Supporting findability of COVID-19 research with large-scale text mining of scientific publications

Luxembourg Centre for Systems Biomedicine



Danielle Welter¹, Carlos Vega¹, Maria Biryukov¹, Valentin Grouès¹, Soumyabrata Ghosh¹, Reinhard Schneider^{1,} Venkata Satagopam¹ ¹ Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Luxembourg

The BioKB platform¹

- Exploits text mining and semantic technologies to help researchers easily access semantic content of thousands of abstracts and full text articles
- Concepts from a range of contexts, including proteins, species, chemicals, diseases and biological processes are tagged based on existing dictionaries of controlled terms
- Co-occurring concepts are classified based on their asserted relationship and the resulting subject-relation-object triples are stored in a publicly accessible human- and machine-readable knowledgebase
- All concepts in the BioKB dictionaries linked to stable, persistent identifiers
 - Resource accession such as an Ensembl², Uniprot³ or PubChem⁴ ID for genes, proteins and chemicals
 - Ontology term ID for diseases, phenotypes and other ontology terms

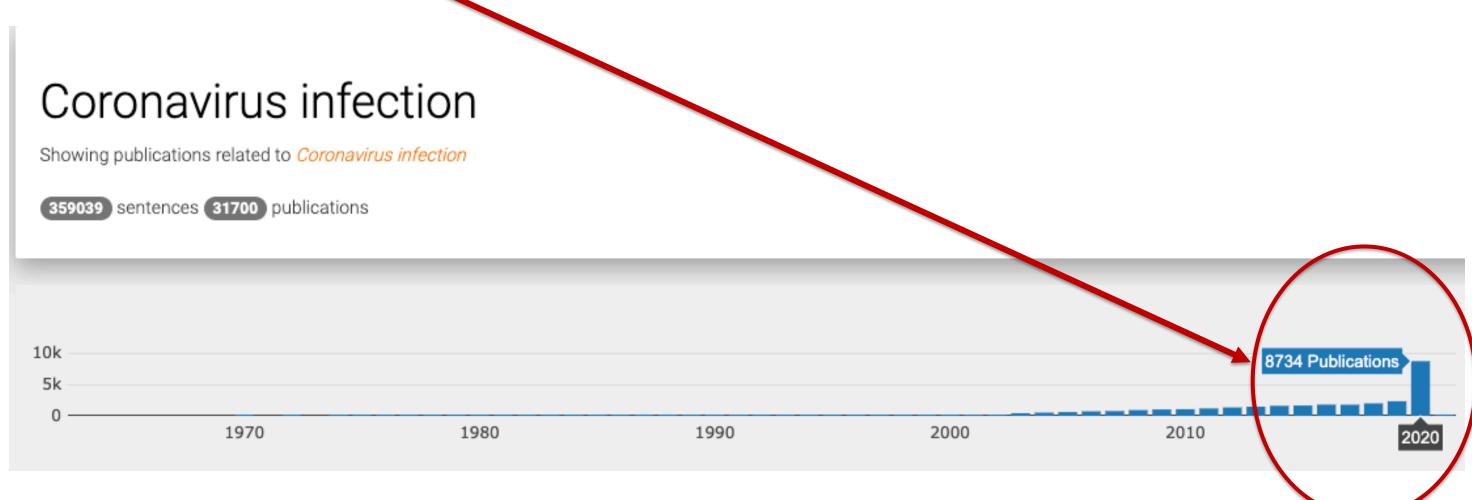
https://biokb.lcsb.uni.lu/

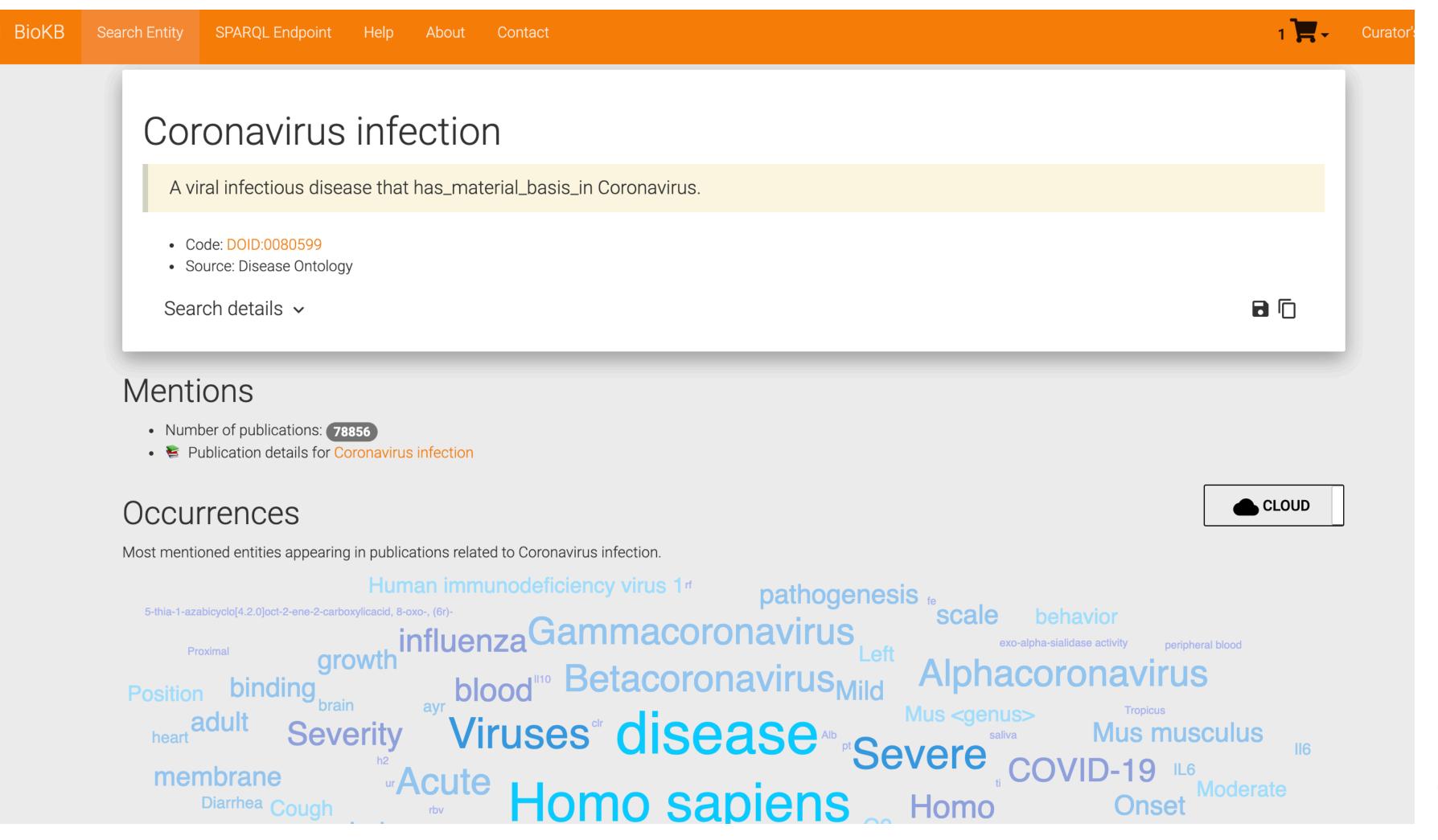
