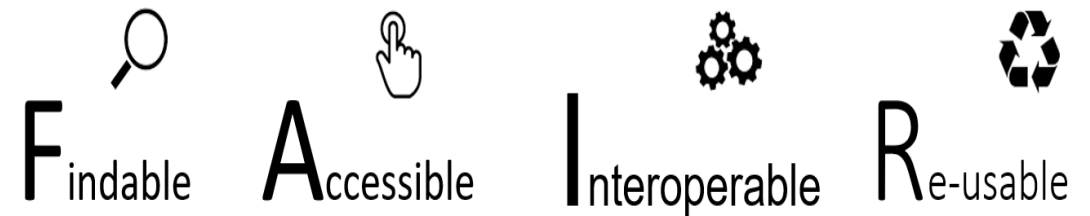
CODO v1.3 [some glimpses (3 of 3)]



CODO Knowledge Graph

Primarily with the following two goals:

1. Transforming COVID-19 data as **FAIR Semantic** data
2. CODO ontology evaluation



CODO KG Development

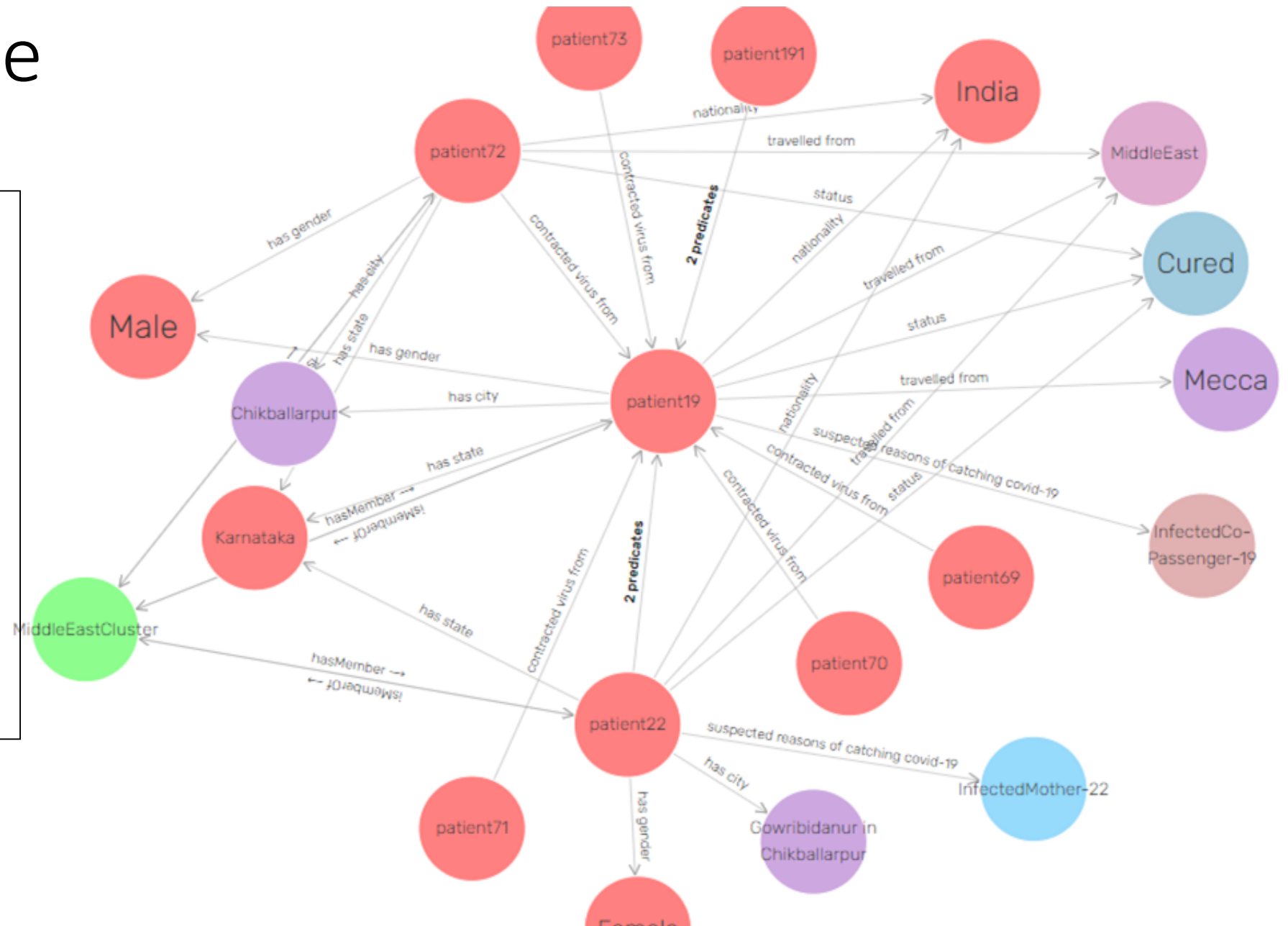
- Dataset collection
 - Karnataka COVID-10 data (*Obtained from https://covid19.karnataka.gov.in/govt_bulletin/en*)
- Data processing
- Data transformation
- KG production

```
10 INSERT (  
11   ?cl codo:hasMember ?p.  
12   ?p codo:isMemberOf ?cl.  
13   ?p codo:travelledFrom ?place.  
14 )  
15 WHERE (  
16   ?p codo:clusterString ?cs.  
17   BIND  
18     (IF (CONTAINS(?cs, 'From Middle East'), codo:MiddleEastCluster,  
19      (IF (CONTAINS(?cs, 'From USA'), codo:USACluster,  
20       (IF (CONTAINS(?cs, 'From South America'), codo:SouthAmericaCluster,  
21        (IF (CONTAINS(?cs, 'From the rest of Europe'), codo:RestOfEuropeCluster,  
22         (IF (CONTAINS(?cs, 'From the Southern States'), codo:SouthernStatesCluster,  
23          (IF (CONTAINS(?cs, 'TJ Congregation from 13th to 18th March in Delhi'), codo:TJCongregation13thTo18thCluster,  
24           (IF (CONTAINS(?cs, 'From Gujarat'), codo:GujaratCluster,  
25            (IF (CONTAINS(?cs, 'From Maharashtra'), codo:MaharashtraCluster,  
26             (IF (CONTAINS(?cs, 'From Rajasthan'), codo:RajasthanCluster,  
27              (IF (CONTAINS(?cs, 'From Rajasthan'), codo:RajasthanCluster,  
28               (IF (CONTAINS(?cs, 'Domestic Travel History Absent'), codo:DomesticTravelHistoryAbsentCluster,  
29                (IF (CONTAINS(?cs, 'International Travel History Absent'), codo:InternationalTravelHistoryAbsentCluster,  
30                 (IF (CONTAINS(?cs, 'From United Kingdom'), codo:UnitedKingdomCluster, owl:Nothing)))))))))))))) AS ?cl)  
31   ?cl codo:hasLocation ?place.  
32   FILTER(?cl != owl:Nothing)  
  
let ((results  
      (sparql:run-sparql "SELECT ?p ?rs WHERE {?p codo:reasonString ?rs.  
FILTER(Regex(?rs, 'Contact of P\\\\\\\\d+-\\\\\\\\d+'))" :results-format :lists)))  
loop for result in results do  
  (let* ((patient-upi (first result))  
         (patient-value (part->terse patient-upi))  
         (rs-upi (second result))  
         (rs (upi->value rs-upi)))  
    (multiple-value-bind (matched-p ids) (match-re "\\d+-\\d+" rs)  
      (let* ((id-string-list (split-re "-" ids))  
             (first-id-num (parse-integer (first id-string-list)))  
             (last-id-num (parse-integer (second id-string-list))))  
        loop for id-index from first-id-num to last-id-num
```

