The Knowledge Graph Conference (KGC)

Tutorial

3rd May 2022

Analysis of the Impact of COVID-19 Ontologies

Sven Groppe (University of Lübeck), Sanju Tiwari (Universidad Autonoma de Tamaulipas), Farah Benamara (IRIT-Université de Toulouse), Soror Sahri (Université Paris Cité)

https://www.ifis.uni-luebeck.de/~groppe/kgc22/

Worldwide: COVID-19 Confirmed Cases Daily

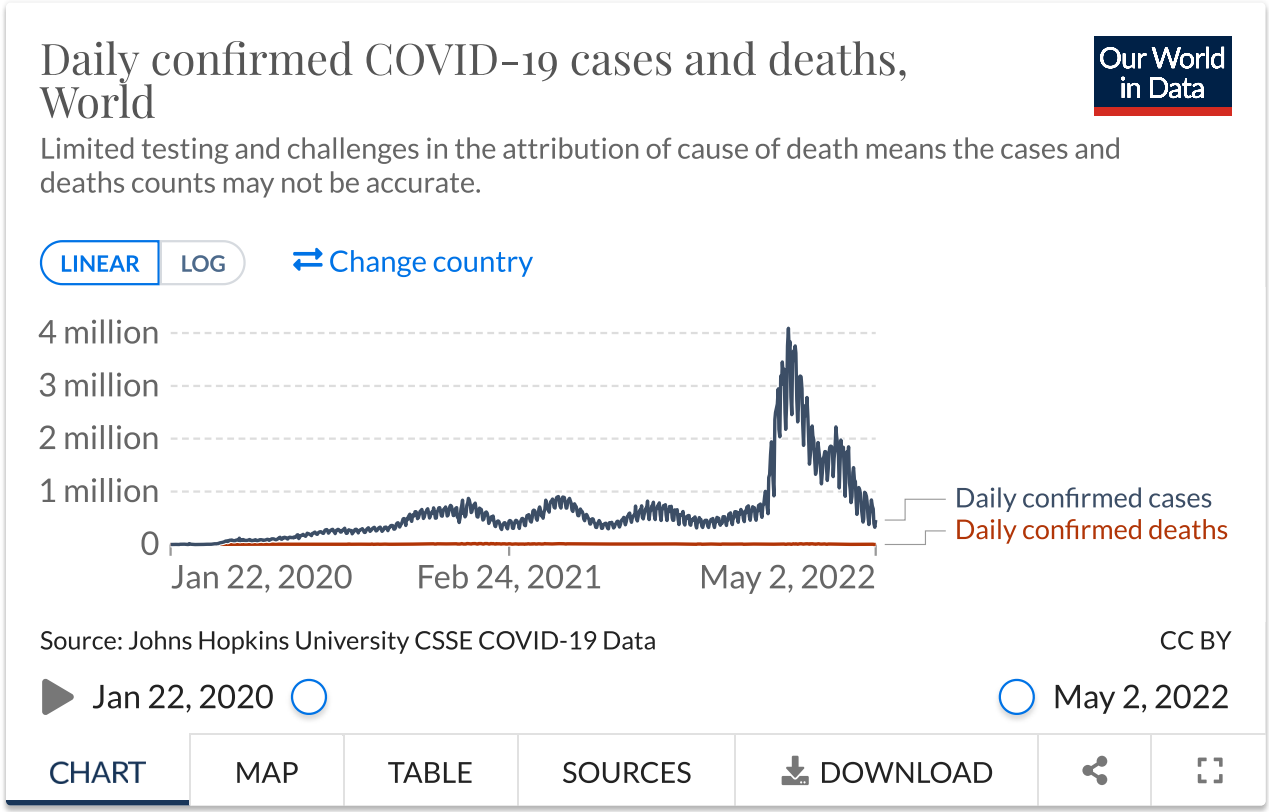
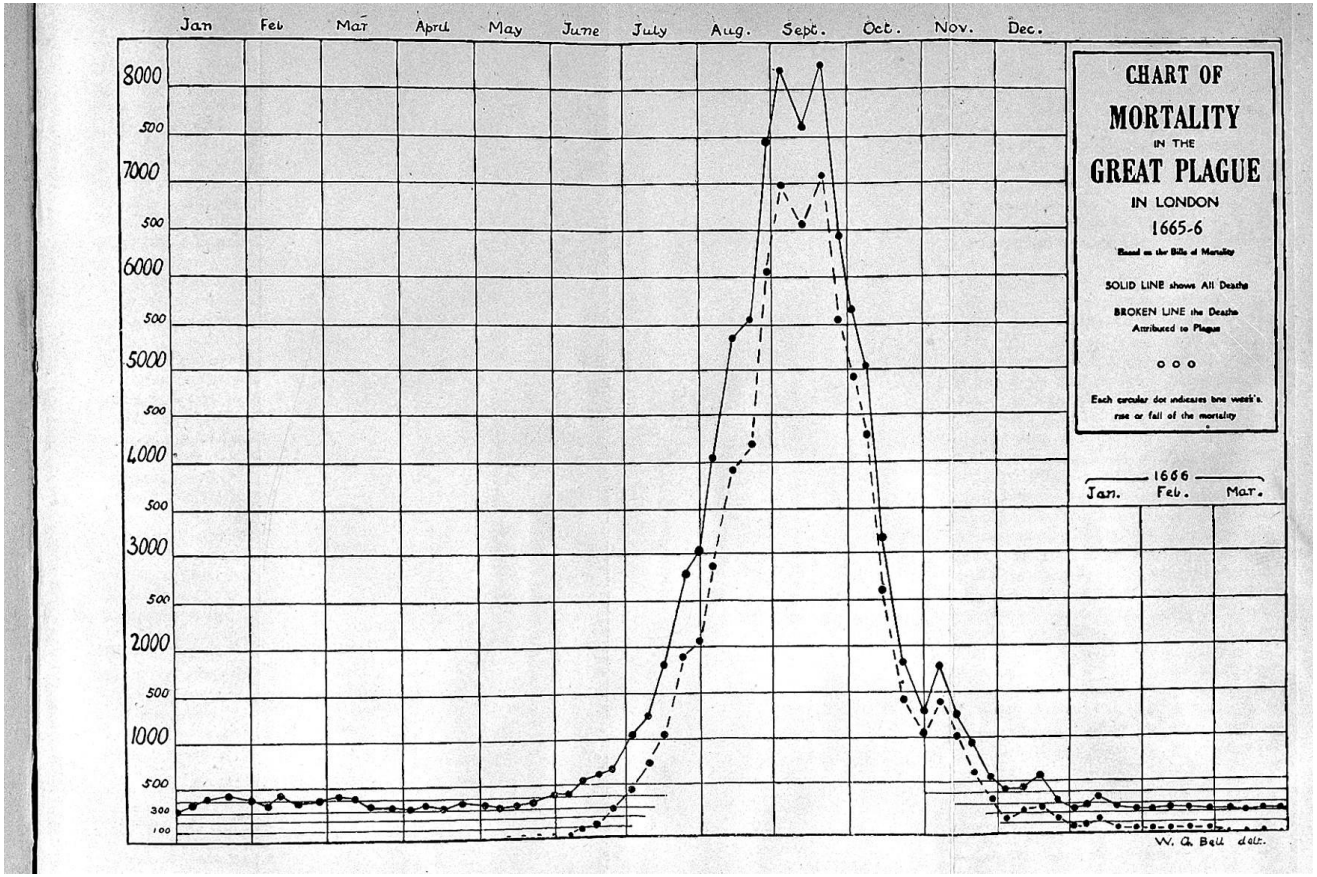
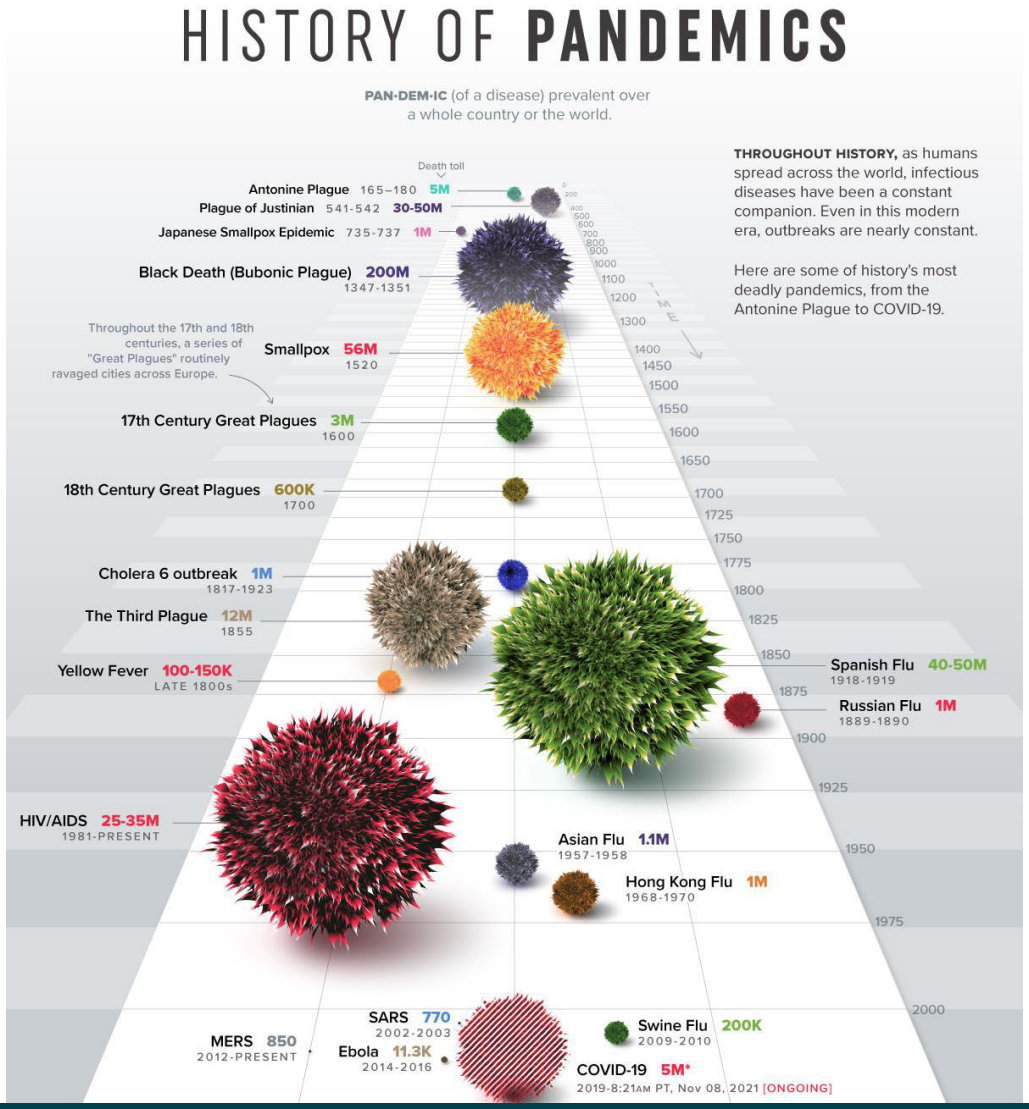


