Deliverable D8.5 Data access points and tools

Project Title (Grant agreement no.):	ELIXIR-CONVERGE: Connect and align ELIXIR Nodes to deliver sustainable FAIR life-science data management services (871075)		
Project Acronym (EC Call):	ELIXIR-CONVERGE (H2020-INFRADEV-2018-2020)		
WP No & Title:	WP8 ELIXIR-CONVERGE European COVID-19 Data Platform		
WP leader(s):	Guy Cochrane (EMBL-EBI)		
Deliverable Lead Beneficiary:	1 - EMBL-EBI		
Contractual delivery date:	31/12/2021	**Actual delivery date:**	31/01/2021
Delayed:	*yes*		
Partner(s) contributing to this deliverable:	*EMBL-EBI*		
Authors: Guy Cochrane (EMBL-EBI), Carla Cummins, (EMBL-EBI), Nadim Rahman (EMBL-EBI), **Contributors:** Marianna Ventouratou (EMBL-EBI) **Acknowledgments (not grant participants):** *N/A*			
Reviewers:	ELIXIR-CONVERGE Management Board (MB) members.		

Log of changes

DATE	Mvm	Who	Description
31/10/2021	0v1	Guy Cochrane (EMBL-EBI)	Initial version
19/1/2022	0v2	Marianna Ventouratou (EMBL-EBI)	Sent to PMO after incorporating internal WP feedback
20/1/2022	0v3	Nikki Coutts (ELIXIR Hub)	Circulated to the MB for final review before submission
31/1/2022	0v4	Name Surname (acronyms)	MB comments addressed
31/1/2022	1v0	Nikki Coutts (ELIXIR Hub)	Final version to be uploaded into EC Portal

Table of contents

1. Executive Summary

Within the scope of deliverable 8.5 which was to enhance access to the COVID-19 Data Portal data, we developed the following data access support tools and services:

COVID-19 Data Portal API: a proxy layer on top of EBI search pulling together a diverse range of resources into a single search service.

COVID-19 Data Portal Download: an interactive download tool, allowing users to select different formats for metadata and data downloads.

OmicsDI Indexing Service: an integrated and open source platform which indexes sources external to EBI and enables display on the COVID-19 Data Portal. Two ELIXIR core resources have been indexed: Cellosaurus and HPA.

Covid-19 Data Portal File Downloader Tool: a user runnable tool for downloading all host and viral sequence data from ENA (European Nucleotide Archive) as seen on the COVID-19 Data Portal. The tool can download data in different formats and can generate scripts that run interactively with local data syncing of COVID-19 data available on the Portal.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

Third-Party Tools: Open data sharing has led to the further development of additional open-source tools that provide support to the wider scientific community. As part of the COVID-19 Data Platform three main examples of such tools have been identified: the COVID-19 Viral Beacon, Nextrain's public phylogeny and report for INSDC-submitted consensus sequences, and Galaxy.

GISAID Metadata Conversion Tool: to support users who have been sharing data with GISAID, this tool converts a metadata spreadsheet used during the GISAID submission process, to a sample XML file that is compatible with ENA's programmatic submission process.

2. Contribution toward project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives/key results:

Objective no. / Key Result no. Description	**Contributed to:**
Objective 1: Develop a sustainable and scalable operating model for transnational life-science data management support by leveraging national capabilities **(WP1, WP5**)	
Key Result 1.1: Established European expert network of data stewards that connect national data centres and similar infrastructures and drive the development of interoperable solutions following international best practice, including national interpretations of the General Data Protection Regulation (GDPR)	**No**
Key Result 1. 2: Development of joint guidelines and common toolkit that are adopted into funder recommendations, with support available nationally and in local languages	**No**
Key Result 1.3: The catalogue of successful national business models incorporated into national strategies	**No**
Key Result 1.4: The developed "sustainable and scalable operating model for transnational life-science data management support" is adopted into national ELIXIR Node	**No**
Objective 2: Strengthen Europe's data management capacity through a comprehensive training programme delivered throughout the European Research Area **(WP2, WP6)**	
Key Result 2.1: A comprehensive ELIXIR Training and Capacity building programme in Data Management, directed at both data managers and ELIXIR users, and connected to the national training programmes in Data Management in the ELIXIR Nodes and prospective ELIXIR Member countries.	**No**

