

Deliverable D3.3

D3.3.1 COVID-19 Data Portal

Project Title (grant agreement No)	BY-COVID Grant Agreement 101046203		
Project Acronym (EC Call)	BY-COVID		
WP No & Title	WP3: COVID-19 integration platform		
WP Leaders	Henning Hermjakob (EMBL-EBI) Mari Kleemola (CESSDA/TAU-FSD)		
Deliverable Lead Beneficiary	1 - EMBL-EBI		
Contractual delivery date	30/09/2023	Actual Delivery date	28/09/2023
Delayed	No		
Partner(s) contributing to this deliverable	EMBL-EBI, CESSDA/TAU-FSD, UOXF, ERINHA, Lygature, ELIXIR-NL/VUmc, UNIMAN		
Authors	Henning Hermjakob (EMBL-EBI) Orcid: 0000-0001-8479-0262 Mari Kleemola (CESSDA/TAU-FSD) Orcid: 0000-0001-8855-5075 Marianna Ventouratou (EMBL-EBI) Orcid: 0000-0001-7964-6127 Allyson Lister (UOXF) Orcid: 0000-0002-7702-4495 Susanna-Assunta Sansone (UOXF) Orcid: 0000-0001-5306-5690 Romain David (ERINHA) Orcid: 0000-0003-4073-7456		



Contributors	Julia Lischke (Lyg) Orcid: 0000-0002-5524-2838 Robin Navest (Lyg) Orcid: 0000-0002-0152-2092 Jeroen Belien (ELIXIR-NL/VUmc) Orcid: 0000-0002-7160-5942 Nick Juty (UNIMAN) Orcid: 0000-0002-2036-8350 Stian Soiland-Reyes (UNIMAN) Orcid: 0000-0001-9842-9718 Carole Goble (UNIMAN) Orcid: 0000-0003-1219-2137)
Reviewers	Dimitra Kondyli, Rafael Buono

Log of changes

Date	Who	Description
11/08/2023	Mari Kleemola (CESSDA/TAU-FSD)	Initial structure based on WP3 discussions.
28/08/2023	Henning Hermjakob (EMBL-EBI)	Text editing
04/09/2023	WP3 team	Text editing, additions, feedback
05/09/2023	Henning Hermjakob (EMBL-EBI), Mari Kleemola (CESSDA/TAU-FSD)	Ready for internal review
19/09/2023	Marianna Ventouratou (EMBL-EBI), Mari Kleemola (CESSDA/TAU-FSD)	Addressed reviewers' comments



Disclaimer

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them. This deliverable is licensed under a Creative Commons Attribution 4.0 International License.

Table of contents

1. Executive Summary	3
2. Contribution towards project objectives	5
3. Introduction	7
4. Description of work accomplished	8
4.1 Discoverability: EBI Search and three-tiered approach	8
4.2 New Data Sources	9
FAIRsharing	9
Infectious Disease Toolkit (IDTk)	11
CESSDA	12
European University Institute	12
European Health Information Portal	13
Dutch National COVID-19 metadata portal	13
Federated European Genome-Phenome Archive	13
EU-OPENSREEN	14
4.3 Website Development	14
4.4 The Pathogens Portal	16
5. Results and discussion	17
6. Next steps	18
7. Impact	18
8. References	20



1. Executive Summary

The BY-COVID project is one of the Horizon Europe projects that has supported the operation and enhancements of the European COVID-19 Data Portal¹, a critical resource that provides access to open literature and data on both the SARS-CoV-2 virus and the disease. For instance, the portal now contains >17.5 million viral sequences and more than one million open scientific articles related to SARS-CoV-2 and COVID-19, as of September 2023.

BY-COVID Work Package 3 is focused on services for the discovery and integration of COVID-19 data by delivering a flexible, tiered metadata discovery system across different domains, metadata standards, and maturity/robustness levels of data sources. This enables the linking of FAIR data and metadata on SARS-CoV-2 and COVID-19, and crucially, on other related viruses and diseases, and on socio-economic consequences, across research fields, from omics, clinical, and epidemiological research, to social sciences and humanities, improving preparedness for future potential emergent disease or pandemic scenarios.

Based on the metadata model developed in D3.1, WP3 has implemented the infrastructure to support the three-tiered discoverability concept in the COVID-19 Data Portal² as described in D3.2. This report provides a status update of the Portal and outlines the next steps.

¹ COVID-19 Data Portal: <https://www.covid19dataportal.org/> [accessed 20/09/2023]

² COVID-19 Data Portal: <https://www.covid19dataportal.org/> [accessed 20/09/2023]