Chart of Mortality, Great Plague in London (1665 to 1666)



Source: https://fr.wikipedia.org/wiki/Fichier:Chart_of_Mortality,_Great_Plague_in_London_Wellcome_L0025769.jpg



Source: https://www.visualcapitalist.com/history-of-pandemics-deadliest/

Analysis of the Impact of COVID-19 Ontologies/Knowledge Graphs

(Sanju Tiwari)

Removed and available upon request...

How to assess KG’s quality? (Soror Sahri)

Removed and available upon request...

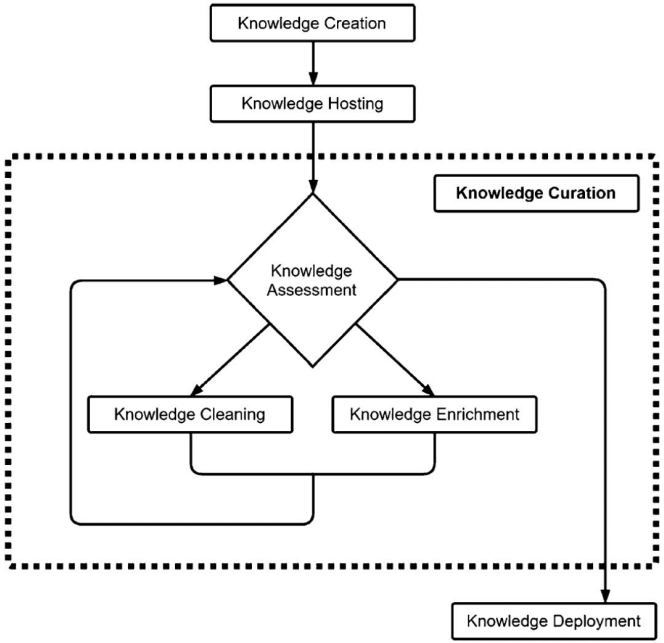
Existing tools for COVID-19 KG’s quality assessment (Soror Sahri)

Removed and available upon request...