CODO KG size

CODO KG consists of (as of March 8, 2022)

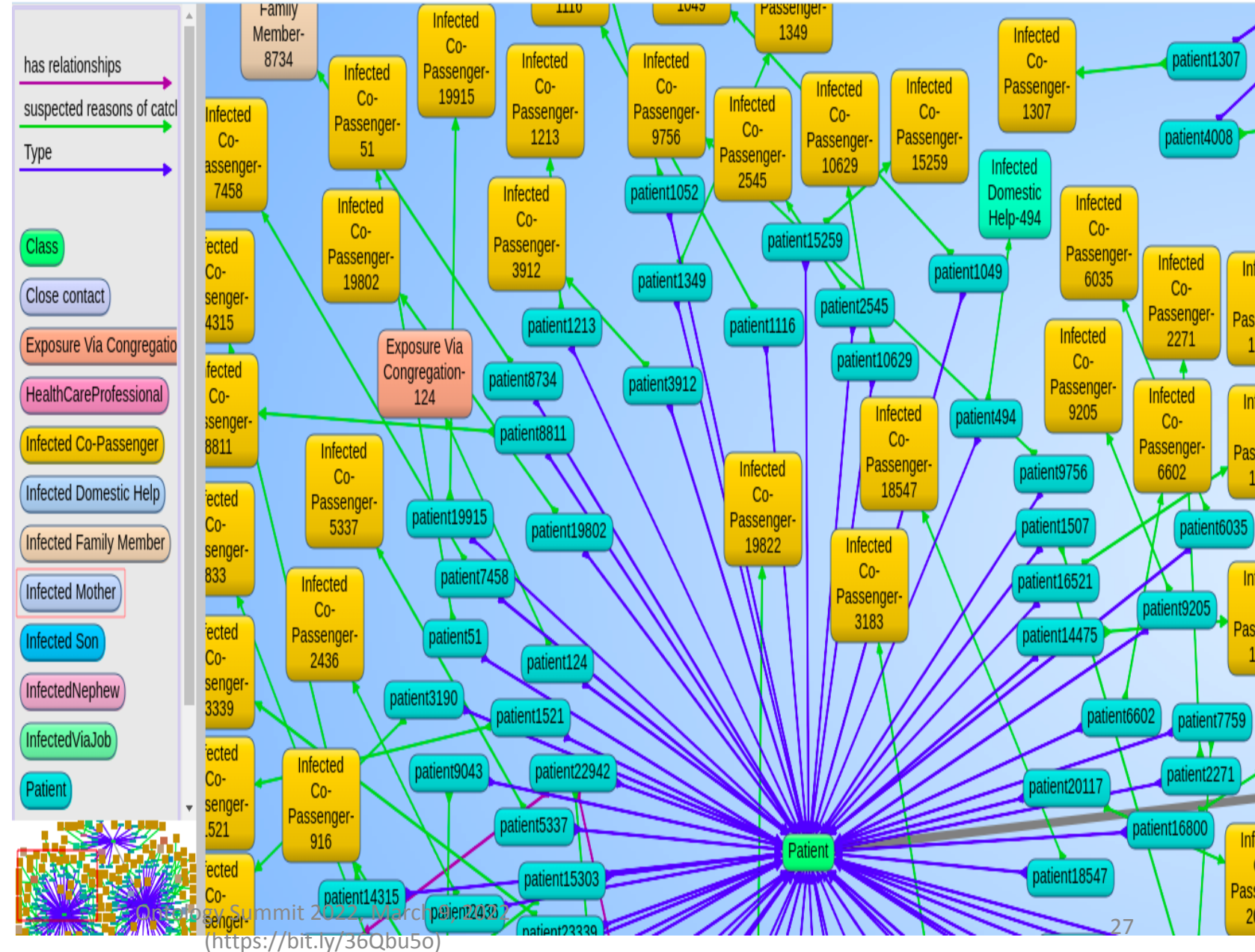
- # of axioms: 1151495
- # of individuals: 89511 (patients (75759), cases, vaccination count, weather, preventive measures, etc.)
- # of classes: 185
- # of object properties: 116
- # of data properties: 95