2. Contribution towards project objectives

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

	Key Result No and description	Contributed
Objective 1 Enable storage, sharing, access, analysis and processing of research data and other digital research objects from outbreak research	1. A research data management practice in European research infrastructures practice that drives discovery, access and reuse of outbreak data and directly links experimental data from HORIZON-INFRA-2021-EMERGENCY-02 transnational access projects into the COVID-19 Data Portal.	Yes
	2. Workflows and processing pipelines that integrate transparent quality management and provenance and are openly shared.	Yes
	3. Research infrastructures on-target training so that users can exploit the platform	No
	4. Engagement so that stakeholders (RI, national centres, policy makers, intergovernmental organisations, funders and end-users) incorporate FAIR and open data in infectious disease guidelines and forward planning.	Yes
Objective 2 Mobilise and expose viral and human infectious disease data from national centres	1. A comprehensive registry of available data with established procedures to collate data governance models, metadata descriptions and access mechanisms in a pandemic scenario.	Yes
	2. Mechanisms for the initial discovery across data sources based on available metadata at the reference collection.	Yes
	3. Demonstrated transnational linking of real-world data from national surveillance, healthcare, registries and social science data that allow the assessment of variants to serve the research needs of epidemiology and public health.	Yes
	4. Demonstrated assessment of emerging SARS-CoV-2 variants against data generated in the on-going European VACCELERATE clinical trials project to investigate vaccine efficacy.	No
Objective 3	1. A platform that links normative pathogen genomes and variant representations to research cohorts and mechanistic studies to understand the biomolecular determinants of	Yes



Link FAIR data and metadata on SARS-CoV-2 and COVID-19	variant response on patient susceptibility, and disease pathways.	
	2. An open and extensible metadata framework adopted cross-domain that supports comprehensive indexing of the infectious disease resources based on mappings across resources and research domains.	Yes
	3. A provenance framework for researchers and policy-makers that enables trust in results and credit to data submitters, workflow contributors and participant resources.	Yes
Objective 4 Develop digital tools and data analytics for pandemic and outbreak preparedness...	1. Broad uptake of viral <i>Data Hubs</i> across Europe deliver an order-of-magnitude increase in open viral variant detection and sharing.	No
	2. Infrastructure and quality workflows mobilised and shared to produce open, normative variant data that is incorporated into national and regional data systems and decision making.	No
Objective 5 Contribute to the Horizon Europe European Open Science Cloud (EOSC) Partnership and European Health Data Space (EHDS)	1. Guidelines and procedures for FAIR data management and access will be established, building on work of other guideline producing consortia such as the Global Alliance for Genomics and Health (GA4GH), the 1Mio Genomes Initiative (1MG) and the Beyond One Million Genomes project (B1MG).	Yes
	2. Services, software, protocols, guidelines and other research objects that are openly accessible for reuse by the EOSC Association and the community at large as a foundation for European preparedness for infectious diseases, leveraging developments in EOSC-Life, SSHOC, EOSC-Future, EGI-ACE and other EOSC projects.	Yes
	3. Alignment (both policy and implementation routes) will have been achieved between the data governance strategies for routinely collected health data in the EHDS initiative, including the TEHDAS Joint Action and future EHDS Pilot Actions.	No
	4. To empower national centres to build capacity and train platform users and data providers (e.g., from life, social or health sciences), and with experts from across partner institutions collaborating to create training materials for the identified gaps, and to exchange experiences and knowledge.	Yes



3. Introduction

The BeYond-COVID (BY-COVID) project is one of the Horizon Europe projects that has supported the operation and enhancements of the European COVID-19 Data Portal³, a critical resource that provides access to literature and data on both the SARS-CoV-2 virus and the disease. For instance, the portal now (August 2023) contains >17.5 million viral sequences and more than one million open scientific articles related to COVID-19.

BY-COVID aims to provide comprehensive open data on SARS-CoV-2, COVID-19, and other viruses and infectious diseases, across scientific, medical, public health and policy domains. The project mobilises existing data resources (i.e catalogues) and marshals the resources for research, connects and exposes the data and resources via the COVID-19 Data Portal, and drives use and analysis by connecting workflows, national portals and analysis environments. The 3-year project brings together 53 partners from 19 countries.

This deliverable is part of BY-COVID Work Package 3 that is focused on services for the discovery, integration and citation of COVID-19 data by delivering a flexible, tiered metadata discovery system across different domains, metadata standards, and maturity/robustness levels of data sources. This enables the linking of FAIR [1] data and metadata on SARS-CoV-2 and COVID-19, on other related viruses and diseases, and on socio-economic consequences, across research fields, from omics, clinical, and epidemiological research, to social sciences and humanities.

Our inter-domain metadata mapping (Task 3.1, D3.1 Metadata standards⁴) supports (meta)data discovery, access, and analysis across fields from molecular biology to social sciences. The harmonised metadata is discoverable through the central index (Task 3.2, D3.2 Implementation of cloud-based, high performance, scalable indexing system⁵) and web portal (Task 3.3 - this deliverable), but also openly accessible for third party applications through web services. We have continuously improved the COVID-19 Data Portal, in particular through the addition of more than 150 related resources⁶ based on the FAIRsharing workflow in collaboration with WP2, five new fully indexed resources including social sciences data from the CESSDA catalogue⁷, and addition of a global search

³ COVID-19 Data Portal: <https://www.covid19dataportal.org/>

⁴ Hermjakob H, Kleemola M, Moilanen K, Sansone S-A, Lister A, David R, Panagiotopoulou M, Ohmann C, Bellen J, Lischke J, Juty N & Soiland-Reyes S. (2022). BY-COVID - D3.1 - Metadata standards. Documentation on metadata standards for inclusion of resources in data portal (V1.0). Zenodo. <https://doi.org/10.5281/zenodo.6885016>