Background

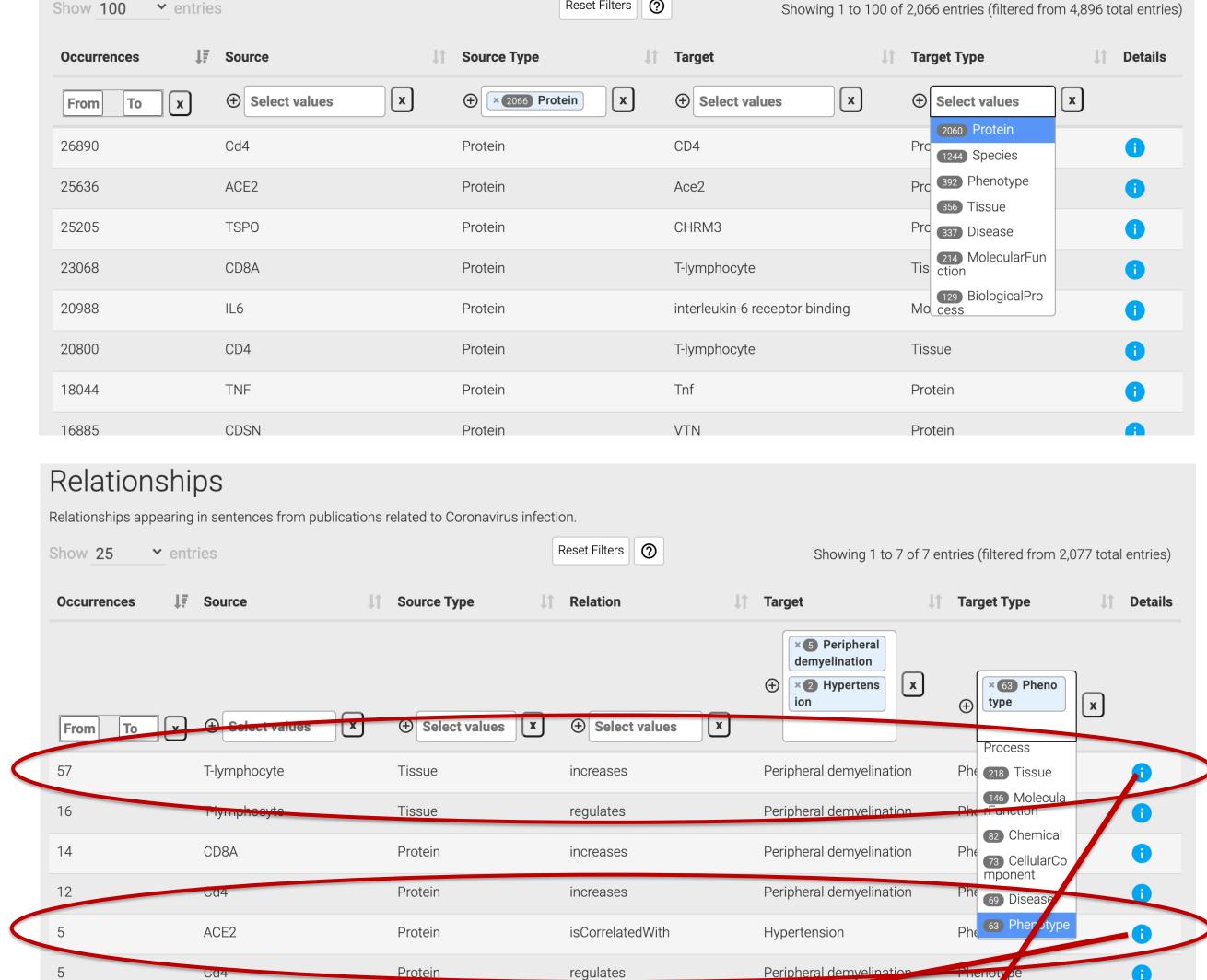
Co-occurrences

Co-occurrences appearing in sentences from publications related to Coronavirus infection

- COVID-19 pandemic lots of research efforts quickly redirected towards studies on SARS-CoV2 and COVID-19 disease
 - Sequencing and assembly of viral genomes
 - Elaboration of robust testing methodologies
 - > Development of treatment and vaccination strategies
 - Flurry of scientific publications around SARS-CoV-2 and COVID-19
 - Increasingly difficult for researchers to stay up-to-date with latest trends and developments in this rapidly evolving field





