EOSC-Life: Building a digital space for the life sciences

D13.1 — Enhanced COVID-19 Portal and data mobilisation reporting

WP13 – Extension of COVID-19 Data Portal
Lead Beneficiary: EMBL-EBI
WP leader: Guy Cochrane
Contributing partner(s): EMBL-EBI

Authors of this deliverable: **Guy Cochrane, Marianna Ventouratou**

Contractual delivery date: **31 August 2021**
Actual delivery date: **3 November 2021**
H2020-INFRAEOSC-2018-2

Grant agreement no. 824087
Horizon 2020
Type of action: RIA

Table of Contents

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

Executive Summary

The overall objective of Work Package 13 is to deliver essential open data infrastructure to support urgent scientific research in the face of the COVID-19 pandemic. Including both the extension of existing infrastructure through software development and the operationalisation of this extended infrastructure through deep support services, the aim is to enable and empower a substantial research effort across European and global research communities. More specifically the objectives of WP13 are to:

a. Mobilise open biomolecular data relevant to COVID-19 research from structured bioinformatics databases and prioritised life science RI resources via the COVID-19 Data Portal
b. Mobilise new SARS-CoV-2 biomolecular data emerging from research laboratories, national public health organisations, EU, and global initiatives through the SARS-CoV-2 Data Hubs to the COVID-19 Data Portal
c. Connect biomolecular data to sensitive contextual clinical and epidemiological data that lie within dispersed controlled access repositories, presenting these through the COVID-19 Data Portal

Progress on objectives: 1 and 2 are progressing well objective 3 is underway and we move forward, although data protection-related issues have presented challenges.

- Objective 1: over 5 million records from across EMBL-EBI resources are displayed on the portal, with many options for search and retrieval
- Objective 2: consensus sequence data from 101 countries, and raw sequencing read data coming in from 79 countries[1]. All public raw sequencing read data being analysed with custom workflows through the data hubs system
- Objective 3: a preliminary cohort browser application (developed under ReCoDID) has been adapted and made available for the purposes of the European COVID-19 Data Platform to allow browsing linked biomolecular and clin-epi data, based on simulated data. While the legal framework for harmonisation and linking of sensitive data continues to be developed, with a corresponding lack of data to show, this functionality is held ready for deployment in the COVID-19 Data Portal

Deliverable 13.1 set out for the COVID-19 Data Portal to be operating systematic data harvesting systems and tools to provide statistics from comprehensive EMBL-EBI data resources and tracking of growing data sets, offering tools and programmatic access having undergone User Experience testing.

[1] https://www.covid19dataportal.org/statistics

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

Project Objectives

With this deliverable, the project has contributed to the following objectives under Work Package 13:

a. Mobilise open biomolecular data relevant to COVID-19 research from structured bioinformatics databases and prioritised life science RI resources via the COVID-19 Data Portal[2]
b. Mobilise new SARS-CoV-2 biomolecular data emerging from research laboratories, national public health organisations, EU, and global initiatives through the SARS-CoV-2 Data Hubs to the COVID-19 Data Portal
c. Connect biomolecular data to sensitive contextual clinical and epidemiological data that lie within dispersed controlled access repositories, presenting these through the COVID-19 Data Portal

All of the work undertaken in this deliverable is contributing to all three EOSC-Life's main objectives, namely:

- Objective 1: Establish EOSC-Life by publishing FAIR life science data resources for cloud use
- Objective 2: Create an ecosystem of innovative life-science tools in EOSC
- Objective 3: Enable ground-breaking data driven research in Europe by connecting life scientists to EOSC

Detailed Report on the Deliverable

1. Detailed description of work carried out during reporting period

1.1. Integration of data via a COVID-19 Data Portal

Under 13.1 Task 1, EMBL-EBI set out to populate the COVID-19 Data Portal that integrates and presents relevant biomolecular data and literature. At the time of writing (September 2021) the following progress has been made:

