EMBL-EBI COVID-19 Action Plan

The world is facing the worst public health crisis with all its consequences for society since many decades. To combat COVID-19 we need to intensify research efforts enabling the scientific community to understand the biology, epidemiology, transmission and evolution of the virus responsible for the outbreak, SARS-CoV-2. Otherwise the world will lack the ability to respond in informed and effective ways, such as through diagnostics, therapeutics, vaccines and public health measures.

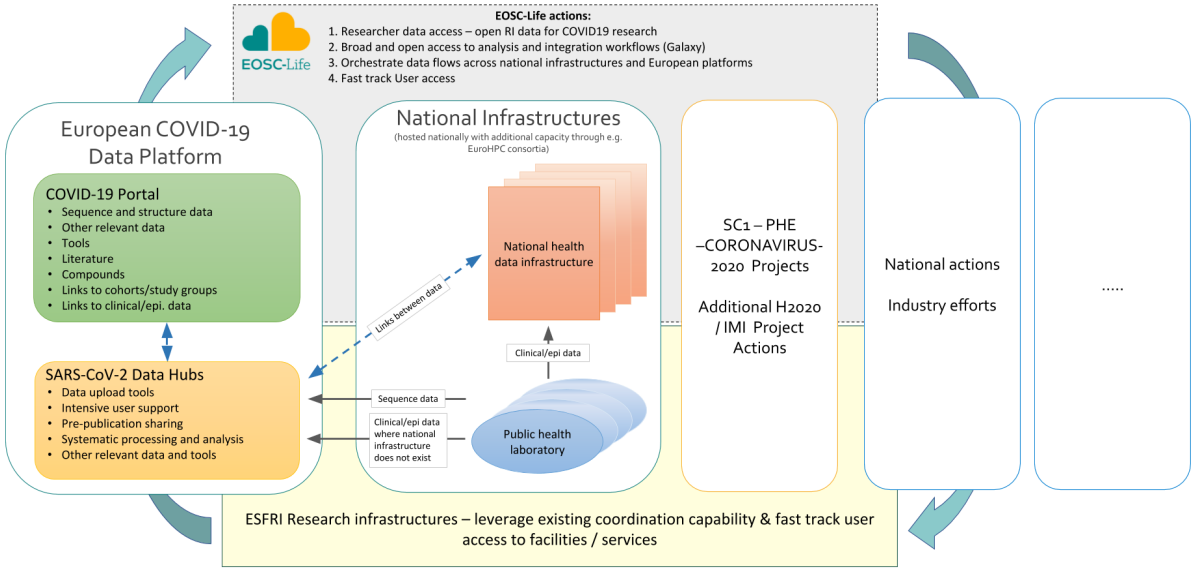
EMBL's European Bioinformatics Institute (EMBL-EBI) has recognised the urgency to identify and consolidate needs around creating a European COVID-19 Data Platform for data/information exchange, connected to the European Open Science Cloud (EOSC)[1].

The goal is to collect and share rapidly available research data from different sources and of different types to enable synergies, cross-fertilisation and use of diverse data sets with different degrees of aggregation, validation and/or completeness so they can be accessed by the research community.

European COVID-19 Data Platform

We envisage the European COVID-19 Data Platform to consist of two major connected components, the SARS-CoV-2 Data Hubs organising the flow of SARS-CoV-2 outbreak sequence data and providing comprehensive open data sharing for the European and global research communities, and one broader COVID-19 Portal (Figure 1).

Figure 1: European COVID-19 Data Platform *Architecture*



[1] https://www.eosc-portal.eu/

To enable a very rapid launch of a working system we plan to assemble existing elements of informatics infrastructure, extend and enhance these elements and leverage initially the key strengths of EMBL-EBI. These strengths are based upon molecular biology data infrastructure and services that EMBL-EBI provides, and its unique position in offering connectivity with national public health data infrastructures, to the EOSC and relevant European Research Infrastructures and research projects, as well as International and National Research organisations. The broad involvement of these and other relevant stakeholders in coordination and governance of the COVID-19 portal is crucial to stimulate the use of this platform by research data providers and users, while ensuring broad, yet where necessary controlled, access.