KG Construction
(Farah Benamara)

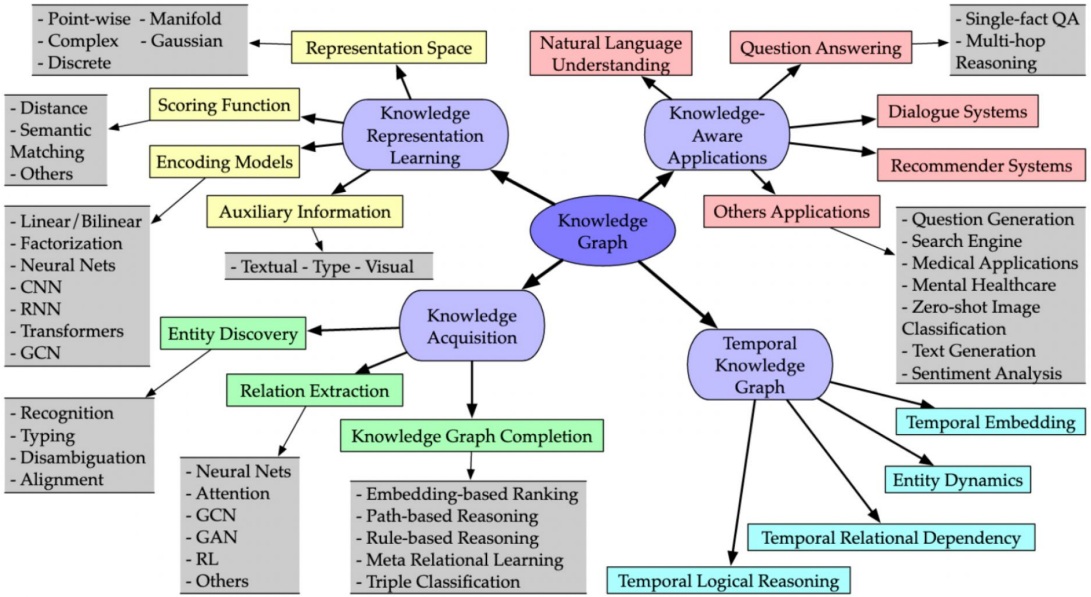
KG Pipeline

- *Knowledge acquisition*: Extract information from different sources, structuring it, and creating useful knowledge
- *Knowledge hosting*: Collect, store, and retrieve semantic annotations
- *Knowledge curation*: Improve data quality via knowledge assessment, cleaning, and enrichment
- *Knowledge deployment:* Publish the KG following some principles (e.g., the FAIR and the Linked Data Principles)



The process of building a KG (Fensel et al., 2020)
Knowledge Graphs - Methodology, Tools and Selected Use Cases

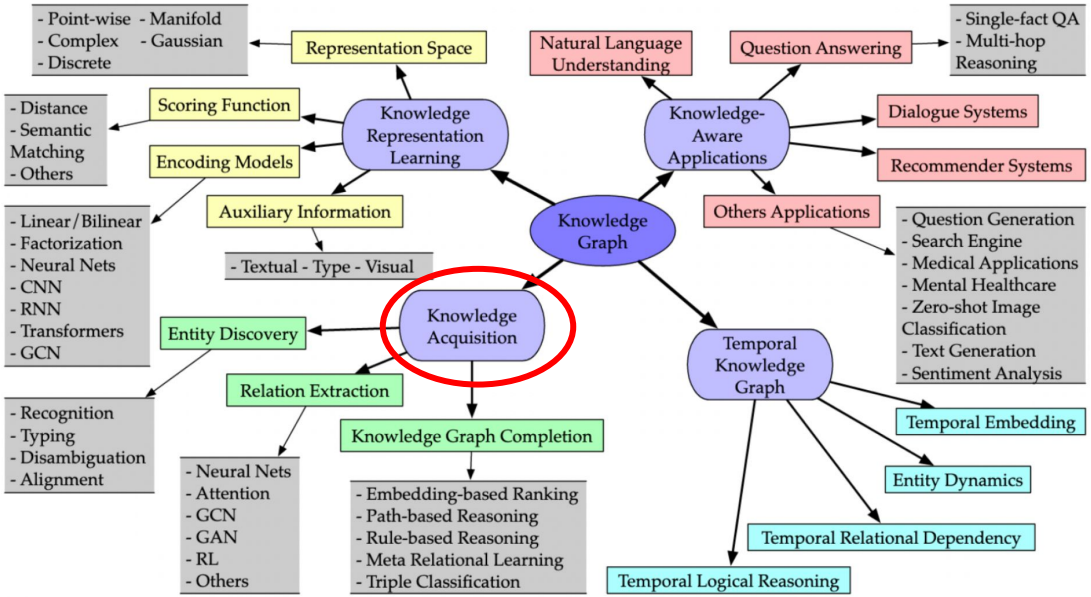
KG Research Topics



Main research topics on KGs (Ji et al., 2020)

A survey on knowledge graphs: Representation, acquisition and applications

KG Research Topics



Main research topics on KGs (Ji et al., 2020)

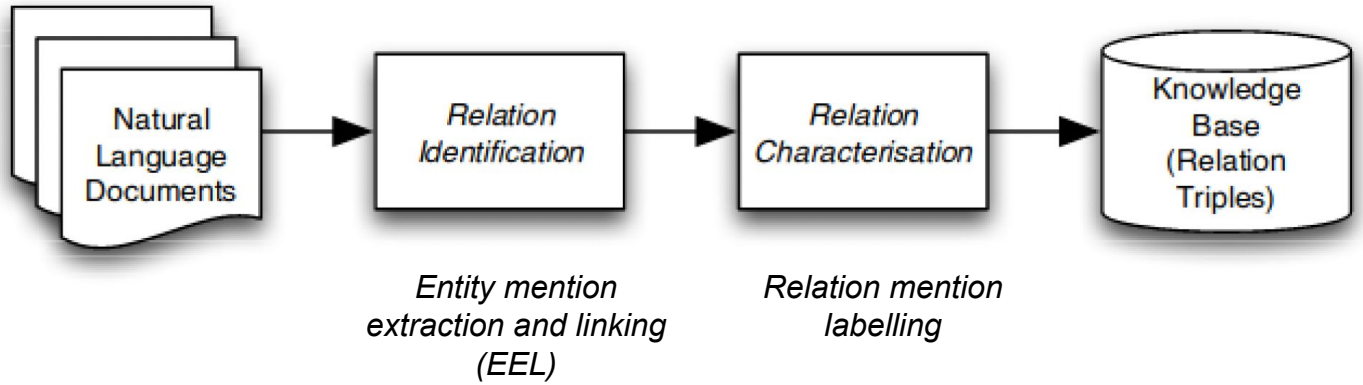
A survey on knowledge graphs: Representation, acquisition and applications

KG: Knowledge Acquisition

- Extract triples from either unstructured or semi-structured data
- Triples are then used for:
  - Creation and enrichment via extraction/mapping techniques from external sources
    - Direct contributions from human editors (crowd-sourcing or collaborative-editing platforms) ⇒ Errors/potential bias: humans are rather employed to verify and curate KGs
    - Textual contents via NLP and IE techniques
  - Refinement to complete and correct the KG
    - *Link discovery* (Nentwig et al., 2017): Predict the existence or the probability of correctness of missing edges
    - *Knowledge cleaning*
      - ***Fact validation*** (Syed et al., 2019): assign a plausibility score to a given edge
      - ***Inconsistency repairs*** (Bonatti et al., 2011): use ontological axioms to resolve inconsistencies