- over 5 million records indexed and presented in the Portal
- cross referencing between the resources is supported and presented, creating a rich, navigable network of information
- fast turnover: when data goes public in its relevant database, it is indexed and displayed on the portal rapidly through daily indexing schemes

Data is sourced dynamically from EMBL-EBI's 54+ specialist databases, of which 21 currently provide data to the Portal, that span the EC's recommended deposition databases for COVID-19 and secondary databases deriving content from the deposition databases. This is achieved

[2] https://www.covid19dataportal.org/

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

through EMBL-EBI's centralised indexing service - EBI Search and provides endpoints to query metadata from all EMBL-EBI services, with many resources creating COVID-19-specific services. The indexes are updated nightly.

COVID-19 data are already richly represented and are rapidly growing across EMBL-EBI data resources, such as ENA, UniProt, PDBe, EMDB, Expression Atlas and EuropePMC covering genes, proteins, structures, electron microscopy data and scientific publications relating to COVID-19.

The Portal will be continually enriched as new data emerge from data submissions into EMBL-EBI deposition databases and furnished with web and programmatic interfaces appropriate to support diverse and extensive research upon the data served. The COVID-19 Data Portal displays the following useful features to users:

- Improved search queries - search box can support complex queries in Apache Lucene syntax (mirroring EBI Search)[3]
- API to query any/all metadata held in the Portal[4]
- Bulk downloader tool to retrieve sequencing data in various formats[5]

1.2. Mobilise sequence data through the SARS-CoV-2 Data Hubs

To enable the mobilisation of sequence data through the SARS-Cov2-Data Hubs and into the COVID-19 Data Portal in order to serve our extensive and diverse user base of research laboratories, national public health organisation and international bodies, we set out in this deliverable to:

1. develop the necessary software and adapt the submission interfaces,
2. provide support to embed programmatic submission services in existing laboratory informatics systems,
3. provide data standards validation and compliance tools, as well as
4. training and assistance for users of web interfaces and support in deploying and accessing analytical workflows within the Data Hubs compute environment.

The main output within this deliverable is that we connected these tools and data from the Data Hubs to the COVID-19 Data Portal.

More specifically we achieved the following to enhance the mobilisation of data from the Data Hubs to the Data Portal:

- Deployed JSON-based submission tool for SARS-CoV-2 submissions, simplifying the validation and submission of consensus sequences
- Developed support for transformation of data already submitted to the GISAID system
- Assigned three data hubs, of which two are linked to national sequencing efforts and one links datasets together
- Integrated analytical workflows for reference-based mapping (developed under VEO). For Illumina datasets, there are unfiltered VCFs, filtered VCFs and consensus sequences
- Archival[6]

[3] https://www.ebi.ac.uk/ebisearch/documentation.ebi#query_syntax
[4] https://www.covid19dataportal.org/api-documentation
[5] https://www.covid19dataportal.org/bulk-downloads

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

- EVA variant accessioning
- Developed a CoVEO variant visualisation application (under VEO)
- Incorporated WHO and Pango lineages
- Maintained COVID-19 phylogeny

Archival

The archived project, mentioned above, includes an umbrella project, along with three child projects. These correlate to the three main pipeline outputs generated from the processing of the raw read data. First of all, non-filtered variant calls are accompanied by additional pipeline output files, such as a coverage file and BAM file of mapped reads. The other two child projects refer to filtered variant calls, outlined following a threshold (e.g. allele frequency of 0.25), and then a consensus sequence, calculated using the filtered variant calls. This structure supports findability for users and downstream services, and also interoperability between other resources. This has been demonstrated by variant accessioning at the European Variation Archive (EVA), providing stable rsIDs for variants identified from the systematic analysis. Work is ongoing to feed these to display in the portal.

Lineages and Phylogeny

The portal includes Pango and WHO lineage assignments for all submitted sequences[7]. Integrating the pangolin workflow and running this on a daily basis on all submitted sequences enables for the assignments to be up-to-date.