Core components to be brought into the system as EMBL-EBI contributions include the long-established EMBL-EBI European Nucleotide Archive (ENA)[2], the open sequence database of record and European node of the celebrated International Nucleotide Sequence Database Collaboration (INSDC)[3]; the COMPARE Data Hubs[4], which provide pathogen-focused data sharing and analysis tools that have been designed and implemented over the last five years under the EU COMPARE project; the European Genome-phenome Archive (EGA)[5], a database for controlled access data, such as anonymised clinical and epidemiological data from research subjects, operated collaboratively by EMBL-EBI and the Centre for Genomics Regulation in Barcelona; Embassy, a cloud compute facility provided by EMBL-EBI; and the Pathogen Portal[6], a web site and set of programmatic interfaces that will be used to create (one of the) access points to SARS-CoV-2 data in the platform.

Concept of the SARS-CoV-2 Data Hubs

Sequence data form an essential foundation for broad and full investigation of infectious disease outbreaks. In the COVID-19 response, without sequence data, the ability of the scientific community efficiently to understand the biology, epidemiology, transmission and evolution of the virus responsible for the outbreak, SARS-CoV-2, will be severely limited and the world will lack the ability to respond in informed and effective ways, such as through diagnostics, therapeutics, vaccines and public health measures.

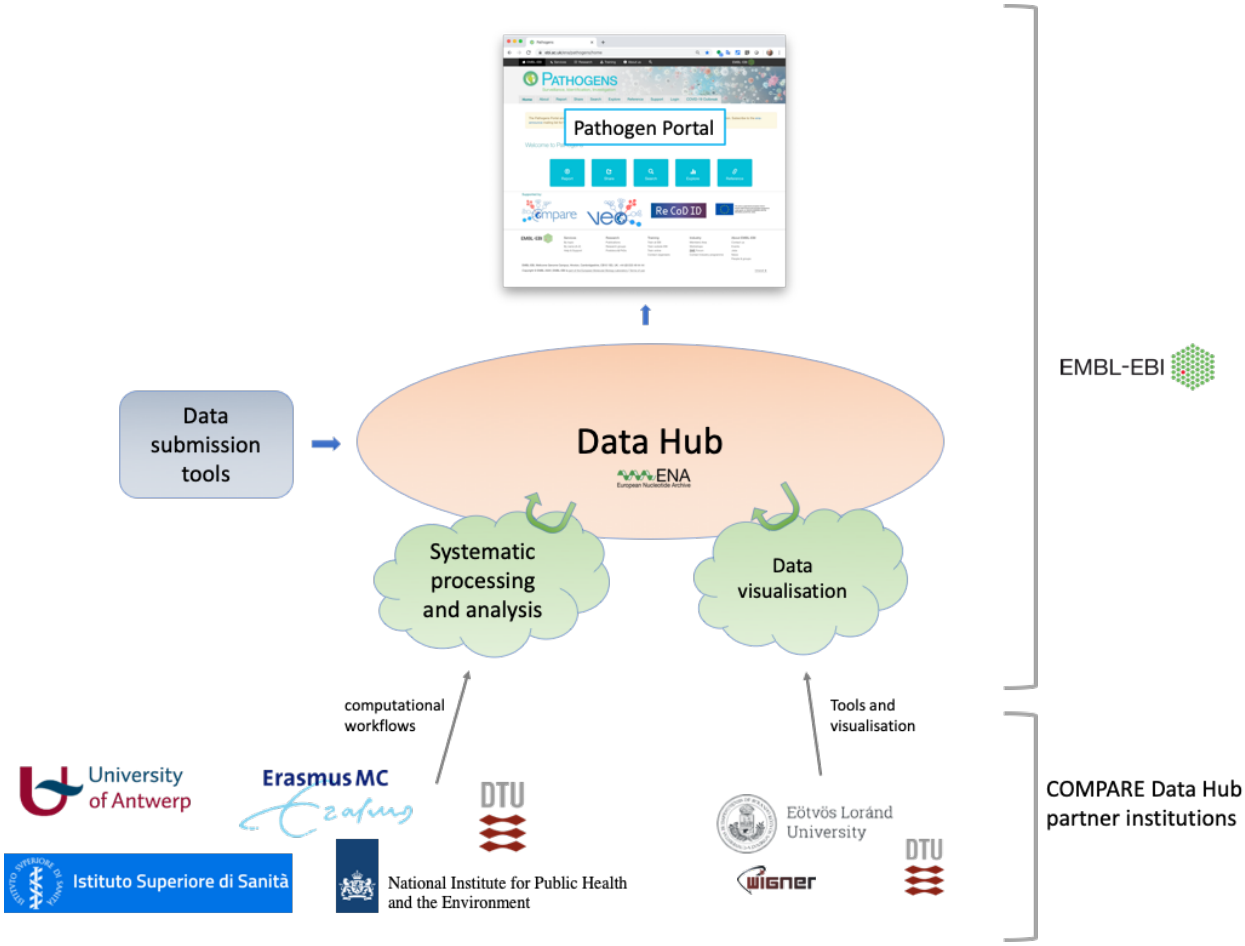
The SARS-CoV-2 Data Hubs will form one of the two components of the European COVID-19 Data Platform. These will be built upon the foundations of existing EMBL-EBI infrastructure, such as ENA and its services and those elements put in place as an extension to this infrastructure under the EU COMPARE project, known as the “COMPARE Data Hubs”.