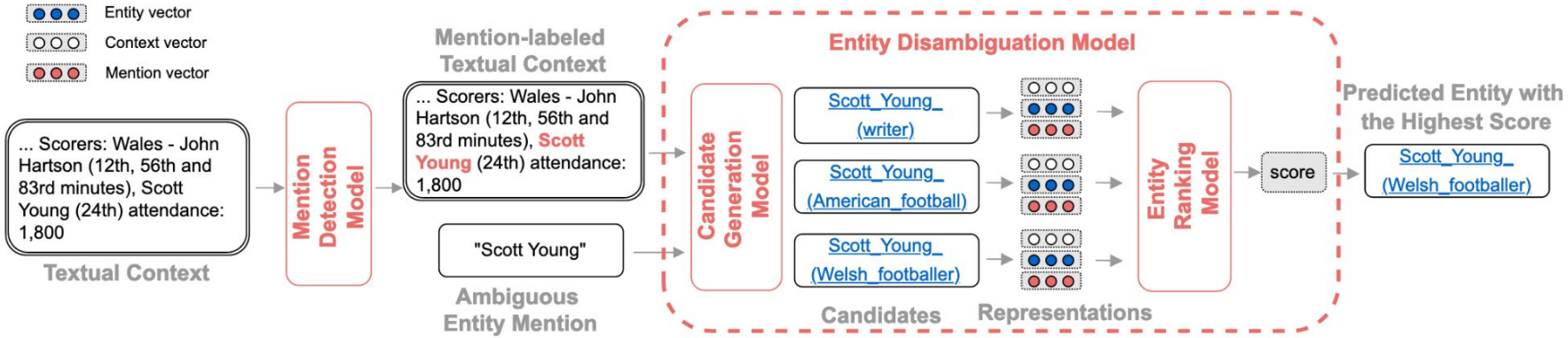
KG: Knowledge Acquisition

- Extract triples from either unstructured or semi-structured data



How Airbus became Boeing's greatest rival ⇒
Compete(Airbus, Boeing)
(Airbus, Competitor, Boeing)

Entity Extraction and Linking



The general architecture of a neural EEL system

Neural Entity Linking: A Survey of Models Based on Deep Learning (Sevgili et al, 2022)

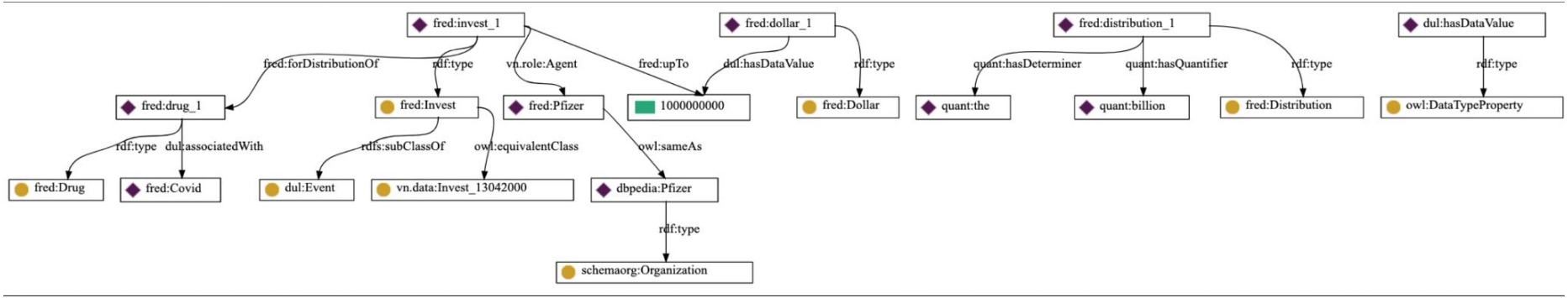
Relation extraction

- ReVerb (Fader et al., 2011) (http://reverb.cs.washington.edu/)
  - Binary RE from flat sentences, verb-based relation phrases.
- OLLIE (Mausam et al., 2012) (https://github.com/knowitall/ollie)
  - Goes beyond verbal-based relations (nouns, adjectives, and more)
- FRED (Gangemi et al., 2017) (http://wit.istc.cnr.it/stlab-tools/fred/)
  - 48 different languages and transform it to linked data
- MinIE (Gashteovski et al, 2017) (https://github.com/uma-pi1/minie)
  - Deals with polarity, modality, attribution, and quantities
- OpenIE (Angeli et al., 2015) (https://stanfordnlp.github.io/CoreNLP/openie.html)
  - Part of the OpenNLP toolkit, process long sentences

For more tools/ressources on Open Information Extraction (OpenIE), visit:
https://github.com/gkiril/oie-resources

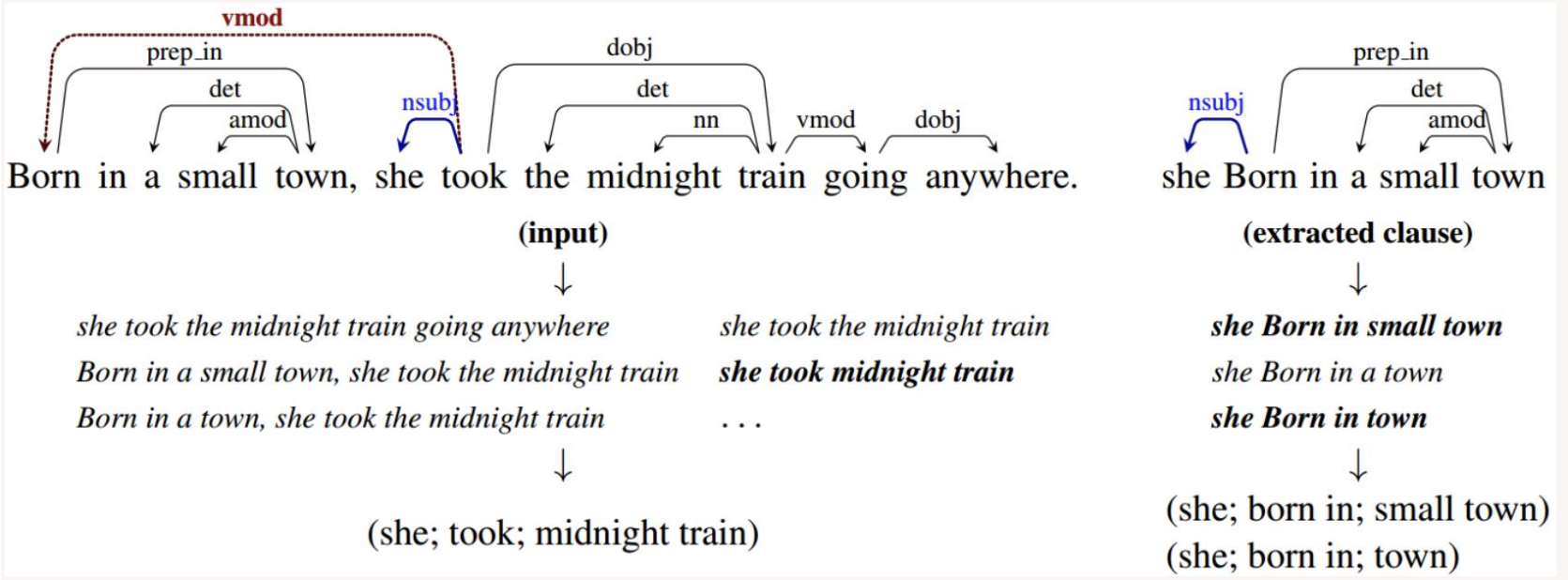
Relation extraction

- The FRED tool
  - Pfizer is investing up to $1 billion for distribution of the COVID drug



