Towards a METADATA platform for COVID-19
(contribution to EOSC-GB)
Giorgio Rossi - Italy

Datasets on SARS-CoV-2 and on COVID-19 are archived in different repositories of reference for the scientific community, both public and private.

Three categories should be identified:

Virus Genome
Human Genome
Human Phenotypes and Clinical Data

DATASETS	TYPE	METADATA STANDARDS
SARS-CoV-2 Virus GENOTYPE	Sequences, genomics	ENA (EBI-EMBL): https://www.ebi.ac.uk/ena/pathogens/covid-19 NCBI VIRUS: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=Wuhan%20seafood%20market%20pneumonia%20virus,%20taxid:2697049 GISAID: https://www.gisaid.org/ (EpiCoV$^{\text{TM}}$)
Virus/Protein structure	Structural data	Structural biology databanks: ELIXIR, INSTRUCT, analystical research infrastructures (synchrotrons, neutron, cryo-TEM, NMR…
Human GENOTYPE	Genome Sequences	https://www.ncbi.nlm.nih.gov/gap/ http://geco.deib.polimi.it/genosurf/
Human PHENOTYPE for COVID-19 and CLINICAL DATA	Comorbidities and response to COVID-19, cardiovascular diseases, diabetes, immunodeficiencies, symptoms at admission/longitudinal, treatment, lab tests…	NO STANDARDS, national/single hospital rules, GDPR, private data banks. Here a list, almost impossible to make it compatible/interoperable at dataset level: https://docs.google.com/document/d/12O6h5EcVCb7y3w8vJPEef1Tpjg0x2fmge9c1uhYTlfo/edit A convergence proposal by COVID-19 Host Genetics Initiative: https://docs.google.com/spreadsheets/d/1RXrJIzHKkyB8qx5tHLQjcBioiDAOrQ3odAuqMS3pUUI/edit#gid=1645477253 (available from https://www.covid19hg.org/data-sharing/)
EPIDEMIOLOGICAL data	Statistical data Territorial analysis, hospitals, retired people in structures,	NO STANDARDS

	retired people at home…	

An example of METADATA developed for genomic data is:
http://geco.deib.polimi.it/genosurf/
https://academic.oup.com/database/article/doi/10.1093/database/baz132/5670757

Starting from the lack of organization in clinical data, an attempt to define a minimum standard in the collection of human phenotype data has been defined in the following document: https://docs.google.com/document/d/1eMdzhO5xk-MACxjz-kOUJLP6Jort5KuwoOa_u-aZPHs/edit

Example of research paths requiring access to datasets with interoperability:
Virus-Genome -> Human Genome -> Human Phenotypes

This is currently hindered by the difficulty of the last step, which requires access to Clinical Data, both due to the legal aspects and to the diversity of databases adopted at all levels. It is largely at this level that a standardized metadata set, even only high-level metadata (i.e., catalogue entries with minimal qualifying description of the dataset contents and methodology of collection) could make the difference in orienting the researchers to a one-to-one negotiation with the data owners.

An Italian project "COVIDTwin" addresses "monitoring, support to decision making , identification of health-system risk in epidemic and pandemic emergencies". It includes HPC resources, models of epidemy, behavior of populations