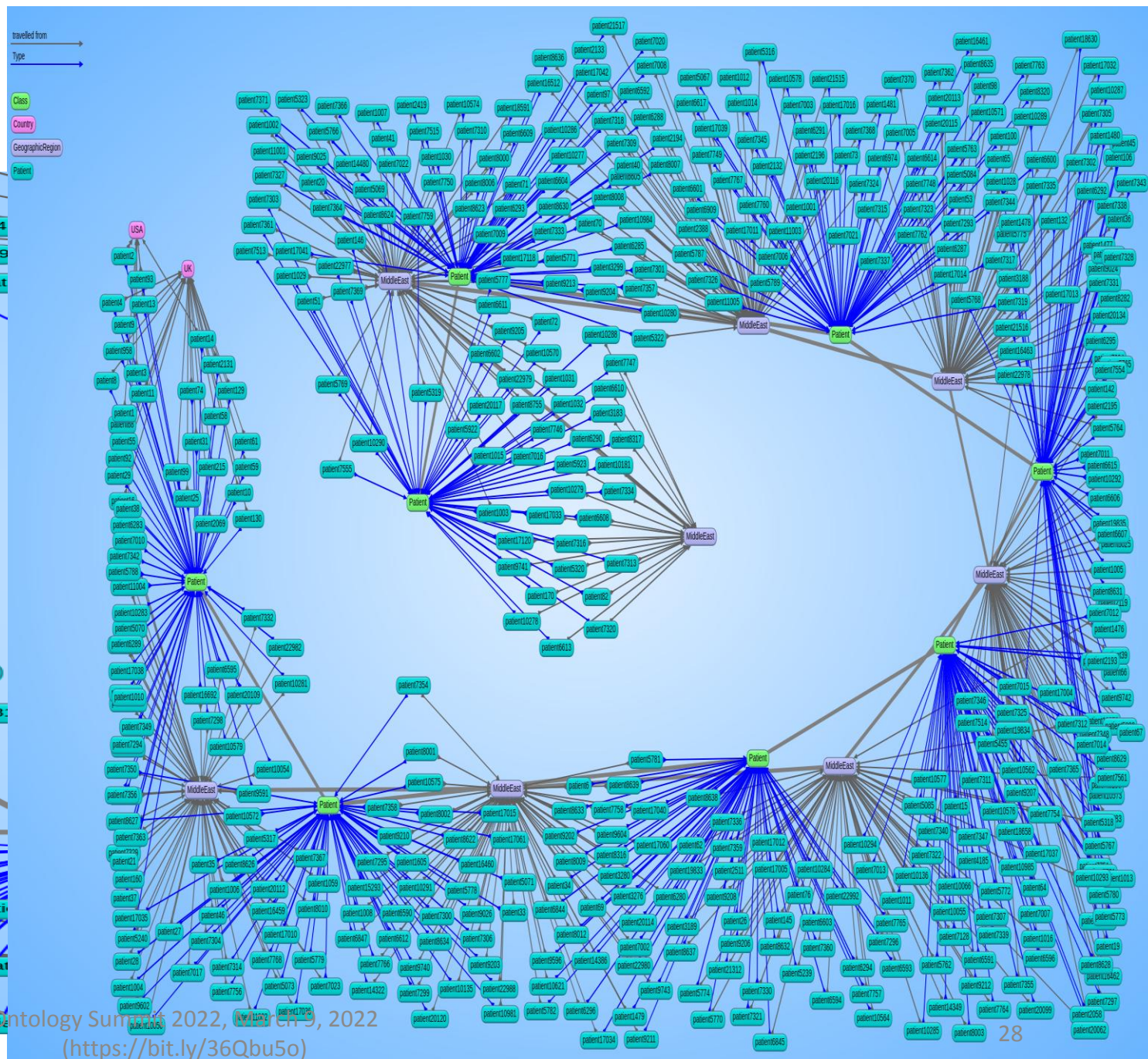
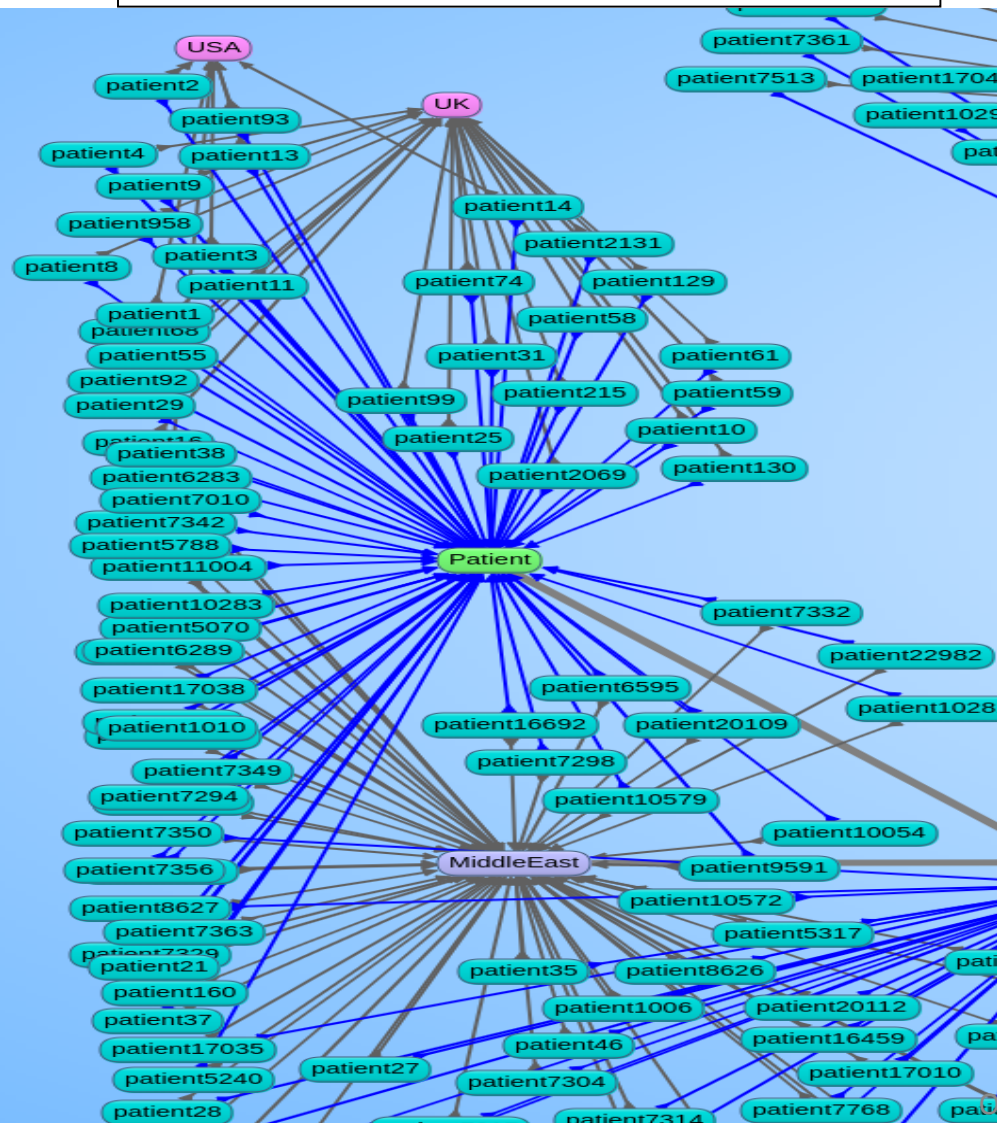
CODO KG Query

