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

Luxembourg Centre for Systems Biomedicine

Danielle Welter[1], Carlos Vega[1], Maria Biryukov[1], Valentin Grouès[1], Soumyabrata Ghosh[1], Reinhard Schneider[1], Venkata Satagopam[1]

[1] Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Luxembourg

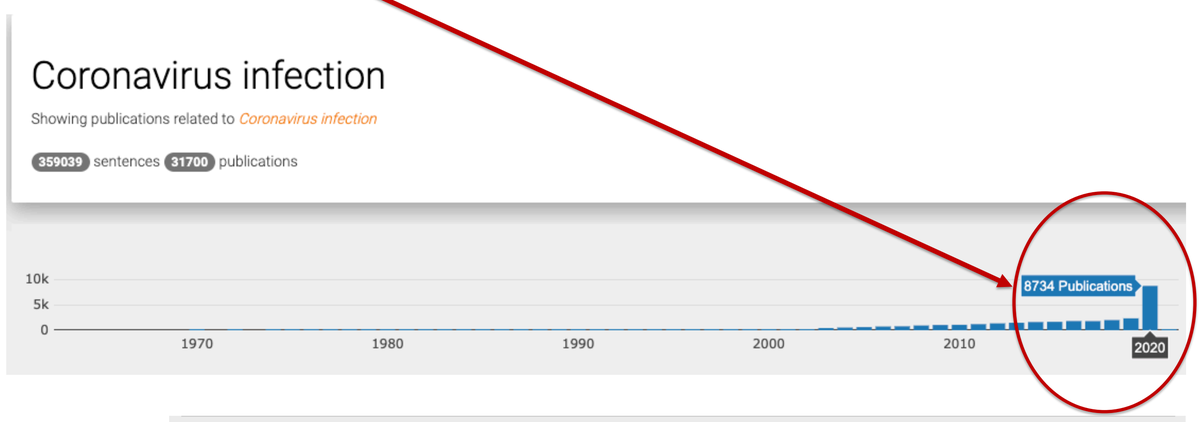
The BioKB platform[1]

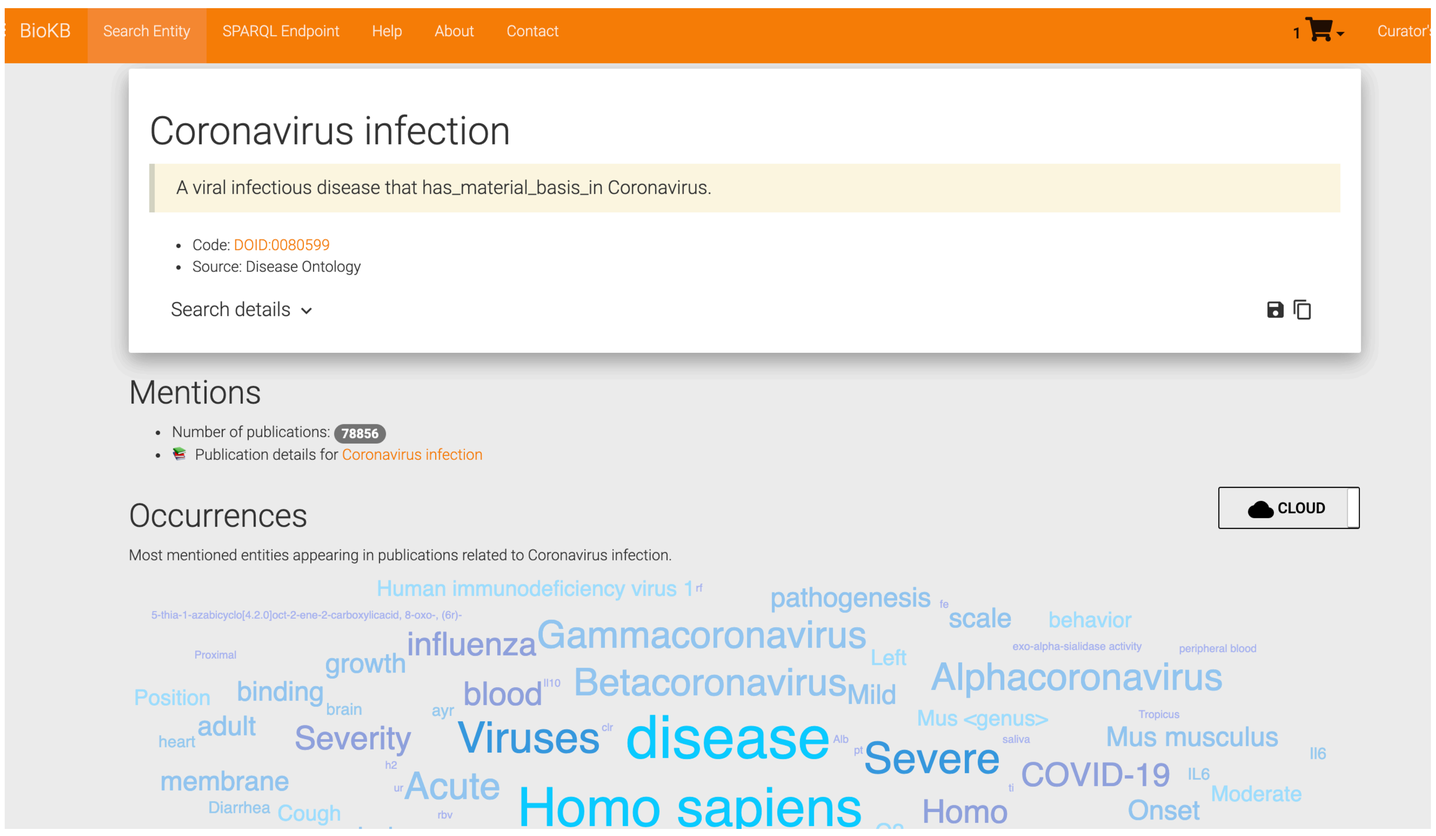
- Exploits text mining and semantic technologies to help researchers easily access semantic content of thousands of abstracts and full text articles
- Concepts from a range of contexts, including proteins, species, chemicals, diseases and biological processes are tagged based on existing dictionaries 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 Ensembl[2], Uniprot[3] or PubChem[4] ID for genes, proteins and chemicals
  - ➢ Ontology term ID for diseases, phenotypes and other ontology terms