⁵ Hermjakob H, Kleemola M, Moilanen K, Tuominen M, Sansone S-A, Lister A, David R, Panagiotopoulou M, Ohmann C, Belien J, Lischke J, Juty N & Soiland-Reyes S. (2022). BY-COVID D3.2: Implementation of cloud-based, high performance, scalable indexing system. Zenodo. <https://doi.org/10.5281/zenodo.7129553>

⁶ COVID-19 Data Portal Related resources: <https://www.covid19dataportal.org/related-resources> [accessed 16/08/2023]

⁷ COVID-19 Data Portal social sciences <https://www.covid19dataportal.org/search/social-sciences> [accessed 16/08/2023]



functionality to the Portal home page. We anticipate development and metadata modelling work for the use-case driven support of complex data sources in collaboration with WPs 2, 4 and 5. Here, we describe developments, current status of the COVID-19 Data Portal, and the next steps.

4. Description of work accomplished

4.1 Discoverability: EBI Search and three-tiered approach

The discoverability concept of the COVID-19 Data Portal is based on the EBI Search system and a three-tiered indexing concept. These are described in detail in D3.2, so here we provide only a short summary.

EBI Search [2] is a cloud-based, high performance, scalable text search engine that provides easy and uniform access to the biological data resources hosted at the European Bioinformatics Institute (EMBL-EBI). EBI Search, based on Apache Lucene, provides easy inter-domain navigation via a network of cross-references. It can be accessed over the web or programmatically using the RESTful Web Services interface. This allows its search and retrieval capabilities to be exploited in workflows and analytical pipelines. The API-based query interface⁸ is available both within and outside EMBL-EBI, and can be used to develop website-specific search/results pages.

Exposing and effectively connecting and linking different data types requires indexing based on cross-mapped metadata, as described in D3.1. Indexing and incorporation of metadata into the COVID-19 Data Portal proceeds via a flexible, **tiered system for metadata integration** (Table 1). We anticipate that some external resources over the course of the project will migrate from tier 3 upwards, as resources for the detailed metadata harmonisation and technical indexing become available.

Table 1. Three-Tiered Indexing Concept

Tier 1	Deepest indexing available, capturing granular, record level identifiers and metadata, support for interoperability use cases.	Aim: Key resources migrate from Tier 3 to Tier 1 over project duration.
Tier 2	Coarse-grained metadata and attributes, focus on record level discoverability.	

⁸ EBI Search API: <https://www.ebi.ac.uk/ebisearch/apidoc.ebi> [accessed 16/08/2023]

Tier 3	Focus on resource level discoverability.	↑
--------	--	---

4.2 New Data Sources

FAIRsharing

FAIRsharing⁹ is a manually curated, informative and educational resource that maps the landscape of community-developed standards, databases (repositories and knowledge bases) and policies across disciplines. As of August 2023, it contains 1657 standards, 2044 databases and 168 policies, citable via DOI. FAIRsharing complements the COVID-19 Data Portal by acting as a catalogue of data sources, describing their characteristics including access terms and protocols, and the standards used at the source to represent the data. Resource metadata (Tier 3) is captured in FAIRsharing records and selectively imported to the COVID-19 Data Portal through the API.

The FAIRsharing BY-COVID Collection¹⁰ is a catalogue and knowledge graph of data sources and their characteristics, including access terms, protocols and standards used to represent the data and metadata. The Collection currently contains 21 data sources (Table 2), and was developed in collaboration between WP2 and WP3, with BY-COVID members from social science and humanities, health and clinical data, images, genomic and phenotypic data and chemical biology. In addition to the BY-COVID Collection, FAIRsharing provides a wider collection of 99 resources potentially relevant to COVID-19 research¹¹. As described in D3.2, we have established a workflow to integrate the resources from both collections into the COVID-19 Data Portal section “Related Resources” (Tier 3)¹², which now contains 156 resources.

Table 2. The BY-COVID FAIRsharing resource collection (as of 15 June 2023)¹³

Domain	Resource and record in FAIRsharing	Type of data
--------	------------------------------------	--------------

⁹ FAIRsharing: <https://fairsharing.org/> [accessed 16/08/2023]

¹⁰ BY-COVID Data Resources in FAIRsharing: <https://fairsharing.org/3773> [accessed 16/08/2023]

¹¹ Wider collection of COVID-19 resources is maintained by the FAIRsharing team and is focused on patient response, clinical trials, virology studies or other related areas. The databases in the collection contain COVID-19 related datasets. <https://fairsharing.org/3538> [accessed 25/08/2023]

¹² COVID-19 Data Portal Related resources <https://www.covid19dataportal.org/related-resources> [accessed 25/08/2023]

¹³ The current status of the BY-COVID FAIRsharing resource collection: <https://fairsharing.org/3773>