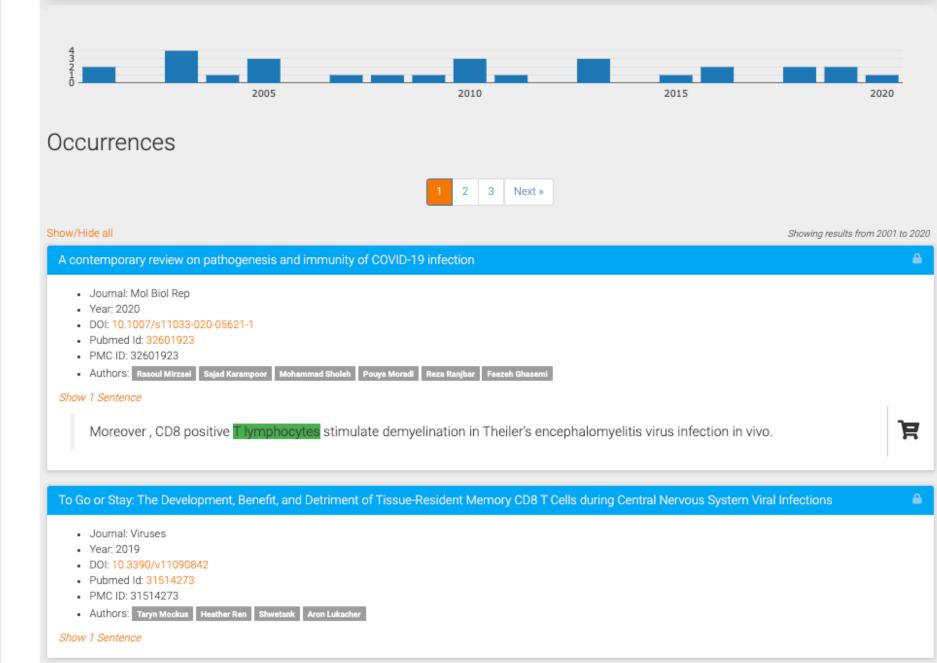


COVID-related platform improvements

Extension of the underlying dictionaries to increase the sensitivity of the text mining pipeline to viral data, including

- > Additional viral species (via NCBI Taxonomy⁵ identifiers)
- Phenotypes from the Human Phenotype Ontology⁶ (HPO)
- COVID-related concepts including clinical and laboratory tests from the COVID-19 ontology⁷
- Additional diseases (DO⁸)
- Biological processes (GO⁹)
- All viral proteins found in UniProt and gene entries from EntrezGene¹⁰

ACE2 - isCorrelatedWith - Hypertension 5 relationships in 5 publications Filtered for: Coronavirus infection Occurrences Year: 2020 DOI: 10.6061/clinics/2020/e1912 Pubmed Id: 32428113 This has been corroborated in Nefhcrex_AT1aRflox/flox_SPEC_7/NNS, indicating that ACE-2 and sACE-2 levels in the brain are associated with the etiology of neurogenic hypertension 20. Journal: Intern Emerg Med Year: 2020 Pubmed Id: 32399954 PMC ID: 32399954 Show 1 Sentence E2 and Microbiota: Emerging Targets for Cardiopulmonary Disease Therapy



erential neurodegenerative phenotypes are associated with heterogeneous voiding dysfunction in a coronavirus-induced model of multiple

T-lymphocyte - increases - Peripheral demyelination

References

- https://biokb.lcsb.uni.lu/
- https://www.ensembl.org/
- https://www.uniprot.org/ https://pubchem.ncbi.nlm.nih.gov/

- https://www.ncbi.nlm.nih.gov/taxonomy
- https://hpo.jax.org/
- https://bioportal.bioontology.org/ontologies/COVID-19/?p=summary
- https://disease-ontology.org/

- http://geneontology.org/
- 10. https://www.ncbi.nlm.nih.gov/gene/

Protein

57 relationships in 28 publications

Filtered for: Coronavirus infection