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

Key Result 2.2: Development of a collective group of trainers that support scalable deployment of Data Management training across ELIXIR Nodes.	**No**
Key Result 2.3: A substantial cohort of data managers, Node coordinators and researchers with specific data management skills, business planning and knowledge of transnational operations across the ELIXIR Nodes	**No**
Objective 3: Align national data management standards and services through a sustainable, scalable and cost-effective data management toolkit **(WP2, WP3, WP5**)	
Key Result 3.1: Assemble a full-stack harmonised common toolkit comprising all aspects of data management: from data capture, annotation, and sharing; to integration with analysis platforms and making the data publicly available according to international standards.	**No**
Key Result 3.2: Provide exemplar toolkit configurations for prioritised demonstrators to serve as templates for future use.	**No**
Key Result 3.3: Establish national capacity in using as well as updating, extending and sustaining the toolkit across the ERA.	**No**
Key Result 3.4: Enable 'FAIR at source' practice for data generation, and analytical process pipeline implementation by flexible deployment of the toolkit in national operations	**No**
Objective 4: Align national investments to drive local impact and global influence of ELIXIR **(WP4,WP6**)	
Key Result 4.1: Development of a Node Impact Assessment Toolkit based on RI-PATHS methodology.	**No**
Key Result 4.2: Adoption of Impact assessment in ELIXIR Nodes, supported by Node coordinators network and feedback on applicability from dialogues with national funders.	**No**
Key Result 4.3: Creation of national public-private partnerships and industry outreach where open life-science data and services stimulate local bioeconomy	**No**
Key Result 4.4: Growth in reach, impact and engagement of stakeholder communication assessed by established ELIXIR Communications metrics	**No**
Key Result 4.5: Initiating and advancing discussions on Membership (EU and international) or strategic partnerships (international countries) following ELIXIR-CONVERGE workshops.	**No**
Objectives - WP8 - ELIXIR-CONVERGE European COVID-19 Data Platform	
O8.1 Data management support for EU projects (Task 8.1)	**No**

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

O8.2 Mobilisation of analysis upon SARS-CoV-2 sequence data (Task8.2)	**No**
O8.3 Enhanced access to data, tools and support (Task 8.3)	**Yes**

3. Introduction

The scope of deliverable 8.5 falls within the remit of Task 8.3 which is to enhance access to the COVID-19 Data Portal data. For that, the aim was to extend and enhance the access points for data in the system by providing specific tools and support for automated synchronisation of all data in the platform into external computational facilities. Additionally, the scope was to work with EMBL-EBI's many networks to enable data flows from the system and the connection of new third-party tools and interfaces. Drawing from and extending the tools and services offered via ELIXIR's Tools, Interoperability and Compute Platforms, the aim was, among others, to enable data and workflow access from within Galaxy, present data in compliance with BioSchemas.org and synchronise data into cloud instances, respectively.

The methodology applied to achieve the above was new and enhanced data access points to SARS-CoV-2 data supporting such operations as full data synchronisation, accessibility from external cloud compute and embedding in third party tools.

4. Description of work accomplished and Results

CV19 Portal API

The COVID-19 Data Portal has an integrated REST API, suitable for querying metadata tables programmatically. Each display in the portal has its own endpoint: https://www.covid19dataportal.org/api-documentation. This API is a proxy layer on top of EBI Search, which pulls together a diverse range of resources into a single search service, and uses the Apache Lucene query syntax (https://www.ebi.ac.uk/ebisearch/documentation.ebi).

Portal Download

An interactive download option is available through the COVID-19 Data Portal (figure 1). Users can select data of interest via checkboxes and click on the 'Download' button above the table to bring up the download menu. Several different formats are available for both metadata and data itself, depending on the data type. For sequences, EMBL and FASTA formats are available. For raw read data, XML and FASTQ formats are available.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