The current COMPARE Data Hubs provide a system for the sharing and analysis of pathogen sequence data across pathogen species and across the clinical, animal health, food and water and environmental domains (Figure 2). Each Data Hub is established by a group of collaborators and configured appropriately for their requirements. Configuration includes the selection of the most appropriate data submission tools in addition to a variety

[2] https://www.ebi.ac.uk/ena/browser/home
[3] http://www.insdc.org/
[4] http://europepmc.org/article/MED/31868882
[5] https://ega-archive.org/
[6] https://www.ebi.ac.uk/ena/pathogens/home

of data analysis and visualisation options. These are offered based on technical and support elements provided by the partner institutions. Users are supported in their use of their Data Hub through the establishment of standards and best practices. Technical elements include a number of computational analysis workflows, the Jupyter Notebook system for data exploration and visualisation, and the Evergreen phylogenetic tree-building package. While all data are ultimately released fully and public through ENA, data owners are given the option to restrict access to those with whom they are collaborating around a given Data Hub. At the time of writing there are 13 COMPARE Data Hubs used across collaborations addressing a breadth of issues, from eukaryotic parasites, through bacterial antimicrobial resistance to sequencing methods improvement.

Figure 2: Architecture of a current COMPARE Data Hub, showing EMBL-EBI components (hardware, operation, user support and coordination) and technical elements provided from partner institutions (computational workflows, tools and visualisation).



We will offer SARS-CoV-2 Data Hubs to those public health agencies and other scientific groups responsible for generating sequence from the virus at national or regional levels. It is expected that there will ultimately be numerous SARS-CoV-2 Data Hubs with similar configurations in terms of analysis, but different preferred submission tools.

While the focus of the SARS-CoV-2 Data Hubs will be sequence data, these will be highly contextualised. Essential metadata, such as sampling tracking identifiers, sampling time,

geographical location, method of sampling, health status of host and sequencing platform/strategy, will be captured alongside sequence data into ENA and the Data Hubs. Since these elements of metadata are highly anonymised they can be shared without the need for controlled access. Alone, these metadata will enable a great many uses of the data, such as geospatial analysis, evolutionary studies, transmission and hotspot investigations.

The SARS-CoV-2 Data Hubs will also provide high connectivity between sequences, essential metadata and deeper clinical and epidemiological data, such as classification of symptoms, time since infection, comorbidities, treatment history and travel/contact history. Such data are however not frequently shareable across national borders. Clinical and epidemiological data are stored within the data management systems of the healthcare systems of individual nations, and usually require controlled access due to the potentially identifiable patient information embedded in them. Important for deeper scientific investigations, such as into the relationship between viral genetic variation and disease severity, is that the SARS-CoV-2 DataHub will be able to connect the relevant clinical/epidemiological data. The sample tracking identifiers captured as part of the essential metadata alongside viral sequence will provide the mechanisms for this. Where appropriate, we will also invoke the EGA archive for clinical/epidemiological data in cases where national health infrastructures request this and national legislation permits it.

