

# Controlled Vocabulary for COVID-19 (COVoc)

A COVID-19 ontology applied to literature triage

Zoë May Pendlington

Ontologist

Samples, Phenotypes and Ontologies Team  
(Parkinson)

EMBL-EBI

# Contents

- COVoc objectives
  - Why was COVoc created?
  - Specific use case
- COVoc development
  - Data collection
  - OBO alignment
  - OWL generation and release pipeline
- Controlled vocabulary – or ontology?
- Future plans

# Why?

Search worldwide, life-sciences literature

[Coronavirus articles and preprints](#) Search examples: ["breast cancer"](#) [Smith J](#)

[Recent history](#) [Saved searches](#)

Search only

1-25 of 55,457 results



Search worldwide, life-sciences literature

[Coronavirus articles and preprints](#) Search examples: ["breast cancer"](#) [Smith J](#)

[Recent history](#) [Saved searches](#)

Search only

1-25 of 7,326 results



Search worldwide, life-sciences literature

[Coronavirus articles and preprints](#) Search examples: ["breast cancer"](#) [Smith J](#)

[Recent history](#) [Saved searches](#)

Search only

1-25 of 20,030 results



“barrier gestures” “sneezing into one's elbow” “contact tracking” “acro-ischemic skin disorder”  
“social distancing” etc

# Objectives



- **COVoc**: Controlled Vocabulary for COVID-19
- Objective:
  - A new ontology fully dedicated to literature triage and curation of COVID-19 related literature

# Development - Data collection

- The team at SIB collected an initial set of terms.
  - 9 axes

Biological/Medical vocabulary

Cell lines

Chemicals

Clinical Trials

Conceptual Entities

Disease and Syndrom

Geographic locations

Organism

Proteins - Genomes

Definition: any sign or symptom and well defined diagnosis associated with covid-19, including comorbidities and virus infection.

A	B	C	D	E	
Definition: any sign or symptom and well defined diagnosis associated with covid-19, including comorbidities and virus infection.					
id	Preferred term	Synonymes	Source	CUI UMLS	Sem
DIS_1	abdominal pain		PUBLIC - Clinical observations extrac		
DIS_2	acetaminophen		PUBLIC - Clinical observations extrac		
DIS_3	acute respiratory distress syndrome	ARDS   respiratory distress	PUBLIC - Clinical observations extrac		
DIS_4	afebrile		PUBLIC - Clinical observations extrac		
DIS_5	albuterol		PUBLIC - Clinical observations extrac		
DIS_6	anaphylaxis		PUBLIC - Clinical observations extrac		
DIS_7	anemia		PUBLIC - Clinical observations extrac		
DIS_8	anxiety		PUBLIC - Clinical observations extrac		
DIS_9	asthma		PUBLIC - Clinical observations extrac		
DIS_10	cardiac complications	cardiac complication			
DIS_11	chest pain		PUBLIC - Clinical observations extrac		
DIS_12	chills		PUBLIC - Clinical observations extrac		
DIS_13	Chronic Obstructive Airway Disease			C0024117	[Dis
DIS_14	Chronic obstructive pulmonary disease	copd			
DIS_15	Coinfection			C0275524	[Dis
DIS_16	Communicable Diseases			C0009450	[Dis
DIS_17	congestion		PUBLIC - Clinical observations extrac		
DIS_18	Coronavirus Infections			C0206750	[Dis
DIS_19	cough	Coughing		C0010200	[Sigr
DIS_20	depression		PUBLIC - Clinical observations extrac		
DIS_21	diabete	diabetes			
DIS_22	diarrhea		PUBLIC - Clinical observations extrac		