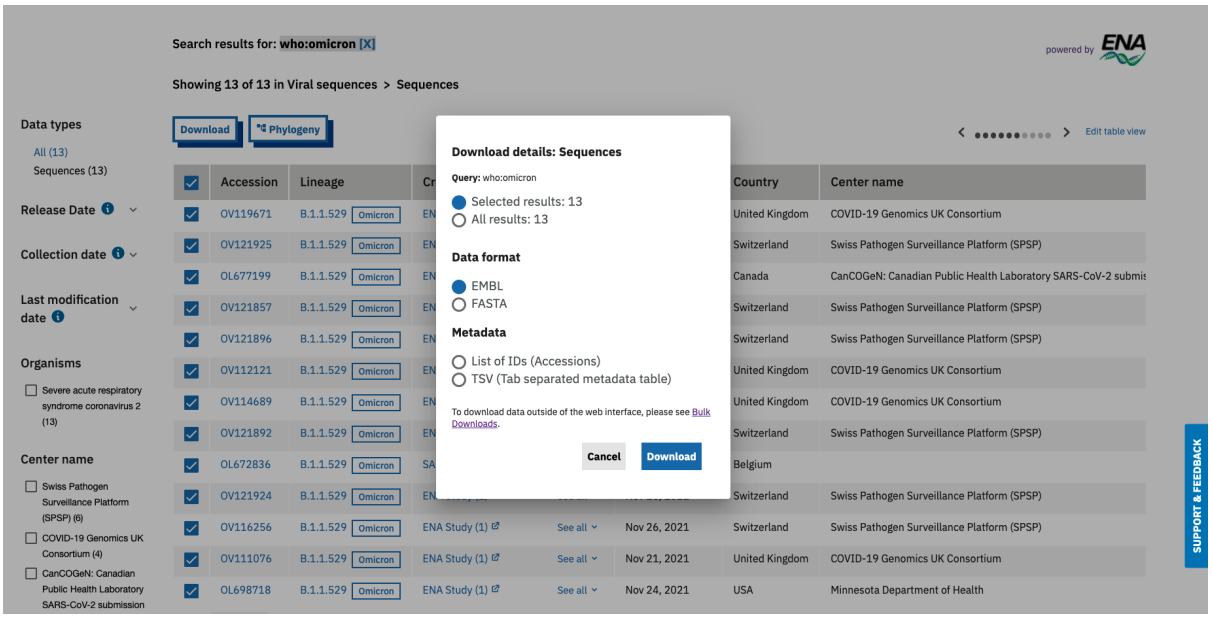

Figure 1: COVID-19 Data Portal web interface for downloading data of interest

OmicsDI Indexing Service

Omics Discovery Index is an integrated and open source platform facilitating the access and dissemination of omics datasets. Through this service, data sources external to EBI can be indexed and become discoverable through EBI search services. This also enables display within the COVID-19 Data Portal.

To date, we have worked with two ELIXIR core resources to index their SARS-CoV-2 data : Cellosaurus (https://web.expasy.org/cellosaurus/) and Human Protein Atlas (https://www.proteinatlas.org/humanproteome/sars-cov-2). Both resources can now be searched through the EBI Search service and have been fully integrated into the COVID-19 Data Portal:

- Cellosaurus: https://www.covid19dataportal.org/samples?db=cellosaurus
- HPA: https://www.covid19dataportal.org/proteins?db=hpa-covid19

CDP File Downloader Tool

The COVID-19 Data Portal (CDP) downloader tool is a user runnable tool used for downloading all host and viral sequence data from ENA (European Nucleotide Archive), as seen on the COVID-19 Data Portal. It is based on Java 8 and follows the same menu structure as the COVID-19 Data Portal, for ease of use.
The tool can download data in many different formats (XML/FASTA/EMBL/FASTQ) using either the FTP or Aspera transfer protocols. Additionally, the tool can generate scripts which can be run iteratively to keep your local data in sync with all data available on the portal. For information on

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

downloading and running this tool, please refer to https://www.covid19dataportal.org/bulk-downloads.

Third-Party Tools

Open data sharing has supported the downstream usage, analysis and development of additional open-source tools that provide further support for the wider scientific community. This helps support the identification of novel variants, aspects of the development process of vaccines and repurposing of drugs, to name a few. Three main examples have been discussed below.

Firstly, the COVID-19 Viral Beacon (https://covid19beacon.crg.eu/), developed by CRG in Barcelona, provides a resource that enables users to browse, filter and explore SARS-CoV-2 variability at the genomic, amino acid and motif levels. As part of this, the plots and browsers are based on (1) submitted INSDC SARS-CoV-2 consensus sequences, and (2) systematically analysed filtered variant calls generated by ENA and partners. This latter aspect is generated from the analysis of raw read datasets within the COVID-19 Data Portal via workflows, developed by partners in VEO, that have been mentioned in previous CONVERGE reports.

Secondly, Nextstrain maintains a public phylogeny and report for INSDC-submitted consensus sequences. This can be viewed on the Nextstrain website: https://nextstrain.org/ncov/open/global. Nextstrain is an open-source project that explores and presents the public health and epidemiological potential of pathogenic genome data. In addition to maintaining and presenting their own visualisations, Nextstrain enables users to create their own reports. The INSDC sequence phylogeny presents Nextclade assignments, which correlate to the WHO lineage classification system. Furthermore, users can browse an interactive map presenting the geographic distribution of sequences, distribution of variation across the SARS-CoV-2 genome and total frequency of circulating variants within the sequence data.

Finally, Galaxy (https://covid19.galaxyproject.org/) provide a global platform to support users in analysing and gathering interpretations from SARS-CoV-2 data that has been shared to the COVID-19 Data Portal. The project provides an environment for users to create bespoke workflows, and utilise open-source tools to analyse datasets. In the SARS-CoV-2 context, Galaxy enables users to pull from a large and mirrored set of various data types: https://covid19.galaxyproject.org/data/. Furthermore the group regularly produce bespoke analyses on datasets of interest to better understand SARS-CoV-2 and its variants, for example: https://virological.org/t/selection-analysis-identifies-significant-mutational-changes-in-omicron-that-are-likely-to-influence-both-antibody-neutralization-and-spike-function-part-1-of-2/771.

GISAID Metadata Conversion Tool

GISAID provides an additional resource that has enabled researchers to share SARS-CoV-2 consensus sequences within the EpiCoV database. Therefore to support users who have been sharing

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

data with GISAID, the Metadata Conversion Tool was developed. This tool converts a metadata spreadsheet used during the GISAID submission process, to a sample XML file that is compatible with ENA's programmatic submission process. To achieve this, metadata fields are mapped between the two database's to appropriately tag metadata.

The codebase is open-source and available at the following GitHub repository: https://github.com/enasequence/ena-content-dataflow/tree/master/scripts/gisaid_to_ena.

5. Impact

This deliverable report presented, among others, cases of open access tools and the impact of this work is in itself not measurable and outside the scope of this WP.

6. Next Steps

N.A.

7. Deviation from Description of Action

N.A.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.