The SARS-CoV-2 Data Hubs will provide systematic data processing and analysis based on such workflows as Jovian[7] (developed by the National Institute for Public Health and the Environment (RIVM)[8] and the Erasmus Medical Centre (Erasmus MC)[9] in the Netherlands), visualisation (curated and operated by Eötvös University[10] in Hungary[11]) and phylogenetic analysis tools from the Danish Technical University (DTU)[12]. While EMBL-EBI provides almost all of the physical infrastructure that serves the data hubs, our partner institutions are involved in the operation and improvement of the system through access to tenancies in our Embassy cloud compute facility. Specific roles will be:

- EMBL-EBI: provision of ENA and its services including all data management and user support and the Pathogen Portal and its programmatic interfaces
- RIVM and Erasmus MC: development, improvement of the computational workflow that provides systematic processing and analysis
- WIGNER/Eötvös University: curation and operation of the data visualisation “Notebooks” that appear to users in the Pathogen Portal
- DTU: development and operation of tree-building visualisation software

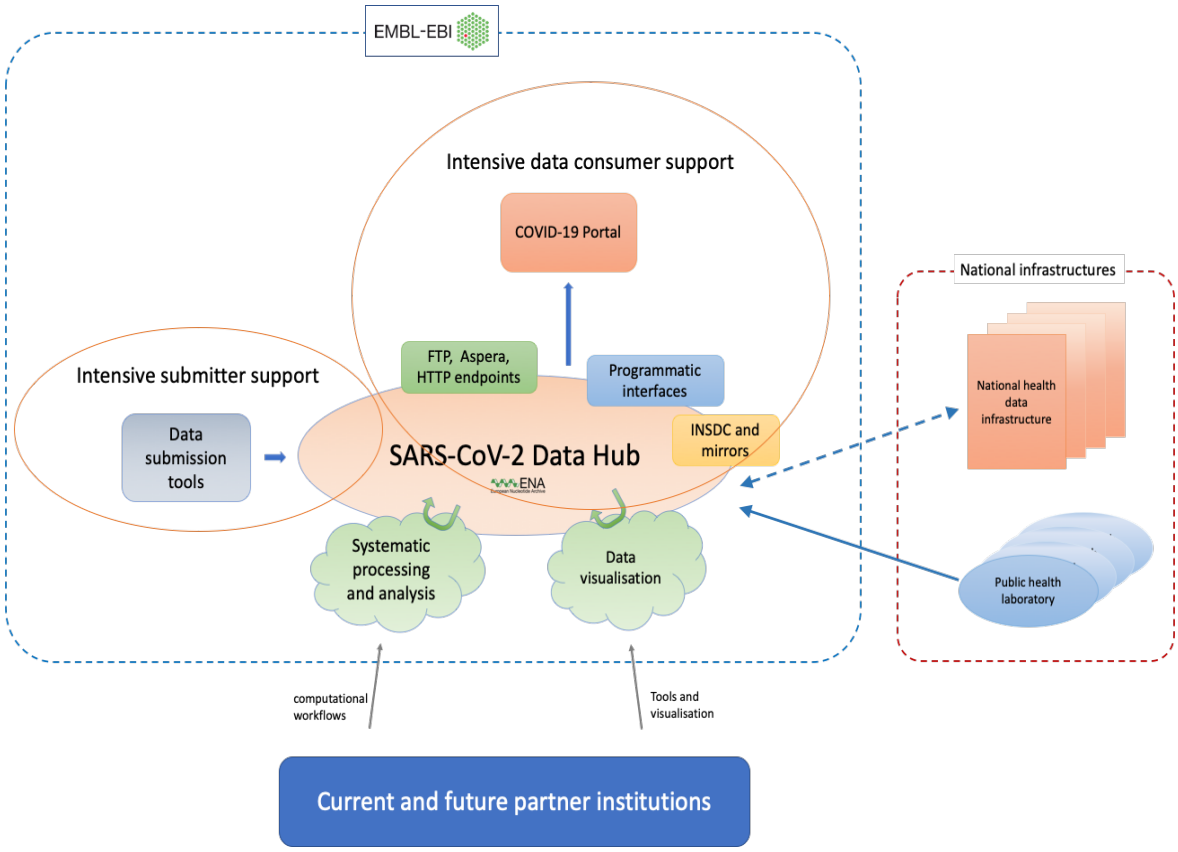
During the lifetime of the project we expect to add new partners with specific expertise to the SARS-CoV-2 Data Hubs development consortium.