Relation extraction

- Stanford OpenIE
  - Relations does not need to be specified in advance



Existing COVID-KG Acquisition Pipelines

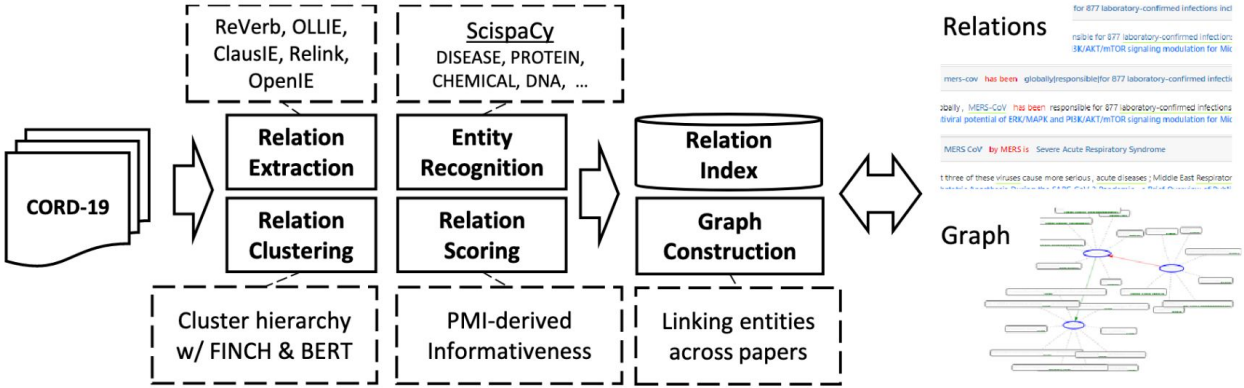
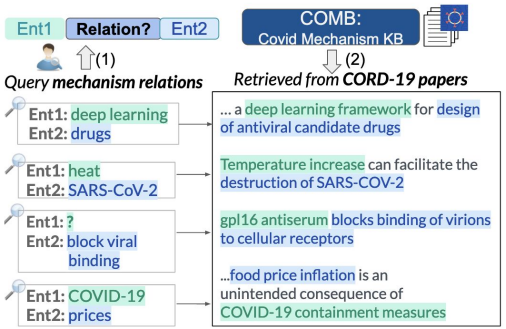


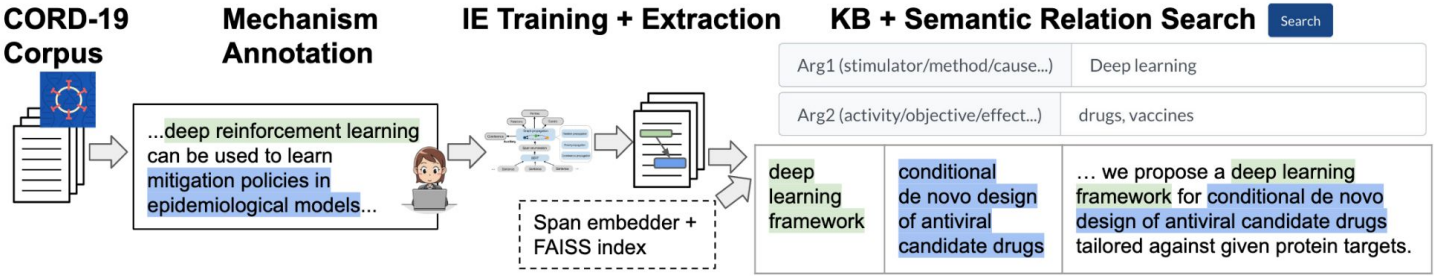
Figure 1: System Overview.

arg_1	rel	arg_2
[MERS-CoV]$_{GGP}$	include	[fever]$_{DISEASE}$, [chills/rigors]$_{DISEASE}$, [headache]$_{DISEASE}$, non-productive [cough]$_{DISEASE}$
[MERS-CoV]$_{GGP}$	is responsible for causing	lower [respiratory infections]$_{DISEASE}$ with [fever]$_{DISEASE}$ and [cough]$_{DISEASE}$

CovRelex: A COVID-19 Retrieval System with Relation Extraction (Tran et al. EACL 2021)
https://www.jaist.ac.jp/is/labs/nguyen-lab/systems/covrelex/home/