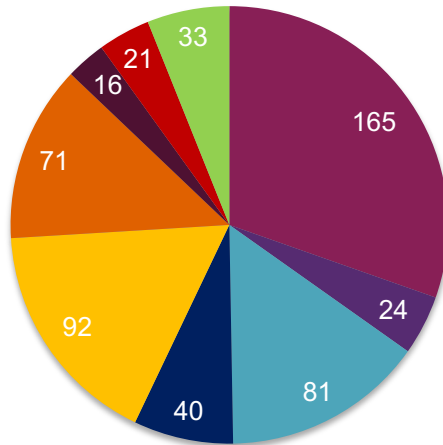
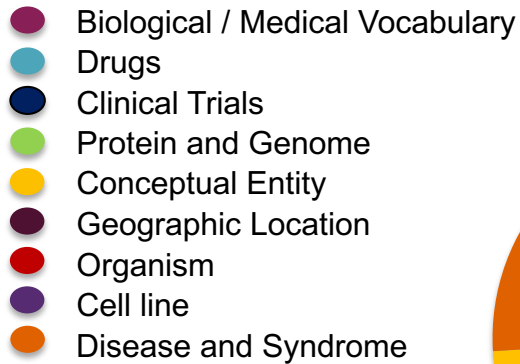
# OBO alignment

**COVOC terms** ☆ 📁 ☁

Fichier Édition Affichage Insertion Format Données Outils Modules complémentaires Aide Dernière modification il y a 4 jours par Paola Roncaglia

	A	B	C	D	E	F	G	H	I
1	Sheet	Definition	COVOC temporary id	id (for template_	ontologies				
2	BiologicalMedical vocabulary	Biological and medical terms	COVOC:001	biomed	bao, chebi, chmo, cl, efo, go, ido, maxo, mondo, ncbitaxon, ncit, obi, omit, pr, uberon				
3	Cell lines	cell lines used in research about C	COVOC:002	cell_lines	clo, efo				
4	Chemicals	substances in link with COVID-19	COVOC:003	chemical	chebi, ncit				
5	Clinical trials	list of clinical trials about COVID-1	COVOC:004	clinicaltrial					
6	Conceptual Entities	generic concepts	COVOC:005	conceptual	efo, obi, pato, chebi				
7	Disease and Syndrome	any sign or symptom and well defi	COVOC:006	disease	hp, chebi, mondo, efo				
8	Geographic locations	a country, a continent, a city or an	COVOC:007	geo	dbpedia				
9	Organism	Organisms (virus or animal) linke	COVOC:008	organism	ncbitaxon				
10	Proteins - Genomes	gene or gene products, including f	COVOC:009	protein_genome	ncit, pr, chebi				

+ ≡ All sheets overview ▾ template\_urls ▾ BiologicalMedical vocabulary (Paola) ▾ Cell lines (Zoe) ▾ Chemicals (Zoe) ▾ Clir ◀ ▶



- **543** terms
- **1625** synonyms
- Mapped and imported from **20** existing ontologies