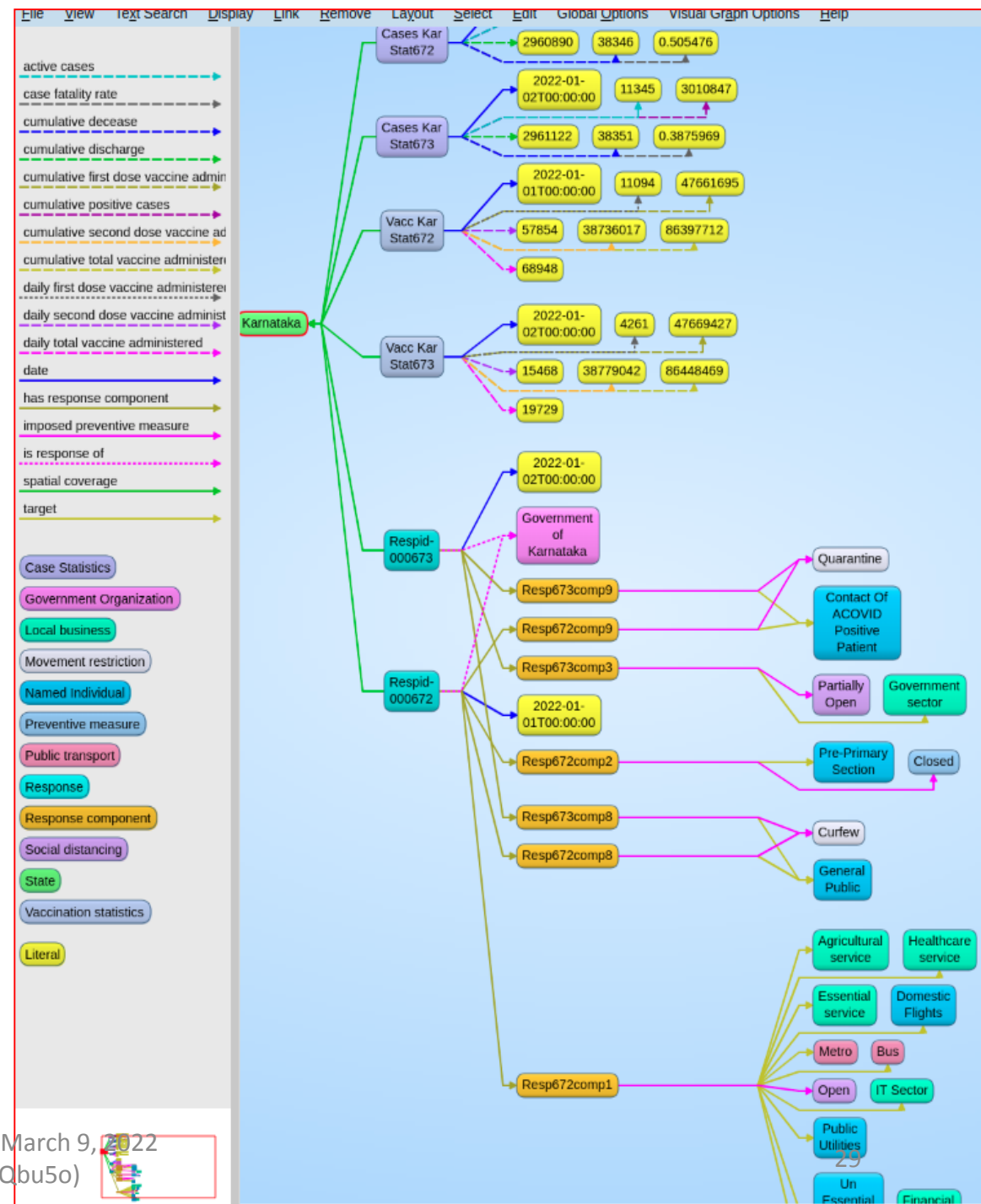
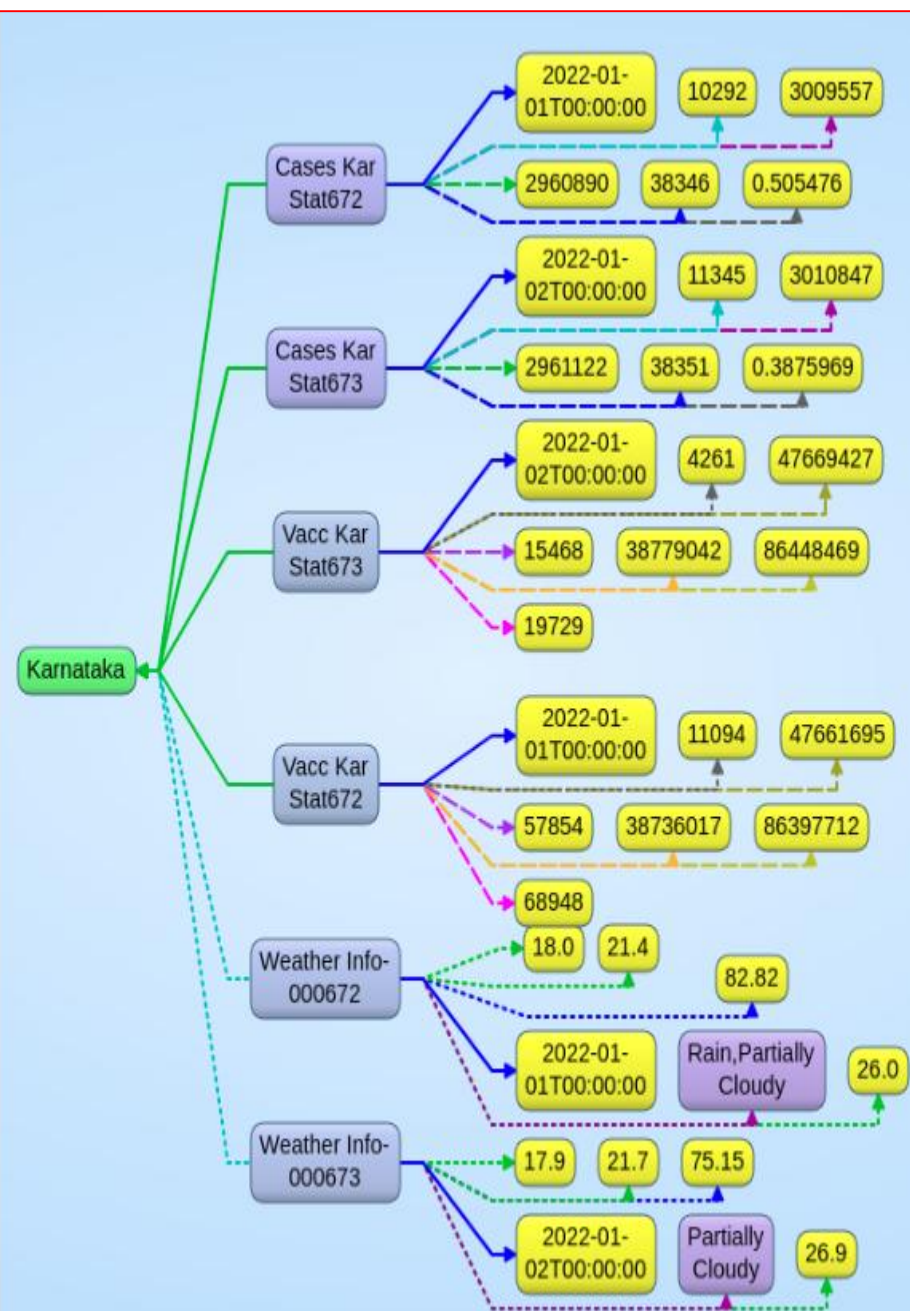
#Show the patients with the possible reasons of catching COVID-19. Also, display the relationships between the patients, if any.