Clinical and health	Health Data Research Innovation Gateway; https://doi.org/10.25504/FAIRsharing.nh1DmP	UK health datasets
	European Health Information Portal (HIP); https://doi.org/10.25504/FAIRsharing.8690f1	European health information
	ECRIN Clinical Research Metadata Repository; https://fairsharing.org/3067	European clinical studies, trial registrations, results summaries, journal articles, protocols
	Dutch National COVID-19 metadata portal; https://doi.org/10.25504/FAIRsharing.527321	Metadata of COVID-related data sets acquired within the Netherlands
	BBMRI-ERIC Directory; https://doi.org/10.25504/FAIRsharing.q9VUYM	Aggregate information about biobanks across Europe
	Dutch National COVID-19 clinical data dashboard; https://doi.org/10.25504/FAIRsharing.71bf06	Clinical data dashboard for the exploration and reuse of clinical data from Dutch university medical centres
	COVID-19 Data Portal; https://doi.org/10.25504/FAIRsharing.f3b7a9	COVID-19 datasets and tools including SARS-CoV-2 sequence data
Genotypic and phenotypic	The European Genome-phenome Archive (EGA); https://doi.org/10.25504/FAIRsharing.mya1ff	Personally identifiable genetic, phenotypic, and clinical data
	European Mouse Mutant Archive (EMMA); https://doi.org/10.25504/FAIRsharing.g2fjt2	Mutant mice strains essential for basic biomedical research provided by INFRAFRONTIER
Social sciences and humanities	EUI COVID-19 social sciences and humanities (SSH) Data Portal; https://doi.org/10.25504/FAIRsharing.97367f	COVID-19-related research in the social sciences and humanities
	Consortium of European Social Science Data Archives (CESSDA) Data Catalogue; https://doi.org/10.25504/FAIRsharing.a12316	European social science data
	European Social Survey (ESS) Data Portal; https://fairsharing.org/4838	European cross-national survey data measuring the attitudes, beliefs and behaviour patterns of diverse populations in more than thirty nations



	Survey of Health, Ageing and Retirement in Europe (SHARE) Research Data Center; https://fairsharing.org/4839	European survey data on the effects of health, social, economic and environmental policies over the life-course of European citizens and beyond
	Open Data Infrastructure for Social Science and Economic Innovations (ODISSEI) Portal; https://fairsharing.org/4841	Metadata from most data collections relevant to the social science community in the Netherlands
Images	Electron Microscopy Public Image Archive (EMPIAR); https://doi.org/10.25504/FAIRsharing.dff3ef	Raw images underpinning 3D cryo-EM maps and tomograms
	Electron Microscopy Data Bank (EMDB); https://doi.org/10.25504/FAIRsharing.651n9j	Electron microscopy density maps of macromolecular complexes and subcellular structures
	Image Data Resource (IDR); https://doi.org/10.25504/FAIRsharing.6wf1zw	Image data from genetic, RNAi, chemical, localisation and geographic high content screens, super-resolution microscopy and digital pathology
	BiImage Archive; https://doi.org/10.25504/FAIRsharing.x38D2k	Biological images that are useful to life-science researchers
Chemical Biology	ChEMBL; https://doi.org/10.25504/FAIRsharing.m3jtpg	Chemical, bioactivity and genomic data to aid the translation of genomic information into effective new drugs
	European Chemical Biology Database (ECBD); https://fairsharing.org/3717	Experimental results from biological screening programs
	COVID 19-NMR; https://fairsharing.org/4850	RNA and protein structural data for SARS-CoV-2 as well as other viruses

Infectious Disease Toolkit (IDTk)

The Infectious Diseases Toolkit (IDTk) is a GitHub-based community effort to expose best practices and showcase solutions to data challenges affecting the response to infectious



diseases outbreaks, created as part of the BY-COVID project¹⁴. It was integrated into the COVID-19 Data Portal as a “Related Resource” (Tier 3) in month 12. Since August 2023, based on the search data file exposed by the ELIXIR toolkit theme¹⁵ used by IDTk, we are indexing the chapter-level metadata of IDTk and thus making it discoverable as a Tier 2 resource. Thus, the IDTk is the first resource that we have upgraded from Tier 3 resource level discoverability to Tier 2 record level discoverability (see section 4.1).

See the Infectious Diseases Toolkit (IDTk) as a facet of literature metadata in COVID-19 Data Portal at <https://dev.covid19dataportal.org/search/literature>¹⁶

CESSDA

The Consortium of European Social Science Data Archives (CESSDA) ERIC provides large-scale, integrated and sustainable data services to the social sciences. As described in detail in a D3.2 case study, the COVID-19 related CESSDA Data Catalogue¹⁷ subset has been integrated into the COVID-19 Data Portal, in month 10. In June 2023 (month 21), the CESSDA team updated the exported metadata file, which was automatically picked up by our indexing workflow, increasing the number of CESSDA records from 432 to 540, and demonstrating the external resource update workflow of the COVID-19 Data Portal.

See CESSDA data as a facet of social sciences & humanities metadata in COVID-19 Data Portal at <https://covid19dataportal.org/search/social-sciences>

European University Institute

The European University Institute (EUI) COVID-19 data portal¹⁸ provides integrated search and discovery over available COVID-19-related data worldwide supporting research in the Social Sciences and Humanities. Building on the toolkit and workflow developed for CESSDA, we prepared the EUI portal data for integration into the COVID-19 Data Portal, adding 522 new records.