The SARS-CoV-2 Data Hubs will support comprehensive sequence data across all platforms/sequencing strategies and along the data life cycle (Figure 3). Raw data, in the

[7] https://github.com/DennisSchmitz/Jovian
[8] https://www.rivm.nl/en
[9] https://www.erasmusmc.nl/en
[10] https://www.elte.hu/en/
[11] The COMPARE partner group from the WIGNER institute, has moved to Eötvös University since the end of the COMPARE project.
[12] https://www.dtu.dk/english

form of sequence “reads”, will be the primary data input into the system for many public health operations. We will operate high-throughput systematic computational processing and analysis of raw data to make available consensus/assembled sequence. We will support consensus/assembled sequences provided both from our computations and generated directly by data providers, in cases where capacity, interest and expertise exist.

Figure 3: Architecture of a SARS-CoV-2 Data Hub, showing EMBL-EBI components (hardware, operation, user support and coordination) and technical elements provided from partner institutions (computational workflows, tools and visualisation)



We will mount an intensive user support operation to reduce barriers to sharing of data through the SARS-CoV-2 Data Hubs. This support will focus on data submissions, for which we will be offering our portfolio of existing submission tools and interfaces as appropriate for the different local contexts, e.g. bioinformatics expertise, volume of data to be transferred, that our data providers will be experiencing. Because the data providers are, by definition, those who are at the heart of testing and response to COVID-19, they have many other priorities and we must reduce data flow friction as much as possible. Support work will include help desk, training, adaptation of tools and services better to suit emerging requirements, software development to build scripts and plug-ins to allow public health laboratories more easily to connect their existing systems to the Data Hubs.

We are currently communicating with EMBL Member States to initiate the mobilisation of data across Europe. Following our first email in mid-March, seven nations have responded thus far, and we have had discussions with two, with many more calls scheduled in the next few days. Of the two, one nation - the United Kingdom - is pushing forward on large-scale

sequence sharing through a national SARS-CoV-2 Data Hub of data from ten different centres within the nation. We will continue to communicate with EMBL and EU Member States, through our many networks, including ELIXIR and, in collaboration with our partners in the INSDC, open our system globally.

While the SARS-CoV-2 Data Hubs will support all forms of sequence data, we note that the GISAID system[13], a controlled access data sharing system set up to address the needs of the WHO's Preparedness for Influenza of Pandemic potential (PIP) Framework, is being used in addition to INSDC - at this point seemingly effectively - for the sharing of consensus/assembled sequences from the outbreak. We will monitor this situation and be ready, if it becomes appropriate, to provide data into the GISAID system as an additional route of dissemination.

SARS-CoV-2 DataHub Plan

1. **Mobilise data** (1.3.20 -- 28.2.22): we will communicate extensively with EU member states, EMBL Member States and other data providers, to establish the flow of raw and consensus/assembled sequence data into ENA; we will adapt and extend existing submission tools and interfaces, providing deep technical support; we will ensure prioritisation of SARS-CoV-19 data in the INSDC multilateral global data exchange;

2. **Mobilise analysis** (23.3.20 -- 28.2.22): we will deploy the full data processing and visualisation components of the COMPARE data hub system (VEO); we will enable and support users of our web and programmatic interfaces to promote external computational analytical systems and projects;

3. **Connect clinical/epidemiological data** (1.3.20 -- 28.2.22): we will integrate top-level cohort/study group descriptions and links to clinical/epidemiological data to allow discovery of data emerging from the 17 SC1-PHE-CORONAVIRUS-2020 projects; we will maintain links to where clinical/epidemiological data are available;

4. **Enhance access** (1.4.20 -- 28.2.22): we will extend and enhance the access points for data in the system, providing tools and support, for example, for automated synchronisation with all data in the platform into external computational facilities; we will work with our many networks to enable data flow from the system and the connection of new third party tools and interfaces.