<https://www.ebi.ac.uk/spot/zooma/>

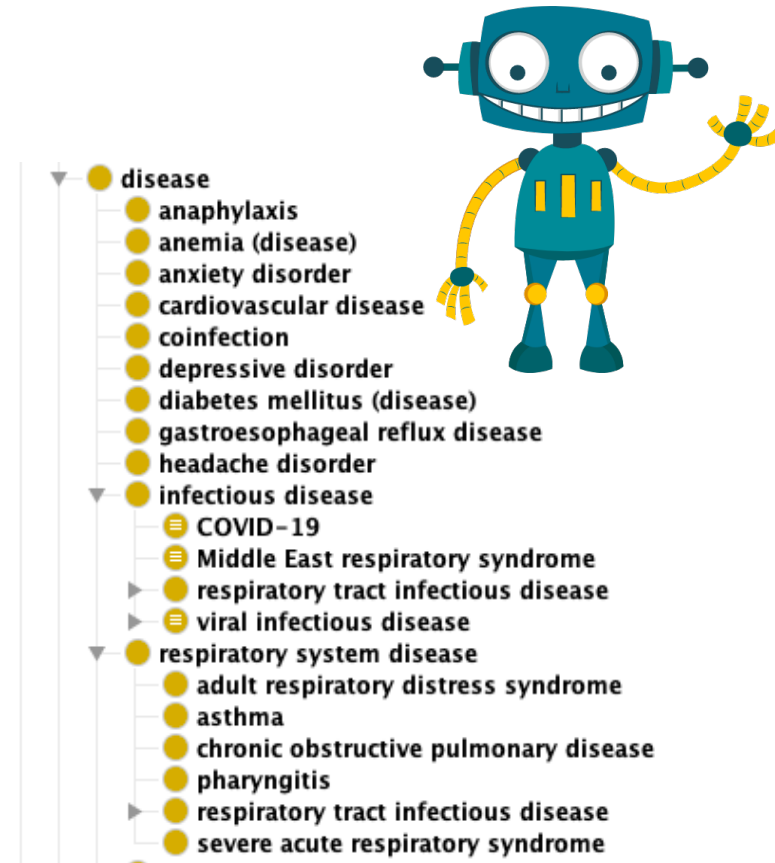
<https://www.ebi.ac.uk/ols/index>

# COVoc - Ontology Sources

Axis	Cross-references with...
Biological/Medical vocabulary	BAO, CHEBI, CHMO, CL, EFO, GO, IDO, MAXO, Mondo, NCBI Taxon, NCIT, OBI, OMIT, PR, UBERON
Cell Lines	CLO, EFO
Chemicals	CHEBI, NCIT
Clinical Trials	No imports
Conceptual entities	EFO, OBI, PATO, CHEBI
Diseases and Syndromes	HP, CHEBI, Mondo, EFO
Geographic Locations	DBpedia, HANCESTRO
Organism	NCBITaxon
Proteins / Genomes	NCIT, PR, CHEBI

# COVoc pipeline – from vocabulary to OWL

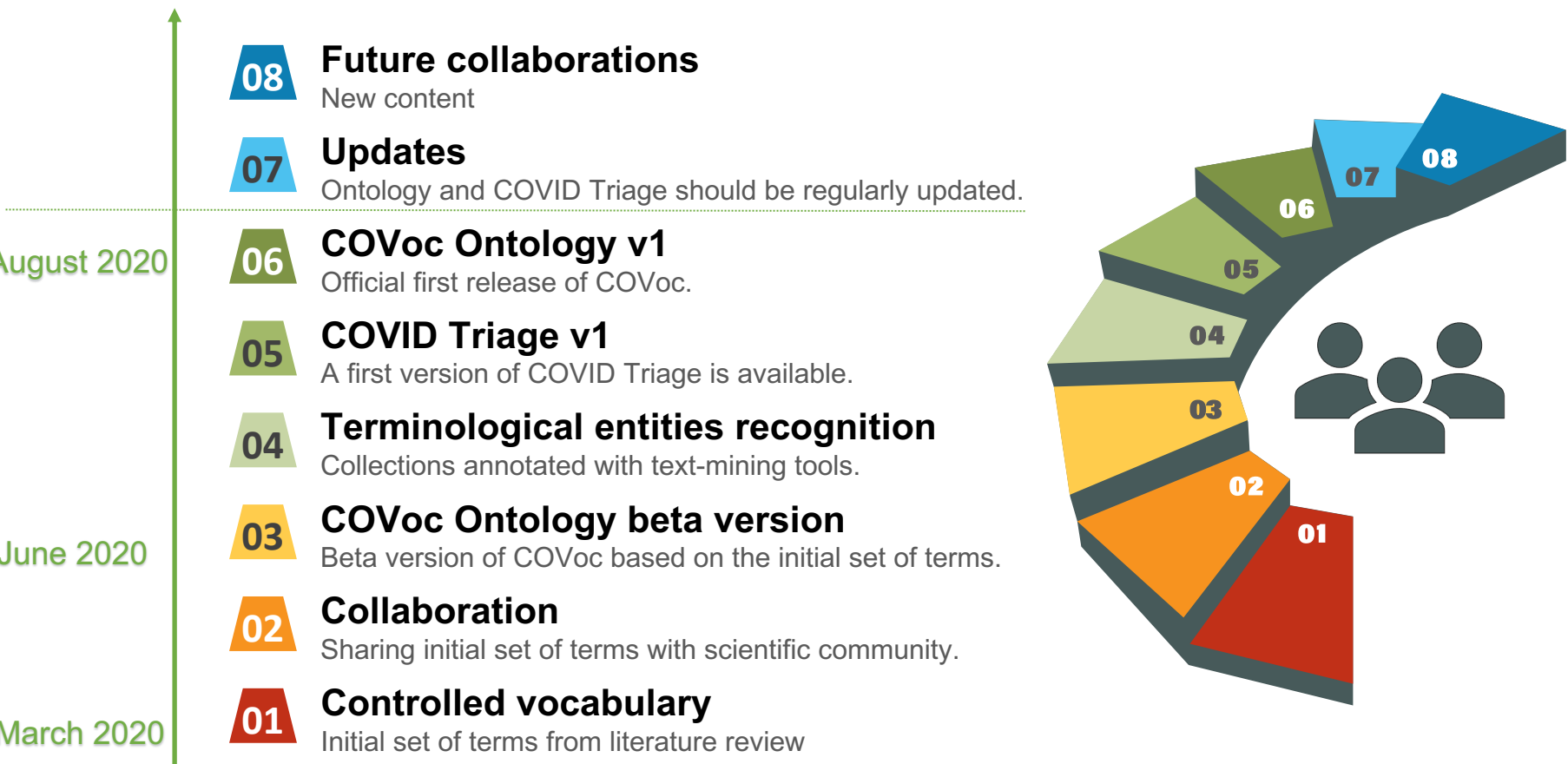
- Manual curation
  - Collecting a list of external ontologies
- Ontology Development Kit
  - Customised pipeline using ROBOT directly working from the original list of terms
  - Creates import files from external ontologies and combines additional information from SIB
  - Set of terms -> OWL files -> COVoc OWL