The COVID-19 Phylogeny[8] still presents phylogenetic trees on submitted SARS-CoV-2 sequences. Sequences have been split by WHO regions, equating to continental world regions. The phylogeny is currently planned for major changes to ensure greater usability.

1.3. Connect Life Science RI data sets to COVID-19 Data Portal

The Life Science RI facilities and repositories are accumulating COVID-19 datasets from the support to European COVID-19 research projects and data cataloguing efforts. Biomolecular data from the RI repositories and data management solutions are interfaced, via the major biomolecular archives into the COVID-19 Data Portal.

The focus in this first phase was on data types that were already partly integrated into the COVID-19 Portal - thereby closing the loop between RI data management practice and the provision of integrated COVID-19 data to the research community.

The following resources were interfaced:

- BioImage Archive: allowing data routed from IDR and Omero to be available through the Portal.
- BBMRI-ERIC: BBMRI Catalogue (Human samples/European biobank data) improving the connectivity between BBMRI Catalogue and Biosamples for the COVID-19 Data Portal.

[6] https://www.ebi.ac.uk/ena/browser/view/PRJEB45555?show=component-projects
[7] https://www.covid19dataportal.org/sequences
[8] https://www.covid19dataportal.org/phylogeny-tree

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

- CSIC: PDB+EMDB (Structural Biology data/Structure based drug design data). 3DBionotes and PDBe Knowledge Base to improve the workflow for cryo-EM data annotations and cryo-EM maps and deposit these in the cryo-EM archive for publication via the COVID-19 Data Portal.
- Fraunhofer: ChEMBL, allowing data (small molecule screening/drug repurposing data to be routed from EUOSECDB to be published via the Portal.

1.4. Connect biomolecular data to external clinical/epidemiological data sets

The SARS-CoV-2 Data Hubs enable high connectivity between locally held data (sequences and essential non-sensitive metadata) and deeper clinical and epidemiological data, such as classification of symptoms, time since infection, comorbidities, treatment history and travel/contact history.

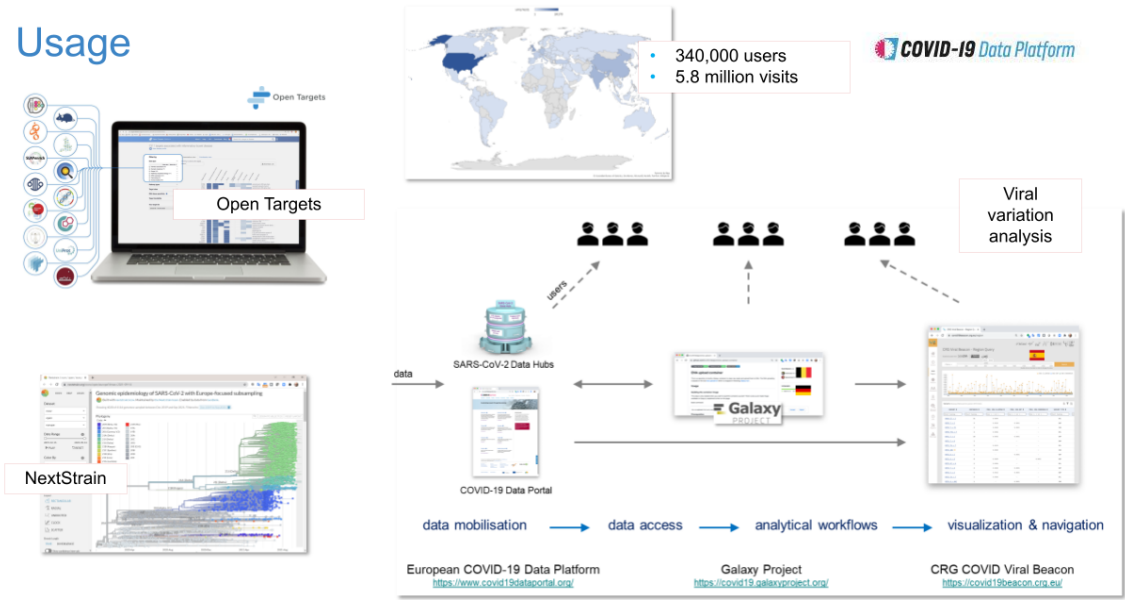
At the time of writing this deliverable the cohort browser is connected and ready to be deployed, but we are still lacking data due to ongoing legal barriers to sharing and linking of sensitive data. A framework and a roadmap have been put in place by partners within the ReCoDID project with GDPR expertise and the data will be added to the cohort browser in due course.