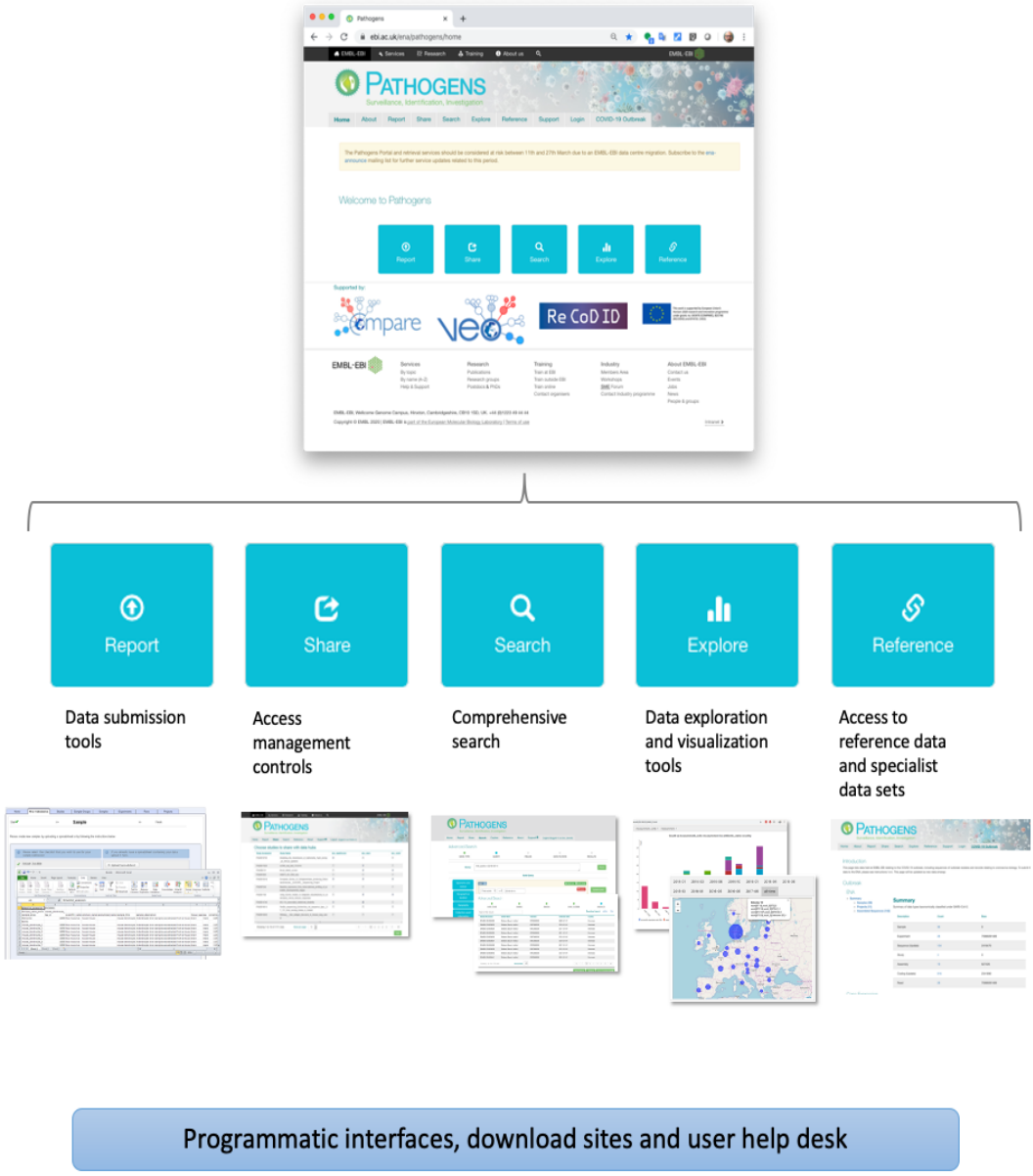
Concept of the COVID-19 Portal

A multitude of research efforts worldwide have been initiated or are shifting their focus towards COVID-19 related activities. Many research outputs lead to datasets submitted to EMBL-EBI and other major centres for the deposition of biomedical data, or in the scientific literature with prePrints being used increasingly for rapid communication.

The public databases run by EMBL-EBI already contain many COVID-19 datasets, and these are open to researchers in academia and industry all over the world. Specifically,

[13] https://www.gisaid.org/

Figure 4: The existing EMBL-EBI Pathogen Portal (https://www.ebi.ac.uk/ena/pathogens/home), showing a variety of its functions, available as web site and through comprehensive programmatic tools



EMBL-EBI's Pathogen Portal (https://www.ebi.ac.uk/ena/pathogens/home) provides access to genes, protein structures, Electron Microscopy data and scientific publications relating to COVID-19 (https://www.ebi.ac.uk/ena/pathogens/covid-19).