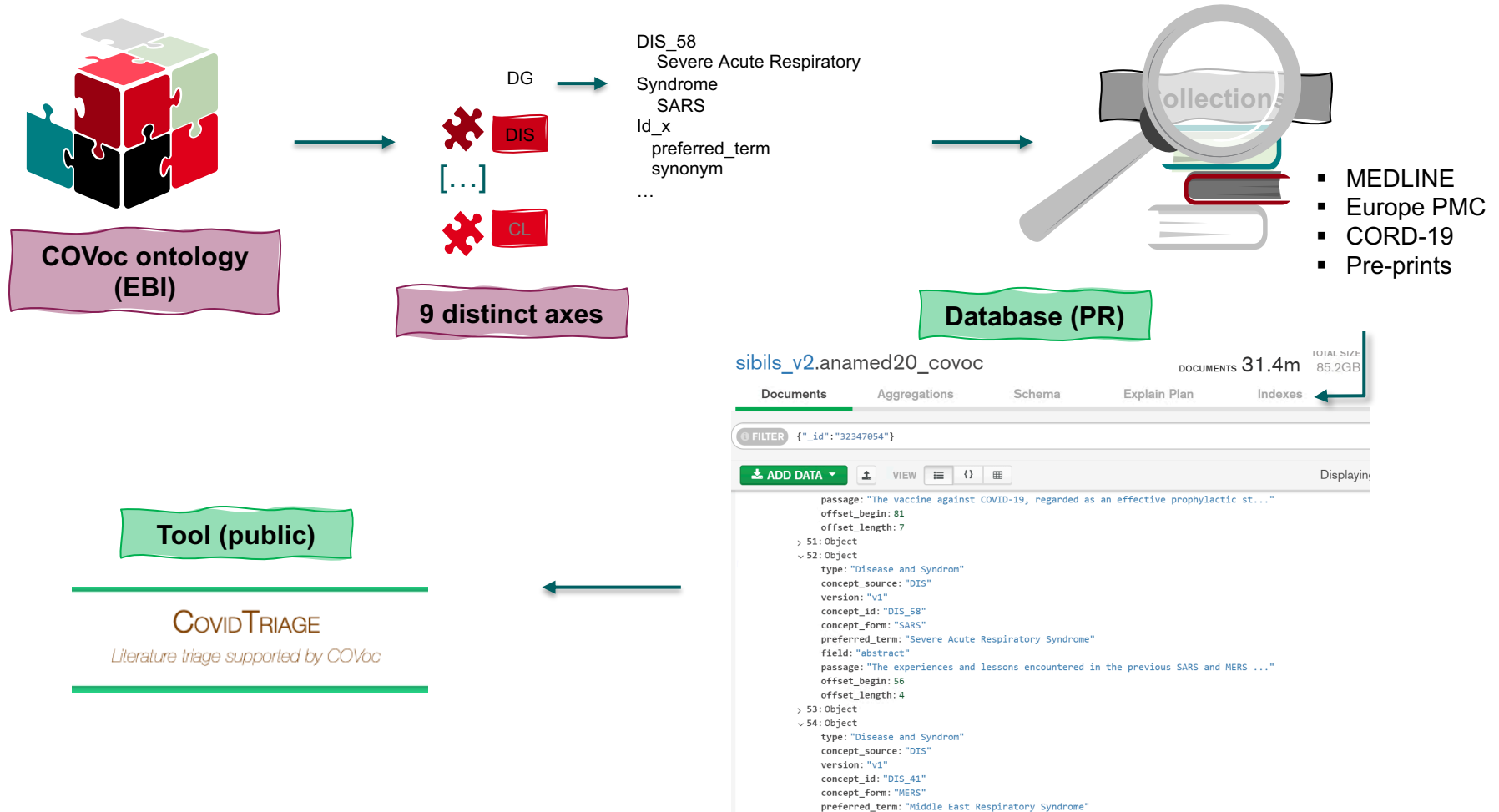
<https://github.com/INCATools/ontology-development-kit>  
<http://robot.obolibrary.org/>