Existing COVID-KG Acquisition Pipelines



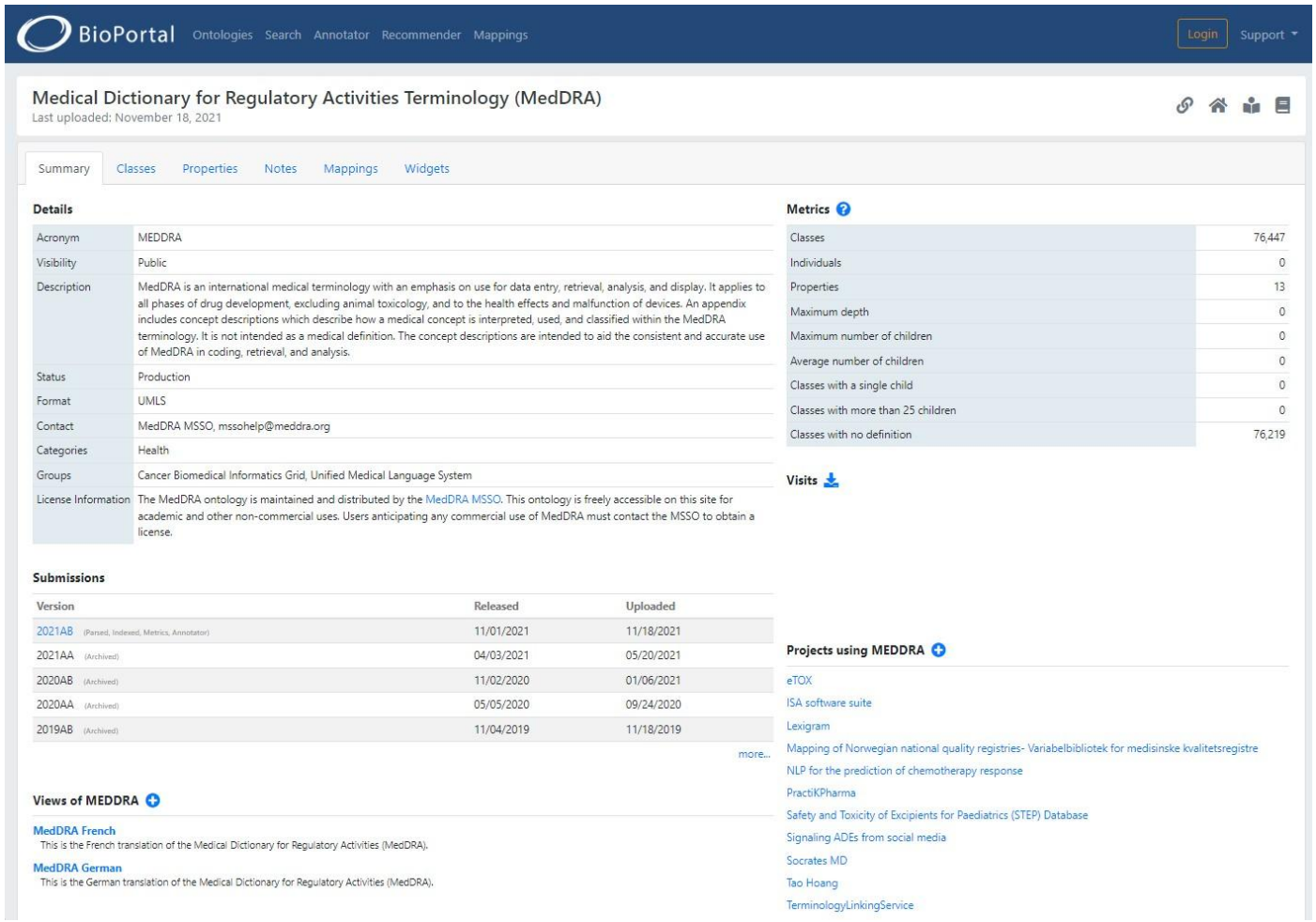


Extracting a Knowledge Base of Mechanisms from COVID-19 Papers (Hope et al. 2020)
https://covidmechanisms.apps.allenai.org/

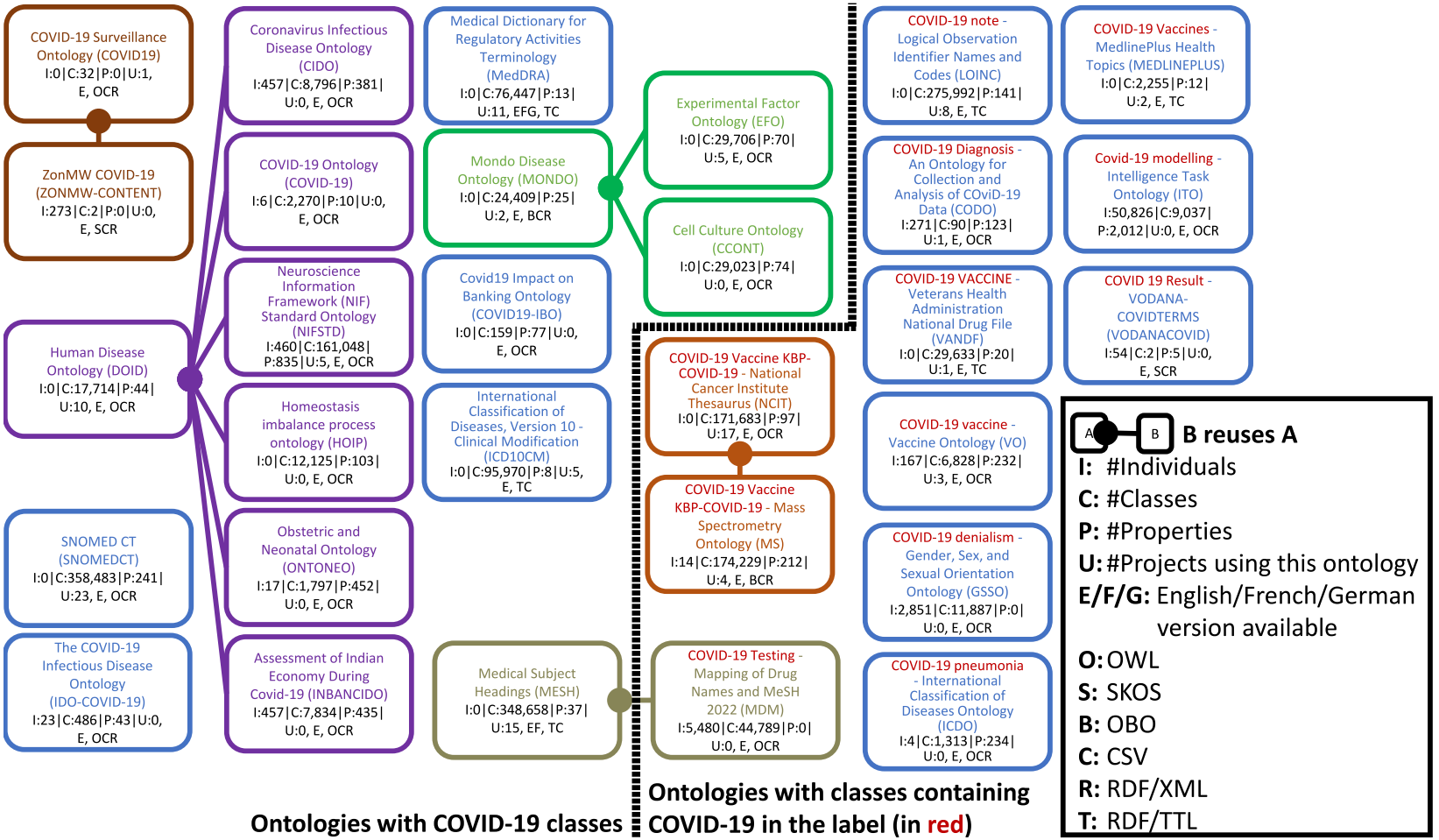
Search for COVID-19 classes on BioPortal

- Search URL: https://bioportal.bioontology.org/search?q=COVID-19

BioPortal information about ontologies



COVID-19 Ontologies on BioPortal



How to measure impact of COVID-19 Ontologies? Number of reuses in ontologies

- High number of reuses in other ontologies
  - is a sign for the popularity of the reused ontology and its classes, – let ontology users stumple over the reused ontology when applying the reusing ontology, and – makes more reusings more likely
- Resulting Ranking:
  1. Class COVID-19 of the Human Disease Ontology (DOID): 6 reuses
  2. Class COVID-19 of the Mondo Disease Ontology (MONDO): 2 reuses
  3. Class COVID-19 of COVID-19 Surveillance Ontology (COVID19), of Medical Subject Headings (MESH) and of National Cancer Institute Thesaurus (NCIT): 1 reuse

Which ontology to choose for your COVID-19 application?

- Not all COVID-19 ontologies will survive
  - Many ontologies will not maintained any more in the future
    - .darkblue[funded project ends $\longrightarrow$ no money any more] for maintenance
    - some ontologies are not widely used
- Choose COVID-19 ontology with high impact
  - $\longrightarrow$ higher probability for being maintained in the future
  - $\longrightarrow$ higher interoperability with other applications

Which ontology to choose for your COVID-19 application?

- Not all COVID-19 ontologies will survive
  - Many ontologies will not maintained any more in the future
    - .darkblue[funded project ends $\longrightarrow$ no money any more] for maintenance
    - some ontologies are not widely used
- Choose COVID-19 ontology with high impact
  - $\longrightarrow$ higher probability for being maintained in the future
  - $\longrightarrow$ higher interoperability with other applications

How to measure impact of COVID-19 Ontologies?