EMBL-EBI plans to launch a COVID-19 Portal based on the existing Pathogen Portal to provide the primary entry point into the functions of the European COVID-19 Data Platform and the data and tools that it makes available (Figure 4). The Portal's primary functions are data upload, access management control for those using pre-publication sharing, powerful search, data exploration and access to reference data and specialist data sets, such as

outbreak sequence data from COVID-19 (available directly from https://www.ebi.ac.uk/ena/pathogens/covid-19) and the “Cohort Browser” to allow search of clinical and epidemiological data. Accompanied by comprehensive programmatic interfaces (RESTful and command-line), all functions are available in addition for programmatic use. The COVID-19 Portal will serve both the sequence data sharing functions of the European COVID-19 Data Platform and the comprehensive presentation of all SARS-CoV-2 resources.

To rapidly populate the COVID-19 Portal (Figure 5) and make it immediately useful we will bring together in a first step all relevant datasets from EMBL-EBI data resources such as ENA, UniProt, PDBe , EMDB, Expression Atlas and EuropePMC on genes, proteins, structures, Electron Microscopy data and scientific publications relating to COVID-19. This will be continually enriched as new data emerge from the established data submission into the EMBL-EBI deposition databases.

The second step will enrich the COVID-19 Portal with datasets and tools from EU projects, in which EMBL-EBI resources like Europe PMC, ChEMBL, ENA, EGA and others are partners and COVID-19 related datasets and literature are already being produced. These will include (aligned EU projects shown in *italics*):

- deeply mined COVID-19 related literature, from Europe PMC;
- development work on this will include COVID-19 specific text mining as an EMBL-EBI contribution to the OpenAIRE initiative (*OpenAIRE*);
- compound screening and assay data relating to ongoing COVID-19 related work (*EUbOPEN* and *eTRANSAFE/NexGETS*);
- chemoinformatics tools that will assist with the integration of compound-related COVID-19 data (*EU-ToxRisk*, *EUbOPEN* and *TransQST*);
- access to tools and interfaces (such as metadata validation and discovery tools) relating to clinical and epidemiological data (*CINECA*);
- “connected” data hubs linking host and viral sequence and other data (*RECODID*);
- support for search and linking of extended data types in the data hubs, to such biological data types as host serology and immunoprofiling and to such non-biological data types such as social media, travel and trade data (*VEO*);
- launch the Cohort Browser (a tool programmed for the RECODID project) to allow top-level discovery and navigation of existing study groups and cohorts, such as from the 17 funded projects of the SC1-PHE-CORONAVIRUS-2020 “Advancing knowledge for the clinical and public health response to the 2019-nCoV epidemic” call. While not directly providing the clinical/epidemiological data, this will allow users to find where data exists and link to them.

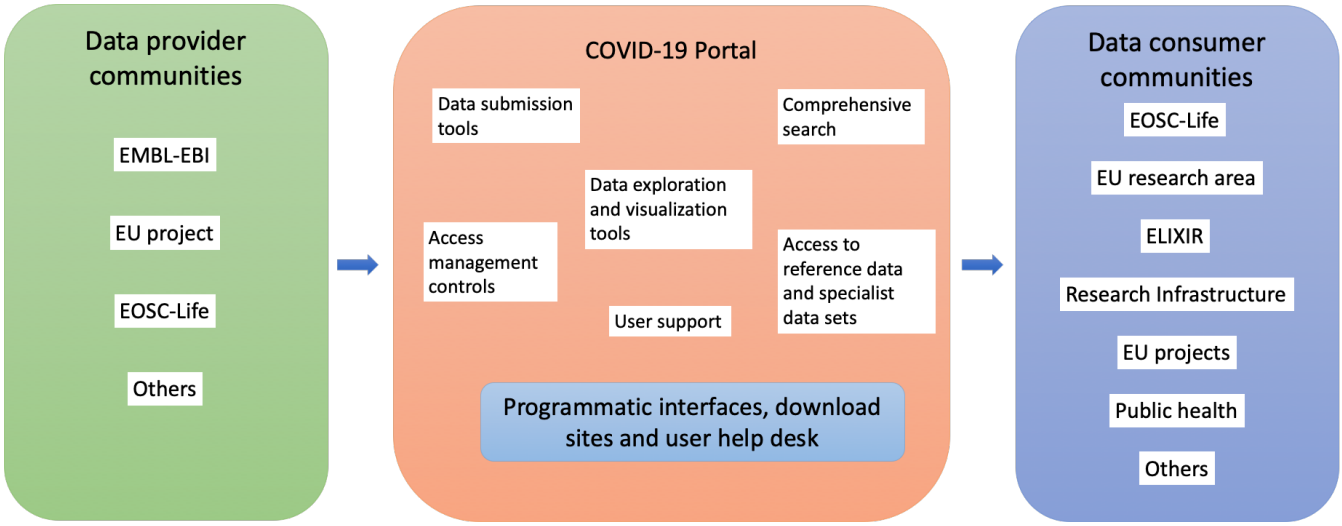
The third step is reaching out to other EU projects in which EMBL-EBI participates to explore the possibility to refocus on COVID-19 related research and to bring these research outputs into the COVID-19 Portal.