# Timeline






# Terminological entities recognition in diverse sources



# RESULTS – entity recognition using COVoc

August 2020

	MEDLINE	Europe PMC	CORD-19
# total of documents	31 362 156	3 065 406	84 132
# total of annotations	239 319 444	429 474 525	26 871 635
Average per document	8	140	319

	P10	MAP	R_prec
Baseline BL3	0.4275	0.1192	0.1983
COVOC3	0.45  + 5.2%	0.1267  + 6.3%	0.2035  + 2.6%

# Covid Triage Interface

## STEP 1 - QUERY AND ANNOTATION AXIS

Search:  \*

Focus:  \*

What terms contain these categories? Find the thesaurus on [github](#)

### ADVANCED OPTIONS

Source:  \*

Published:  to:

Max nb of publications to retrieve:

Keywords to exclude:  Delimiter: semicolon

## STEP 2 - SELECT A RELEVANT PUBLICATION

Back

Keyword:remdesivir Axis:Drug Date:2000-2020 Source:PubMed Central

### RELEVANT PUBLICATIONS

Publication id	Title	Year	Relevance	Annotations nb	Abstract
<input type="checkbox"/> <a href="#">PMC7457617</a>	Elucidation of remdesivir cytotoxicity pathways through genome-wide CRISPR-Cas9 screening and transc...	2020	23.1	140	<a href="#">[Show]</a>
<input type="checkbox"/> <a href="#">PMC7459246</a>	Remdesivir: First Approval	2020	22.0	165	<a href="#">[Hide]</a>

☐ Show concept links

PMCA viewer v1.2

Drugs - 01-09-2020 - (1):1-9. - PMCID: [PMC7459246](#) - DOI: [10.1007/s40265-020-01378-w](#)

Author(s): Lamb YN<sup>1</sup>.