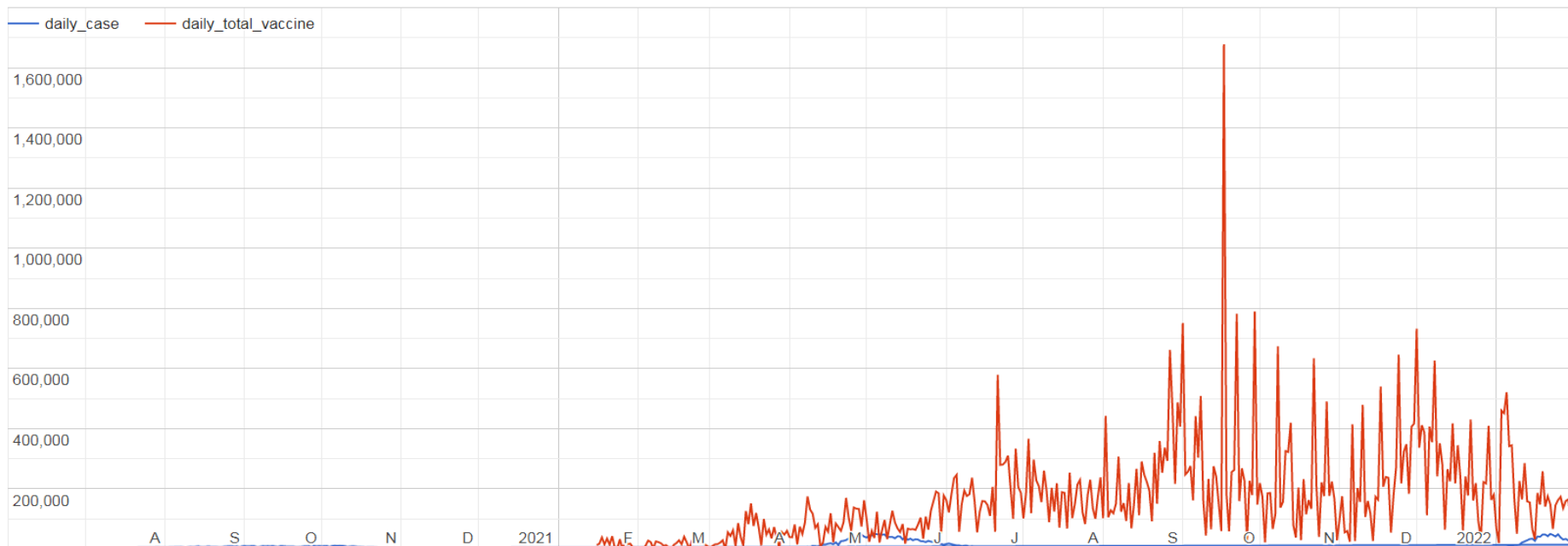
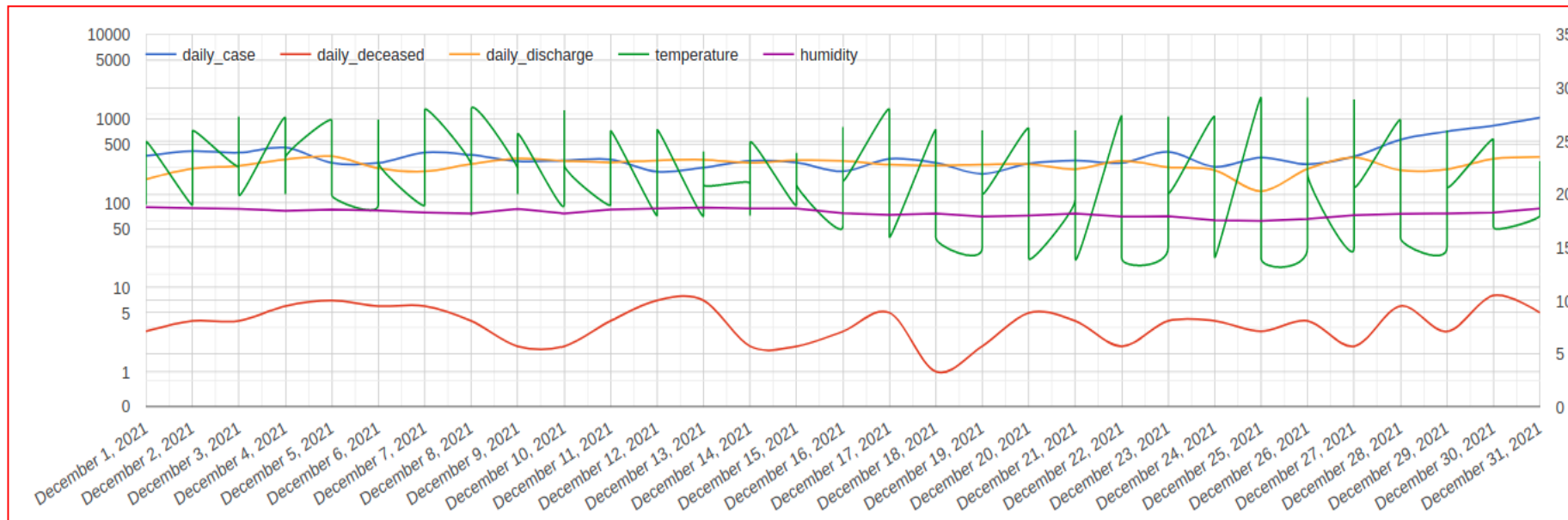
```
SELECT ?p ?r ?l
WHERE {
  ?p rdf:type schema:Patient.
  ?p
    codo:suspectedReasonOfCatchingCovid-19 ?r.
  OPTIONAL{?p codo:hasRelationship ?l.}
} LIMIT 150
```



```
SELECT ?p ?tf ?tt
WHERE {
    ?p rdf:type schema:Patient.
    ?p codo:travelledFrom ?tf.
    OPTIONAL{?p codo:travelledTo ?tt.}
} LIMIT 2000
```