See EUI as a facet of social sciences & humanities metadata in COVID-19 Data Portal at <https://covid19dataportal.org/search/social-sciences>

¹⁴ IDTk: <https://www.infectious-diseases-toolkit.org/> [accessed 25/08/2023]

¹⁵ GitHub ELIXIR toolkit theme: <https://github.com/ELIXIR-Belgium/elixir-toolkit-theme> [accessed 25/08/2023]

¹⁶ COVID-19 Data Portal: In development [19/09/2023]. When in production: <https://covid19dataportal.org/search/literature> [accessed 20/09/2023]

¹⁷ CESSDA Data Catalogue <https://datacatalogue.cessda.eu/> [accessed 18/09/2023]

¹⁸ COVID-19 SSH Data Portal: <https://covid19data.eui.eu/> [accessed 25/08/2023]



European Health Information Portal

The aim of the Health Information Portal is to provide access to population health and healthcare data across Europe. The portal is the gateway for researchers and policy makers to make use of the services of the Population Health Information Research Infrastructure (PHIRI). The portal provides a FAIR Data Point¹⁹ that exports RDF²⁰ data following the DCAT V2 specification²¹. We have developed a library to query the available dcat:Dataset records from the data point, access the DCAT-formatted data provided through this API, and have imported the COVID-19 relevant record metadata into the COVID-19 Data Portal.

See European Health Information Portal data as a facet of research infrastructures metadata in COVID-19 Data Portal at

<https://covid19dataportal.org/search/research-infrastructures>

Dutch National COVID-19 metadata portal

The Dutch National COVID-19 metadata "search and request" portal²² allows exploration and reuse of observational (clinical/health) data from COVID-19 observational projects funded by ZonMw, Dutch university medical centres (UMCs) data sources, as well as collaborating clinical care and other regional hospitals. The dashboard provides researchers with a clear overview of available clinical/health data, allowing them to search for specific data and simplifying access to such data where the necessary ethical and legal conditions have been met. Re-using the tools developed for the European Health Information Portal, we have added the relevant records of the Dutch National COVID-19 metadata portal to the COVID-19 Data Portal, benefitting from the DCAT standard and FAIR Data Point specifications to minimise development time. As the Dutch portal evolves, we expect to index additional metadata from the portal to increase discoverability.

See Dutch National COVID-19 data as a facet of Research Infrastructures metadata in

COVID-19 Data Portal at <https://covid19dataportal.org/search/research-infrastructures>

Federated European Genome-Phenome Archive

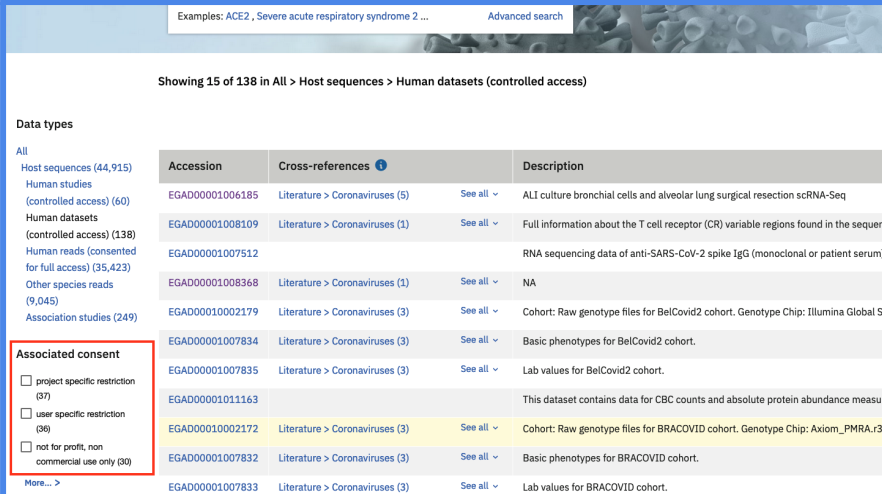
Protected data from the Federated European Genome-Phenome Archive has been part of the COVID-19 Data Portal from its inception. In the initial discussions on the metadata schema (D3.1), access restrictions were highlighted as a topic of interest for integration. WP2 partners have added Data Use Ontology (DUO) [3] annotation to the metadata set exported to the COVID-19 Data Portal, which is now indexed and accessible as a facet (Figure 1), establishing a first step towards harmonised access restriction annotation.