Affiliation(s):  
1 grid.420067.7 0000 0004 0372 1209 Springer Nature, Private Bag 65901, Mairangi Bay, Auckland, 0754 New Zealand

## Remdesivir : First Approval

### Abstract

The [antiviral agent remdesivir](#) (Veklury®; Gilead Sciences), nucleotide analogue prodrug, has broad-spectrum activity against [viruses](#) from several families. Having demonstrated potent [antiviral activity](#) against [coronaviruses](#) in preclinical studies, [remdesivir](#) emerged as a candidate [drug](#) for the [treatment](#) of the [novel coronavirus disease 2019](#) ( [COVID-19](#) ), caused by [severe acute respiratory syndrome coronavirus 2](#) ( [SARS - Co V -2](#) ) [infection](#) , during the current [global pandemic](#) . Phase III evaluation of [remdesivir](#) in the [treatment](#) of [COVID-19](#) commenced in early 2020 and has thus far yielded promising results. In late May 2020, [Taiwan](#) conditionally approved the use of [remdesivir](#) in [patients](#)

### COVoc Ontology

- ☒ BMV - BioMedical Vocabulary (232)
- ☒ CE - Conceptual Entities (36)
- ☒ CL - Cell Lines (2)
- ☒ DG - Drugs (164)
- ☒ DIS - Diseases&Syndromes (107)
- ☒ LOC - Geographic Locations (18)
- ☒ PG - Proteins&Genomes (4)
- ☒ SP - Organisms (119)

### Other terminologies

☐ A.T.C. (21)

Search:

Drug - Remdesivir DG:DG\_62  
Biological / Medical vocabulary - antiviral BMV:BMV\_6  
Biological / Medical vocabulary - antiviral agent BMV:BMV\_6  
Drug - remdesivir DG:DG\_62  
Biological / Medical vocabulary - viruses BMV:BMV\_163  
Biological / Medical vocabulary - antiviral BMV:BMV\_6  
Biological / Medical vocabulary - antiviral activity BMV:BMV\_6  
Biological / Medical vocabulary - coronaviruses BMV:SP\_5  
Organism - coronaviruses SP:SP\_5

<http://candy.hesge.ch/CovidTriage/>

# Controlled vocabulary – or ontology?



## Controlled Vocabulary for COVID-19

- Historical reasons
  - COVoc originally was a pure vocab
  - EBI ontologized and aligned with wider OBO efforts and EFO

# Why not use CIDO?

- March 2020
  - Development of COVoc began
    - tandem with CIDO
  - CIDO paper published in June 2020
  - COVoc Beta released in June 2020
- Crisis response meant **rapid** development solution
- Very open to aligning our efforts!
- Primary use case for COVoc is literature triage NOT ontological representation

June 2020

03

## **COVoc Ontology beta version**

Beta version of COVoc based on the initial spreadsheet.

02

## **Collaboration**

Sharing spreadsheet with scientific community.

March 2020

01

## **Google Spreadsheet**

File creation and first relevant terms.

# Next

- Ontology maintenance
  - Addition of new terms and synonyms required to further support triage and curation
- Improve interoperability
  - Grow cross references (OBO Foundry)
  - Align efforts of others
- Improve consistency of classification
  - What are the curators' needs?

OLS:

<https://www.ebi.ac.uk/ols/ontologies/covoc>

Access to the ontology:

<https://github.com/EBISPOT/covoc/>

COVoc deployed in Triage tool:

<http://candy.hesge.ch/CovidTriage/>

# Acknowledgements



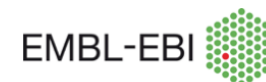
Donat Agosti  
Ruschel Tatiana



Swiss Institute of  
Bioinformatics



Patrick Ruch  
Déborah Caucheteur  
Julien Gobeill  
Pierre-André Michel  
Anaïs Mottaz  
Luc Mottin  
Nona Naderi



Helen Parkinson  
Paola Roncaglia  
Nicolas Matentzoglou  
David Osumi-Sutherland