How to measure impact of COVID-19 Ontologies? Number of usages in projects

- Number of usages in projects
  - obvious measure for the ontology impact
  - in BioPortal is incomplete
    - more a sign for a good maintenance of the ontology & motivated ontology developers and project members pointing out these usages
    - remains as a good metric for the ontology impact
- Resulting Ranking:
  1. 32 projects: SNOMED CT
  2. 17 projects: NCIT
  3. 15 projects: MESH
  4. 11 projects: MedDRA
  5. 10 projects: DOID
  6. 8 projects: LOINC
  7. 5 projects: NIFSTD, EFO, ICD10CM
  8. 4 projects: MS
  9. 3 projects: VO
  10. 2 projects: MONDO, MEDLINEPLUS
  11. 1 project: COVID-19 Surveillance Ontology, CODO, VANDF

How to measure impact of COVID-19 Ontologies? Number of direct and indirect usages in projects

- Number of direct and indirect usages in projects
  - **direct usage:** the usage of a given ontology in projects
  - **indirect usage:** the project usage of an ontology reusing the given ontology
- Resulting Ranking:
  1. 32 projects: SNOMED CT
  2. 21 projects: NCIT
  3. 15 projects: MESH
  4. 15 projects: DOID
  5. 11 projects: MedDRA
  6. 8 projects: LOINC
  7. 7 projects: MONDO
  8. 5 projects: NIFSTD, EFO, ICD10CM
  9. 4 projects: MS
  10. 3 projects: VO
  11. 2 projects: MEDLINEPLUS
  12. 1 project: COVID-19 Surveillance Ontology, CODO, VANDF

How to measure impact of COVID-19 Ontologies? Weighted Combinations of Number of Reuses and Projects

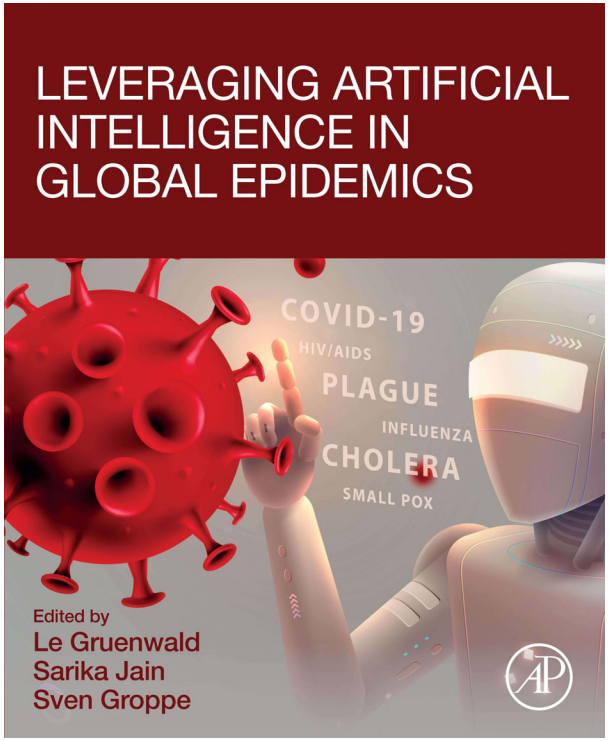
- Both metrics (#reuses and projects) are independent in theory, but
  - a high number in using projects often results in an increased number of reusing ontologies and vice versa in practice
  - BioPortal is incomplete $\rightarrow$ #reuses and projects are sometimes extremely different
- Idea: Calculating a balanced metric for these extreme cases
  - Finding a good balanced metric based on rigorous analysis open challenge for future work

How to measure impact of COVID-19 Ontologies? Open Challenges and Future Work

- Datasets (other than BioPortal) for usages of ontologies and projects
- Metrics based on other properties
  - Searches in ontology search engines
  - Number of instances of ontology classes in knowledge graphs
  - Number of applications using these ontologies
  - Number of accesses to instances of ontologies in applications
  - ...
- Impact of knowledge graphs and datasets

Further Reading

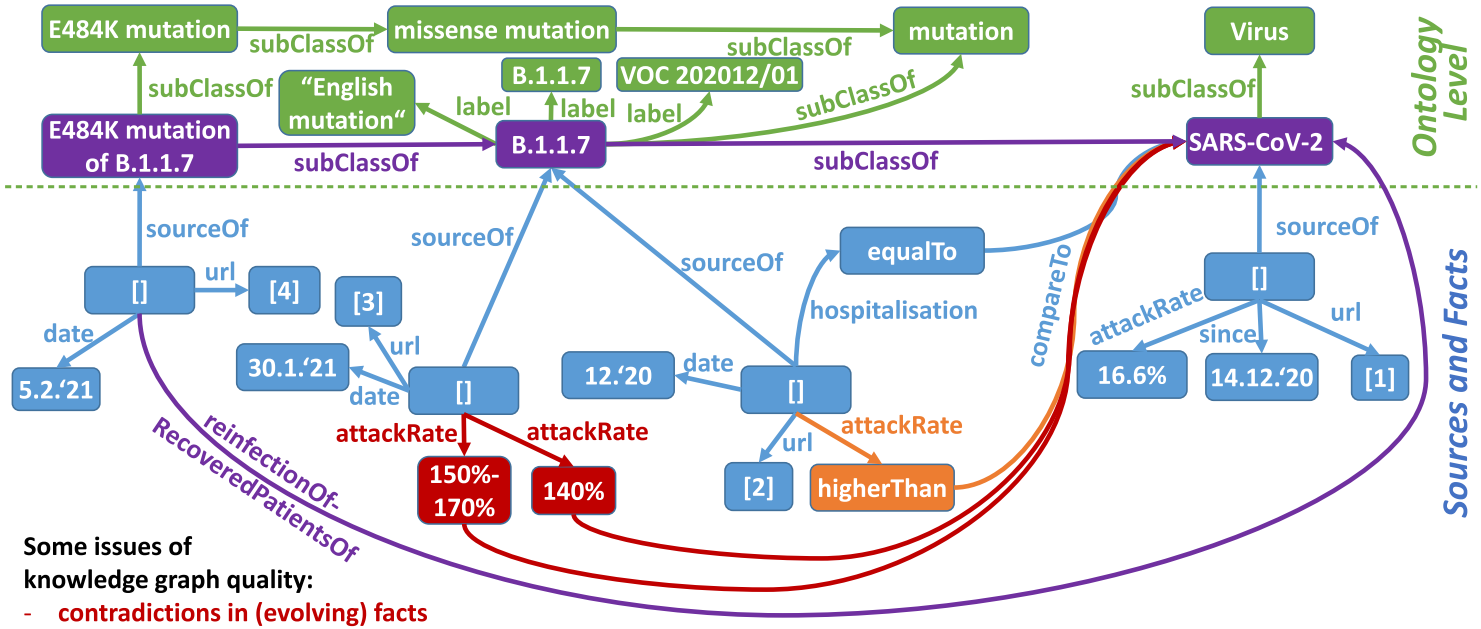
Table of Contents



1. An overview of global epidemics and the challenges faced
2. Leveraging artificial intelligence and digital tech to help citizens, societies, and economies survive and strive during pandemics
3. Towards an alternative to lockdown: Pandemic management leveraging digital technologies and artificial intelligence
4. Exploratory study of existing approaches for analyzing epidemics
5. A data science perspective of real-world COVID-19 databases
6. Preparing with predictions: forecasting epidemics with artificial intelligence
7. The worldwide methods of artificial intelligence for detection and diagnosis of COVID-19
8. The role of AI in digital contact tracing
9. Covid-19 accelerating the dynamics of Artificial Intelligence disruption
10. Use of artificial intelligence in pharmacovigilance for social media network
11. System-level knowledge representation for artificial intelligence during pandemics