¹⁹ FAIR Data Point website: <https://www.fairdatapoint.org/> [accessed 25/8/2023]

²⁰ Resource Description Framework: https://en.wikipedia.org/wiki/Resource_Description_Framework [accessed 03/09/2023]

²¹ Data Catalog Vocabulary (DCAT): <https://www.w3.org/TR/vocab-dcat-2/> [accessed 25/08/2023]

²² Health-RI COVID-19 initiatives: <https://covid19initiatives.health-ri.nl/> [accessed 19/09/2023]



Examples: ACE2 , Severe acute respiratory syndrome 2 ... Advanced search

Showing 15 of 138 in All > Host sequences > Human datasets (controlled access)

Data types

All

- Host sequences (44,915)
- Human studies (controlled access) (60)
- Human datasets (controlled access) (138)
- Human reads (consented for full access) (35,423)
- Other species reads (9,045)
- Association studies (249)

Associated consent

- project specific restriction (37)
- user specific restriction (39)
- not for profit, non commercial use only (30)

More... >

Accession	Cross-references	Description
EGAD00001006185	Literature > Coronaviruses (5)	ALI culture bronchial cells and alveolar lung surgical resection scRNA-Seq
EGAD00001008109	Literature > Coronaviruses (1)	Full information about the T cell receptor (CR) variable regions found in the sequenc
EGAD00001007512	Literature > Coronaviruses (1)	RNA sequencing data of anti-SARS-CoV-2 spike IgG (monoclonal or patient serum).
EGAD00001008368	Literature > Coronaviruses (1)	NA
EGAD00010002179	Literature > Coronaviruses (3)	Cohort: Raw genotype files for BelCovid2 cohort. Genotype Chip: Illumina Global Sc
EGAD00001007834	Literature > Coronaviruses (3)	Basic phenotypes for BelCovid2 cohort.
EGAD00001007835	Literature > Coronaviruses (3)	Lab values for BelCovid2 cohort.
EGAD00001011163	Literature > Coronaviruses (3)	This dataset contains data for CBC counts and absolute protein abundance measur
EGAD00010002172	Literature > Coronaviruses (3)	Cohort: Raw genotype files for BRACOVID cohort. Genotype Chip: Axiom_PMR3
EGAD00001007832	Literature > Coronaviruses (3)	Basic phenotypes for BRACOVID cohort.
EGAD00001007833	Literature > Coronaviruses (3)	Lab values for BRACOVID cohort.

Figure 1. Facet ‘Associated consent’ using Data Use Ontology

See metadata in COVID-19 Data Portal at

<https://www.covid19dataportal.org/search/host-sequences?db=human-studies> and

<https://www.covid19dataportal.org/search/host-sequences?db=human-datasets>

EU-OPENSREEN

EU-OPENSREEN²³ integrates high-capacity screening platforms throughout Europe. All compound structures and primary screening data are published in the open-access European Chemical Biology Database (ECBD)²⁴, where they are made available to a wide scientific audience. Based on the agreed metadata schema, we have integrated COVID-19 relevant metadata records into the COVID-19 Data Portal.

See EU-OPENSREEN data as a facet of Research Infrastructures metadata in COVID-19 Data Portal at <https://covid19dataportal.org/search/research-infrastructures>

4.3 Website Development

All new metadata resources are discoverable through the COVID-19 Data Portal global search. The addition of the new resources, as well as requests from project partners and users, resulted in regular website updates, in particular the new “Research Infrastructures” section (Figure 2) and improved visibility of the “National Data Portals” page (Figure 3).

²³ EU-Openscreen website: <https://www.eu-openscreen.eu/> [accessed 25/08/2023]

²⁴ ECBD website: <https://ecbd.eu/> [accessed 25/08/2023]



COVID-19 Data Portal

About Tools FAQ Related Resources Bulk Downloads Submit Data

Viral Sequences Host Sequences Expression Proteins Networks Cohorts More

Research infrastructures

Accelerating research through data sharing

Search Search

Examples: ... [Advanced search](#)

Showing 9 of 139 in All > Research infrastructures

Data types

All

Research infrastructures (139)

European Health Information Portal (12)

EU-OPENSREEN (5)

Dutch National COVID-19 metadata portal (122)

European Health Information Portal 12 results

COVID-19 Consolidated deaths

Cross references: [Ontology Lookup Service \(OLS\) \(1\)](#) [↗](#) Source: phiri

Emergency preparedness register for COVID-19

Cross references: [Ontology Lookup Service \(OLS\) \(3\)](#) [↗](#) Source: phiri

Impact of the COVID-19 pandemic on drug use and addictions

Cross references: [Ontology Lookup Service \(OLS\) \(3\)](#) [↗](#) Source: phiri

[View all 12 results in European Health Information Portal](#)

EU-OPENSREEN 5 results

Single dose inhibition of SARS-CoV-2 cytopathic effect in VeroE6

Source: [ecbd-covid19](#)

Figure 2. Research Infrastructures page:

<https://www.covid19dataportal.org/search/research-infrastructures>

COVID-19 Data Portal

About Tools FAQ Related Resources Bulk Downloads Submit Data

Viral Sequences Host Sequences Expression Proteins Networks Cohorts More

Partners

The organisations and funders behind the European COVID-19 Data Platform efforts

COVID-19 Data Platform partners National Data Portals Funding Projects Participating Infrastructures Working groups

The European Covid19 Data Portal is result of the efforts of multiple organisations.

The Covid-19 National Data Portal network

Map showing participating countries: Estonia, Greece, etc.