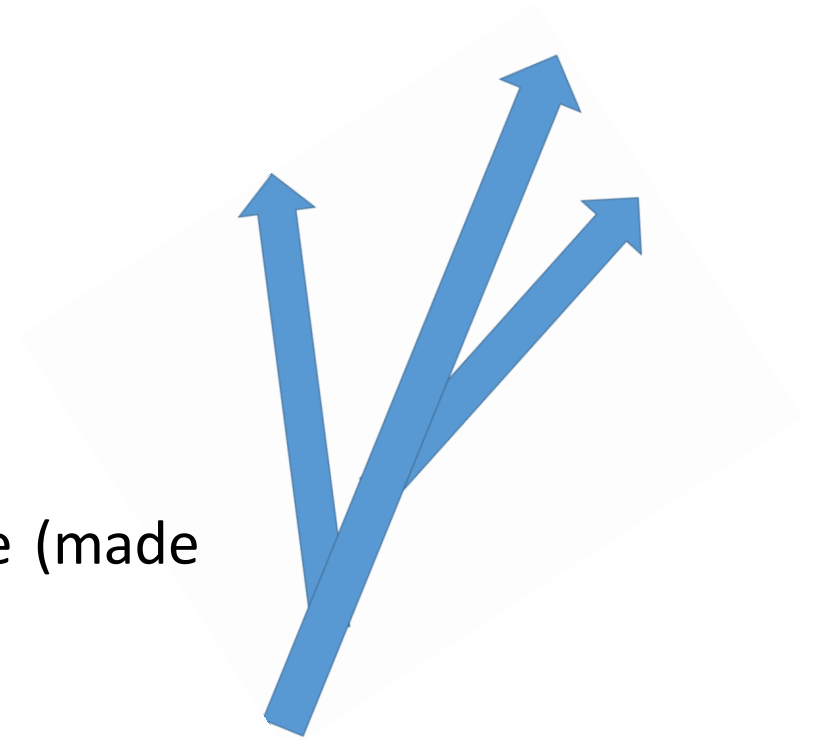


Experience/ challenges

- Data availability (e.g., clinical data)
- Most of the datasets are not directly consumable, not suitable to the graph
 - This made the data transformation complicated, time consuming
- Inconsistent data (e.g., sometimes P1342, sometimes 1342. Sometimes P132-P134 and sometimes "P132 and P133 and P134")
- Typo errors
 - Data misplacement
 - Spelling errors (mostly the place names and this complicates the linkage with the external resources)