The fourth step will be driven by ELIXIR. ELIXIR intends to catalogue the current open datasets, workflows and protocols relevant for COVID-19 and the many efforts in ELIXIR nodes and in other relevant Research Infrastructures as well as across EOSC-life. We plan to cross-link between the ELIXIR catalogue and COVID-19 Portal and envisage ELIXIR and EOSC-life prioritising out of this ELIXIR catalogue the relevant datasets, workflows and protocols to be brought directly into the COVID-19 Portal. We will here also rely on the

ELIXIR community to connect to the open COVID-19 data resources, for example through the European Galaxy instances.

The fifth step will depend on the input from additional European stakeholders and on leveraging our international connections with major bioinformatics data and service providers in the USA and Asia. This effort will be in close coordination with ELIXIR.

Figure 5: The future COVID-19 Portal showing data sources and a variety of its functions



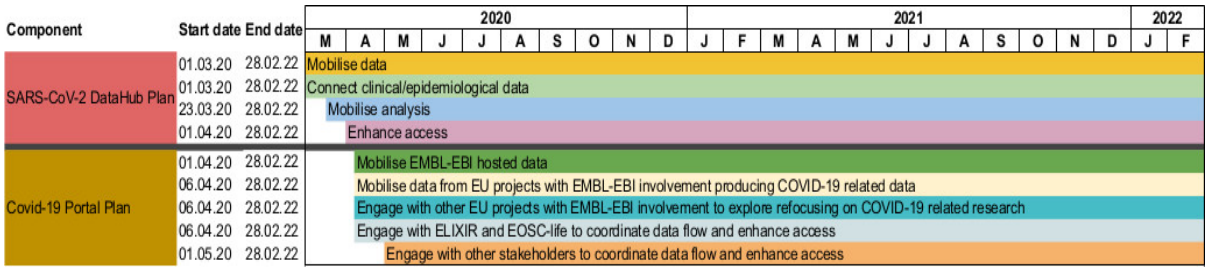
COVID-19 Portal Plan

1. **Mobilise EMBL-EBI hosted data** (1.4.20 - end of project): launch the COVID-19 Portal with all relevant datasets from EMBL-EBI data resources;
2. **Mobilise data from EU projects with EMBL-EBI involvement producing COVID-19 related data** (6.4.20 - end of project);
3. **Engage with other EU projects with EMBL-EBI involvement to explore refocusing on COVID-19 related research** (6.4.20 - end of project);
4. **Engage with ELIXIR and EOSC-life to coordinate data flow and enhance access** (6.4.20 - end of project): We will cross-link between the ELIXIR catalogue and COVID-19 Portal and envisage ELIXIR and EOSC-life prioritising the relevant datasets, workflows and protocols. We will in coordination with ELIXIR and EOSC-life priorities extend and enhance the access points for data in the system, providing tools and support, for example, for automated synchronisation with all data in the COVID-19 Portal with all relevant datasets from EMBL-EBI data resources into external computational facilities;
5. **Engage with other stakeholders to coordinate data flow and enhance access** (1.5.20 - end of project): we will work with our many networks to enable data flow from the system and the connection of new third party tools and interfaces, especially from additional European stakeholders and from the major bioinformatics data and service providers in the USA and Asia.

European COVID-19 Data Platform Timelines

EMBL-EBI started already to work on the European COVID-19 Data Platform and the project is expected to run for two years; longer if required. All activities will be started as soon as possible but in a slightly staggered mode, and are expected to run until the end of the project, as shown in Figure 6.

Figure 6: Timeline for the European COVID-19 Data Platform



EMBL-EBI Contact

Dr Rolf Apweiler
Director, EMBL-EBI
contact emails:
apweiler@ebi.ac.uk
contact@virusresource.embl.org