National Data Portals

- Estonia
- Greece

SUPPORT & FEEDBACK

Figure 3. National Data Portals page:

<https://www.covid19dataportal.org/partners?activeTab=National%20Data%20Portals>



Funded by
the European Union



4.4 The Pathogens Portal

While COVID-19 variants remain a significant public health risk, the focus of the international research community, as well as policy makers, is shifting to broader pandemic preparedness, as anticipated in the BY-COVID strategy. In collaboration with the EU funded VEO project (H2020 grant agreement no. 874735), we have initiated the transition of development focus from the COVID-19 Data Portal to the broader Pathogens Portal²⁵, using the codebase, experience, strategies, and established workflows where possible. On July 5, 2023, we formally released the Pathogens Portal (Figure 4). The list of pathogens featured in the portal was collated using the UK's Health and Safety Executive's list of approved biological agents²⁶ and the WHO's global priority pathogens list²⁷. It includes well-known pathogens that affect humans, including HIV, influenza, Hepatitis B, and the malaria parasite *Plasmodium falciparum*. It also covers lesser-known pathogens affecting humans, such as *Lassa mammarenavirus*, the cause of Lassa hemorrhagic fever, which can lead to deafness and even death in severe cases. The portal also contains hundreds of pathogens that affect other animals, which makes it a useful tool for food security and biodiversity, and is relevant to the One Health approach²⁸ to pandemic preparedness.

The Pathogens Portal currently presents nucleotide sequences, raw genomic data, sample metadata, and relevant scientific literature. We will integrate additional data types, including further molecular data types, as well as broader health and social sciences data. The Pathogens Portal provides a valuable platform for research by many pathogen-specific communities, but also an essential basis for rapid, targeted data provision in case of new potential pandemic threats as in the case of the recent global Mpox outbreak.

²⁵ Pathogens portal website: <https://www.pathogensportal.org/> [accessed 25/08/2023]

²⁶ Everything in Hazard group 2 or higher from the HSE list are incorporated. The list of taxonomy names and IDs:

https://github.com/enasequence/ena-content-dataflow/blob/master/classifications/pathogen_taxonomy/Priority_pathogens_taxonomy.csv [accessed 19/09/2023]

²⁷ WHO publication: <https://www.who.int/publications/i/item/WHO-EMP-IAU-2017.12> [accessed 5/09/2023]

²⁸ WHO - One Health: <https://www.who.int/news-room/questions-and-answers/item/one-health> [accessed 03/09/2023]



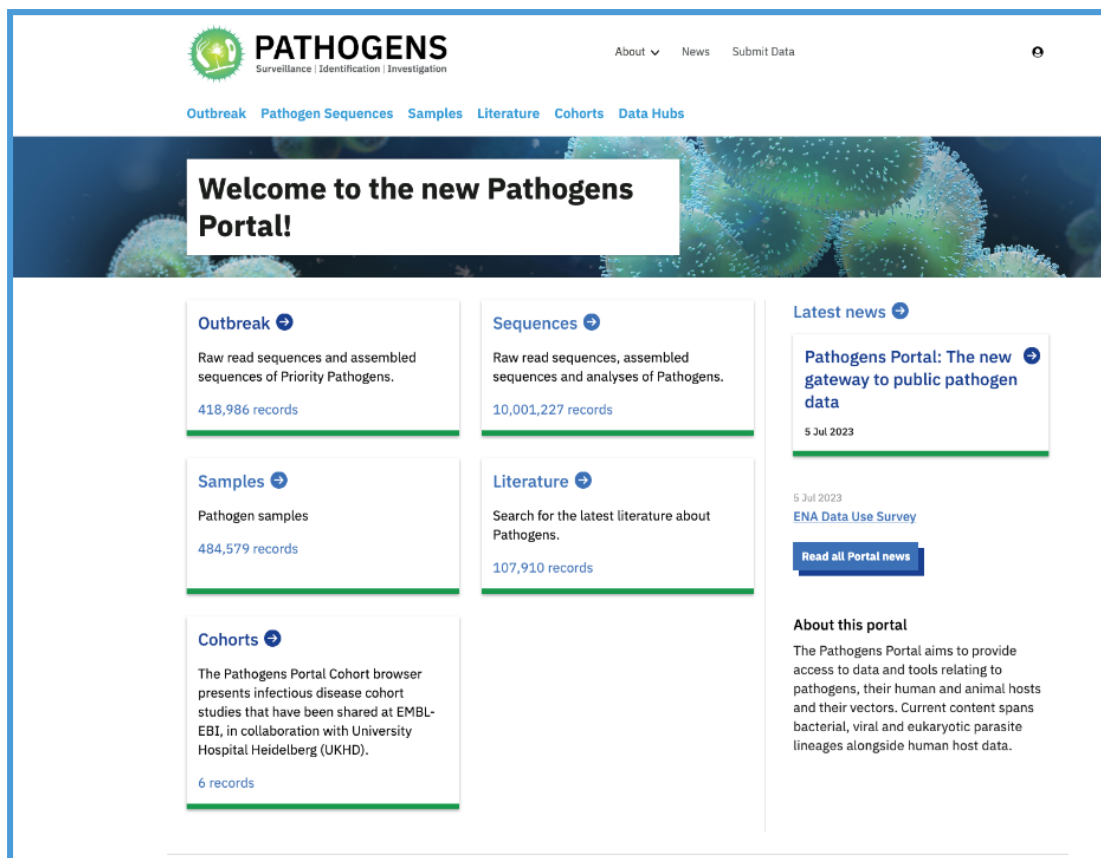


Figure 4. Pathogens Portal home page.