Further research

- CODO 1.4 (Expected to publish soon)
- Mapping and reuse of other related schema
- SPARQL endpoint for schema and data service
- Improvement of our current data integration procedure (made some progress)
- Applications
- API development



Acknowledgement

- Project funded by Indian Statistical Institute (<https://www.isical.ac.in/>)
- Government of Karnataka India (<https://covid19.karnataka.gov.in/>)
- Franz Inc. (<http://www.allegrograph.com>)
- Protégé (<https://protege.stanford.edu/>)
- Cellfie plugin (<https://github.com/protegeproject/cellfie-plugin>)
- GraphDB (<https://www.ontotext.com/>)

References

1. <https://medium.com/virtuoso-blog/linked-data-ontologies-and-knowledge-graphs-a3d0ad6d6f66>
2. Singhal, A. (2012). Introducing the Knowledge Graph: things, not strings. <https://www.blog.google/products/search/introducing-knowledge-graph-things-not/>
3. Idehen, Kingsley U. (2020). Linked Data, Ontologies, and Knowledge Graphs. <https://www.linkedin.com/pulse/linked-data-ontologies-knowledge-graphs-kingsley-uyi-idehen/>
4. Blumauer, A. and Kiryakov, A. (2020). Knowledge Graphs: 5 Use Cases and 10 Steps to Get There. (<https://www.ontotext.com/knowledgehub/webinars/knowledge-graphs-5-use-cases-and-10-steps-to-get-there/>)
5. W3C, 2014. RDF 1.1: Concepts and Abstract Syntax. W3C Recommendation. <https://www.w3.org/TR/rdf11-concepts/>
6. W3C, 2014a. RDF Schema 1.1. W3C Recommendation. <https://www.w3.org/TR/rdf-schema/>
7. W3C, 2012. Web Ontology Language Document Overview (Second Edition). W3C Recommendation. <https://www.w3.org/TR/owl2-overview/>
8. Gruber, T.R. (1993), “A translation approach to portable ontologies”, Knowledge Acquisition, Vol. 5 No. 2, pp. 199-220.
9. DuCharme, Bob, 2011. Learning SPARQL. O’Reilly.
10. W3C, 2004. SWRL: A Semantic Web Rule Language Combining OWL and RuleML. W3C Member Submission. <https://www.w3.org/Submission/SWRL/>
11. Dutta, B., Chatterjee, U. and Madalli, D. P. (2015). YAMO: Yet Another Methodology for Large-scale Faceted Ontology Construction. In Emerald Journal of Knowledge Management. Vol. 19, no. 1, pp. 6 – 24.
12. Dutta, B. and DeBellis, M. (2020). CODO: an ontology for collection and analysis of COVID-19 data. In Proc. of 12th Int. Conf. on Knowledge Engineering and Ontology Development (KEOD), Lisboa, Portugal, 2-4 November 2020, vol. 2, pp. 76-85.
13. DeBellis, M. and Dutta, B. (2021). Developing the Covid-19 CODO Knowledge Graph: An Agile Approach From Ontology to Knowledge Graph. In: Villazón-Terrazas B., Ortiz-Rodríguez F., Tiwari S., Goyal A., Jabbar M. (eds) Knowledge Graphs and Semantic Web. KGSWC 2021. Communications in Computer and Information Science (CCIS), Springer, Cham, vol. 1459, pp. 153-168.
14. Graph: Enterprise Knowledge Graph. 2020. (Available from https://www.youtube.com/watch?time_continue=5&v=MJuRnuA0hrM&feature=emb_logo)

References

15. Bender, Edward A.; Williamson, S. Gill (2010). Lists, Decisions and Graphs. With an Introduction to Probability.
16. Natasha Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, And Jamie Taylor (2019). Five diverse technology companies show how it's done. ACM.
17. Dieter Fensel, [...], Alexander Wahler (2020). Knowledge Graphs: methodology, tools and selected use cases. Springer.
18. Blumauer, A. and Nagy, H. (2020). The Knowledge Graph Cookbook: recipes that work. Edition Mono/Monochrom, Vienna, Austria.
19. Dutta, B. (2020). Knowledge Graph and the current pandemic COVID-19. Available from: <https://www.isibang.ac.in/~bisu/paper/KEOD%20CODO%207August2020-Preprint.pdf>
20. Yongqun He, et al. (2020). CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. Scientific Data, 7, 181. <https://doi.org/10.1038/s41597-020-0523-6>
21. Kejriwal, M. (2020). Knowledge Graphs and COVID-19: Opportunities, Challenges, and Implementation. Harvard Data Science Review. <https://doi.org/10.1162/99608f92.e45650b8>
22. Aasman, J. (2020). Stanford CS 520 Knowledge Graphs. <https://web.stanford.edu/class/cs520/abstracts/Aasman.pdf>

Thank you!!!

Collaboration!!!

Biswanath Dutta, Ph.D.

Email: dutta2005@gmail.com

bisu@isibang.ac.in

Twitter: @biswanath_dutta

<https://sites.google.com/view/bdutta>