Within the ReCoDID project, the core legal challenges revolve around the appropriate assignment of GDPR data controller vs processor roles throughout the entire workflow, the status of international organisations in the context of GDPR, the possibility of derogation due to public interest according to Article 49 GDPR, and varying interpretation of GDPR depending on national guidance and supervisory authorities. These points must be taken into consideration when designing the workflow, formulating a roadmap and putting a range of necessary documents for the legal framework in place. This is done collaboratively within a multidisciplinary team consisting of specialised legal experts, scientists, clinicians, bioinformaticians, epidemiologists as well as local DPOs and legal teams of the partner institutions. Consequently, it is a time-consuming effort to establish the legal framework that needs to be implemented prior to submitting sensitive patient data into the system, resulting in a significant impact on the deployment of a bespoke and fully functional cohort browser.

2. Impact

The data made available through the COVID-19 Data Portal has been used from various global resources to create their toolkits and workflows; similarly, the Portal uses input from global resources. The following are some indicative examples of the reach and impact of the Portal and the interconnectedness of the databases and toolkits:

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.



- Nextstrain[9] use open data from the COVID-19-Data Portal (raw data, consensus sequences and variant calls) to provide an open-source toolkit enabling SARS-CoV-2 bioinformatics and visualization.
- Open Targets retrieves data from such resources as Ensembl, UniProt and ChEMBL and provide analysis and curation to derive prioritisation lists of compounds with potential activities against genomics targets known, or suspected, to be of relevance to COVID-19. Provided as the Open Targets "Target Prioritisation Tool[10]" - now also made available from the COVID-19 Data Portal, this offers a useful service to those in the drug discovery and development worlds.
- Galaxy CRG provide a collection of Galaxy workflows for the detection and interpretation of sequence variants in SARS-CoV-2 on their Global platform for SARS-Cov-2 analysis. Galaxy use the raw data obtained from the COVID-19 Portal/ENA to analyse and generate consensus sequences/variant calls and feed it into their CRG Viral Beacon variant browser[11], providing analyses and visualizations to the global community.

3. Next Steps

Our next steps include the deployment of the cohort browser as soon as it is available. We also aim to connect databases beyond ELIXIR and add further apps and tools to the COVID-19 Data Portal.

[9] https://nextstrain.org/ncov/open/global
[10] https://covid19.opentargets.org/
[11] https://covid19beacon.crg.eu/

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

References

Gaia Cantelli, Guy Cochrane, Cath Brooksbank, Ellen McDonagh, Paul Flicek, Johanna McEntyre, Ewan Birney, Rolf Apweiler, *The European Bioinformatics Institute: empowering cooperation in response to a global health crisis*, Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D29–D37, https://doi.org/10.1093/nar/gkaa1077

Peter W Harrison, Rodrigo Lopez, Nadim Rahman, Stefan Gutnick Allen, Raheela Aslam, Nicola Buso, Carla Cummins et al., *The COVID-19 Data Portal: accelerating SARS-CoV-2 and COVID-19 research through rapid open access data sharing*, Nucleic Acids Research, Volume 49, Issue W1, 2 July 2021, Pages W619–W623, https://doi.org/10.1093/nar/gkab417

Delivery and Schedule

The delivery was delayed for 2 months due to urgent work on COVID-19 and further delays caused by COVID-19 consequences such as unavailability of staff due to sickness & work overload of remaining staff.

Adjustments

Adjustments made: none

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.