5. Results and discussion

We have added a broad range of new metadata resources to the COVID-19 Data Portal, extending its scope beyond its original biomolecular focus, to include chemical screening, large-scale population health studies, social sciences, and a large number of related resources. Building on the metadata standards work in D3.1, metadata have been integrated from their original representation in resource-specific formats, FAIRsharing, and community standard FAIR Data Points and DCAT, demonstrating the flexibility of the underlying indexing platform.

We have also established workflows for the routine update of the metadata from external sources with the CESSDA use case, and implemented the concept of “promoting” resources from resource-level indexing (Tier 3) to record-level indexing (Tier 2), for both IDTk and EU-OPENSREEN. The “Global Search” provides an entry point for metadata discoverability across the entire spectrum of COVID-19 Data Portal domains.

Building on source code, strategies, and experience from the COVID-19 Data Portal, we have initiated the transition to general pandemic preparedness, and have developed the Pathogens Portal, released July 5, 2023. Similar to the early stages of the COVID-19 Data

Portal, the Pathogens Portal is currently focussed on biomolecular, and mainly sequence data, but will evolve towards the broad domain coverage now implemented in the COVID-19 Data Portal. Due to the often fragmented nature of pathogen research, as well as ongoing discussion on equitable data access, integration of additional data sources may be as challenging as it was for COVID-19, but we are now building on strong experience both in technical terms, and from a community building perspective.

6. Next steps

To ensure smooth integration of partner-provided metadata, we anticipate re-running our “Discoverability hackathon” in the future and will continue to evolve our metadata format and presentation of the COVID-19 Data Portal. We anticipate significant development and metadata modelling work for the use-case driven support of complex data sources in close collaboration with WPs 2, 4 and 5.

Over the course of the project, emerging national data portals will be registered in FAIRsharing and linked to the COVID-19 Data Platform, building a federated digital space for infectious disease data. Work towards integrating open research data from the ISIDORE²⁹ project has been initiated to prepare for efficient integration of ISIDORE data as it becomes available [4].

As discussed in the previous section, a key effort now is the transition from COVID-19 to broader pathogen research and pandemic preparedness, and thus the recently released Pathogens Portal will gradually move to be the focus of development and data integration. A follow-up deliverable D3.4 is due in July 2024.

The project also addresses analysis transparency, sharing and trusted exchange to support reproducibility (Task 4.3) and usage indicators as well as attribution and credit for data submitters and workflow developers (Task 3.5).

7. Impact

Making a range of infectious disease data sources widely discoverable, accessible and interoperable is important for research and innovation, which is increasingly multidisciplinary in nature. For example, pathogen research is accelerated by the availability of data from clinical trials, biobanks, behavioural and socioeconomic studies, particularly if the data is combined with host and pathogen omics information. Multidisciplinary data is also critical for public health decision-making, where policy

²⁹ ISIDORE project website: <https://isidore-project.eu/> [accessed 29/08/2023]



questions are complex and evidence from biomolecular research, clinical studies and social sciences must be taken into account.

One lesson from the COVID-19 pandemic is that data-driven decision-making needs high quality, real-time data from many research disciplines and geographic areas in an integrated format. The BY-COVID project is building on these lessons and creating solutions for COVID-19 that can be extended to other pathogens. Resources like the COVID-19 Data Portal and FAIRsharing are central to meet these goals.

The flexible, tiered metadata discovery system across different domains, metadata standards, and maturity/robustness levels of data sources, enables the linking of FAIR data and metadata on SARS-CoV-2 and COVID-19, on other related viruses and diseases, and on socio-economic consequences, across research fields, from omics, clinical, and epidemiological research, to social sciences and humanities. This has the potential to accelerate infectious disease research, surveillance and outbreak investigation.

Since its launch, the COVID-19 Data Portal has been accessed by almost 300,000 users in 187 countries and geographical areas.

The strategies that were adopted, and the experience gained, in the development of the COVID-19 data portal have been integral to the establishment of Pathogens Portal³⁰, and provide a data integration framework that supports ongoing pathogen research, as well as a platform that is ready to be scaled up rapidly in the case of any future epidemic or pandemic.

³⁰ Pathogens Portal <https://www.pathogensportal.org/> [accessed 15/09/2023]

8. References

1. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3: 160018.
2. Madeira F, Pearce M, Tivey ARN, Basutkar P, Lee J, Edbali O, et al. Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res*. 2022. doi:10.1093/nar/gkac240
3. Lawson J, Cabili MN, Kerry G, et al. The Data Use Ontology to streamline responsible access to human biomedical datasets. *Cell Genomics*. 2021 Nov;1(2):None. DOI: 10.1016/j.xgen.2021.100028.
4. David R, Richard AS, Connellan C, et al., (2023b). Umbrella Data Management Plans to integrate FAIR data : lessons from the ISIDORE and BY-COVID consortia for pandemic preparedness. *Data Science Journal*. IN PRESS. Preprint: <https://doi.org/10.5281/zenodo.7998443>.