COVID-19 KG Open Challenges:

For example: Vague Formulations ↔ NLP



Some issues of knowledge graph quality:

- **contradictions in (evolving) facts**
- **checking vague formulations and compare them with other given information**
- **errors with ambiguity: E484K mutation of B.1.1.7 is a SARS-CoV-2 virus, but reinfection with E484K is only possible for patients recovered from SARS-CoV-2 virus not mutating E484K (→ introduce class "SARS-CoV-2 without E484K mutation")**

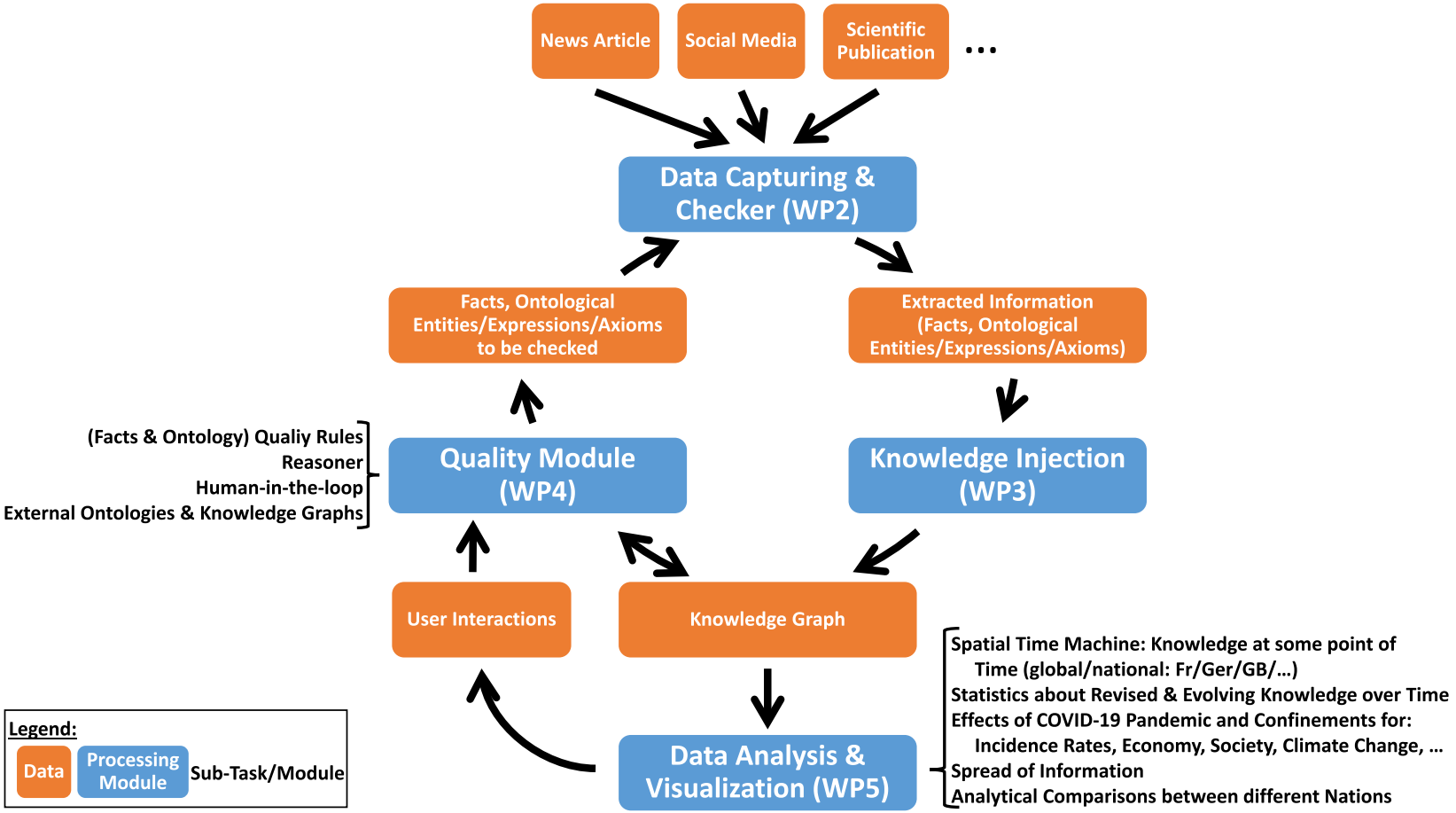
[1] https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2774102
[2] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/959361/Technical_Briefing_VOC202012-2_Briefing_2.pdf
[3] https://www.ruhr24.de/service/corona-britische-mutation-neue-studie-mutante-virus-toedlicher-sterberate-deutschland-90184403.html
[4] https://www.bmj.com/content/372/bmj.n359

High Quality Knowledge Graphs from recent English, French and German Emergent Trends with the example of COVID-19
Main Objectives:

- project starting soon (universities of Paris, Toulouse and Lübeck)
- Generate high-quality Kowledge Graph for emergent English, French and German trends with the example of COVID-19
  - and make the resulting graph publicly available
- Compare the facts extracted from different data sources
  - e.g., last year scientific publications, news articles and headlines, social media like twitter, existing KGs
- identify conflicting assertions as well as complementary ones
- Investigate differences between En/Fr/Ger data sources
- An extensive data analysis & visualization of research findings based on the time machine
- Evaluate the quality of the KG throughout the process of KG enrichment and its querying

This work is jointly funded by the DFG – Project-ID 490998901, and the French National Agency QualityOnt ANR-21-CE23-0036-01

High Quality Knowledge Graphs from recent English, French and German Emergent Trends with the example of COVID-19



This work is jointly funded by the DFG – Project-ID 490998901, and the French National Agency QualityOnt ANR-21-CE23-0036-01

Expected Results

- Data analysis and visualization tools in order to deal with the following issues:
  - extensive statistics about the COVID-19 pandemic especially with focus on facts evolving over time and differences of knowledge in different nations like:
    - number of contradicting and revised facts,
    - number of changes of COVID-19 confinements,
    - calculate effects (using machine learning approaches) of COVID-19 pandemic and confinements for incidence rates, economy, society and climate change.
  - a visualization tool for visualizing the results obtained by the statistics module by:
    - an easy-to-use tool, but
    - which offers flexible ways for querying the data in order to support sophisticated analysis,
    - with satisfactorily answering queries with consideration of users' requirements on data quality as well as the fitness of data to meet those requirements

This work is jointly funded by the DFG – Project-ID 490998901, and the French National Agency QualityOnt ANR-21-CE23-0036-01

Summary - COVID-19 Pandemic

- Statistics
  - Health: incidence rates, global health security index
- Timeline of discoveries and tech trends
- Predictions of incidence rates and other COVID-19 data in time series
- COVID-19 ontologies, knowledge graphs & data sets
  - Overview over existing ontologies, knowledge graphs and data sets
  - Quality assessment
  - Knowledge graph construction
- Further reading