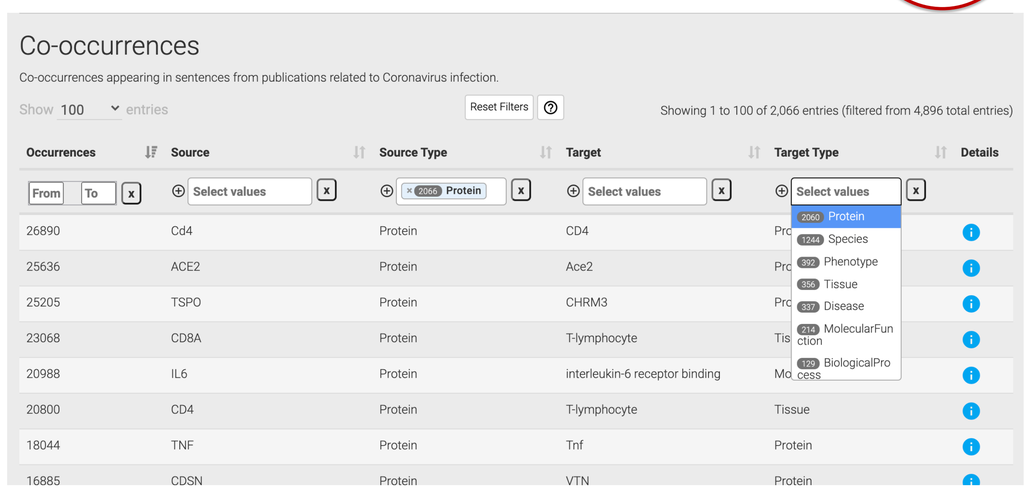
https://biokb.lcsb.uni.lu/

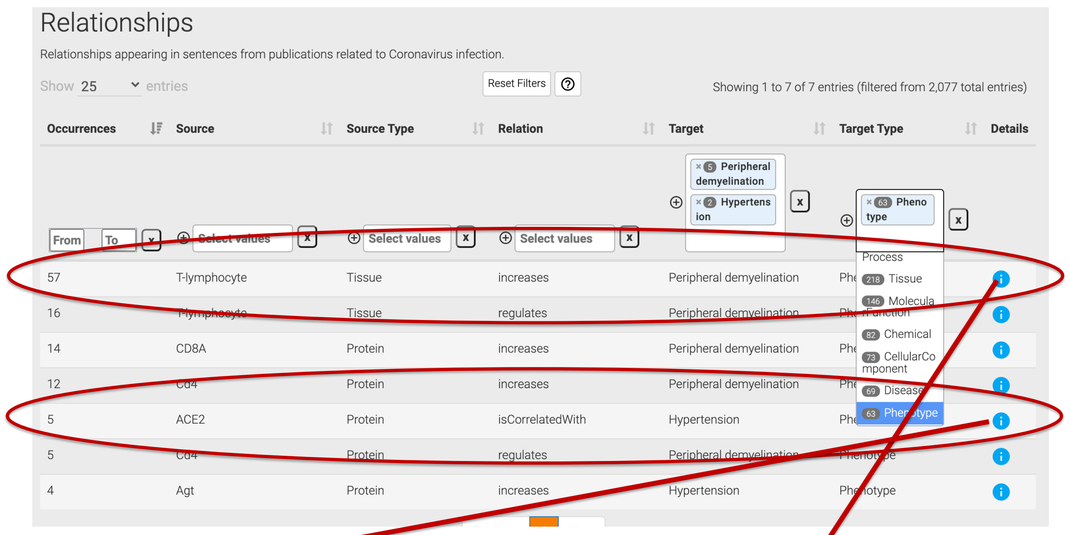
Background

- 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





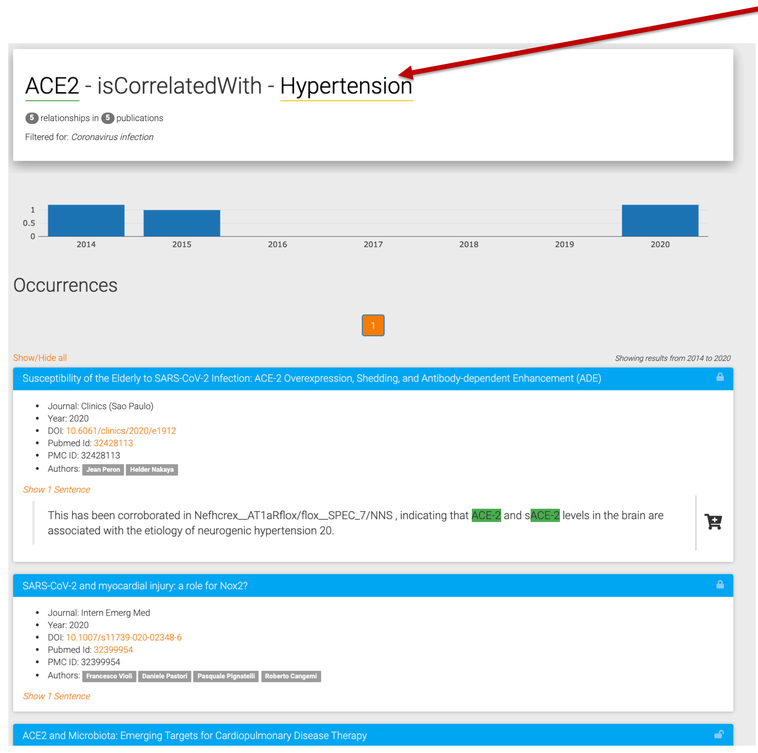


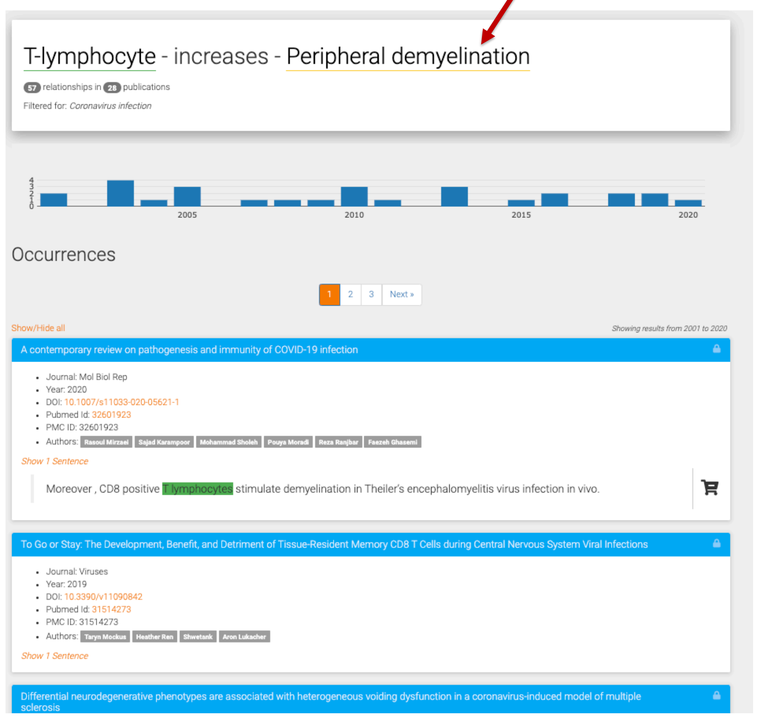


COVID-related platform improvements

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

- ➢ Additional viral species (via NCBI Taxonomy[5] identifiers)
- ➢ Phenotypes from the Human Phenotype Ontology[6] (HPO)
- ➢ COVID-related concepts including clinical and laboratory tests from the COVID-19 ontology[7]
- ➢ Additional diseases (DO[8])
- ➢ Biological processes (GO[9])
- ➢ All viral proteins found in UniProt and gene entries from EntrezGene[10]





References

1. https://biokb.lcsb.uni.lu/
2. https://www.ensembl.org/
3. https://www.uniprot.org/
4. https://pubchem.ncbi.nlm.nih.gov/
5. https://www.ncbi.nlm.nih.gov/taxonomy
6. https://hpo.jax.org/
7. https://bioportal.bioontology.org/ontologies/COVID-19/?p=summary
8. https://disease-ontology.org/
9. http://geneontology.org/
10. https://www.ncbi.nlm.nih.gov/gene/