BY-COVID D3.2: Implementation of cloud-based, high performance, scalable indexing system.

Henning Hermjakob, Mari Kleemola, Katja Moilanen, Markus Tuominen, Susanna-Assunta Sansone, Allyson Lister, Romain David, Maria Panagiotopoulou, Christian Ohmann, Jeroen Belien, et al.

Deliverable D3.2

Implementation of cloud-based, high performance, scalable indexing system.

Project Title (grant agreement No)	Beyond 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	30/09/2022	**Actual Delivery date**		30/09/2022
Delayed	No			
Partner(s) contributing to this deliverable	ELIXIR Hub, EMBL-EBI, CESSDA/TAU-FSD, ERINHA, UOXF, ECRIN, ELIXIR-NL/VUmc, ELIXIR-NL/Lygature, ELIXIR-UK / UNIMAN			
Authors	Henning Hermjakob (EMBL-EBI) Orcid: 0000-0001-8479-0262 Mari Kleemola (CESSDA/TAU-FSD) Orcid: 0000-0001-8855-5075 Katja Moilanen (CESSDA/TAU-FSD) Orcid: 0000-0002-7668-5427 Markus Tuominen (CESSDA/TAU-FSD) Orcid: 0000-0002-3092-1690			

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

Contributors	Susanna-Assunta Sansone (UOXF) Orcid: 0000-0001-5306-5690) Allyson Lister (UOXF) Orcid: 0000-0002-7702-4495 Romain David (ERINHA) Orcid: 0000-0003-4073-7456 Maria Panagiotopoulou (ECRIN) Orcid: 0000-0002-4221-7254 Christian Ohmann (ECRIN) Orcid: 0000-0002-5919-1003 Jeroen Belien (ELIXIR-NL/VUmc) Orcid: 0000-0002-7160-5942 Julia Lischke (ELIXIR-NL/Lygature) Orcid: 0000-0002-5524-2838 Nick Juty (ELIXIR-UK/UNIMAN) Orcid: 0000-0002-2036-8350 Stian Soiland-Reyes (ELIXIR-UK/UNIMAN) Orcid: 0000-0001-9842-9718
Acknowledgements (not grant participants)	NA
Reviewers	Project Management Board

Log of changes

Date	Mvm	Who	Description
2022-08-21		Henning Hermjakob (EMBL-EBI)	Initial structure
2022-08-24		Susanna Assunta-Sansone (U Oxford)	FAIRSharing update
2022-08-25		Romain David, (ERINHA)	Review and adds, comments
2022-09-12		Mari Kleemola, Katja Moilanen, Markus Tuominen (CESSDA/TAU-FSD)	CESSDA case study, general comments and edits
2022-09-15		Henning Hermjakob (EMBL-EBI)	Final draft for internal review

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

Table of contents

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

1. Executive Summary

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 will enable 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.

Building on the metadata format developed in D3.1, in a series of work package meetings and a workshop, with participation from all other work packages, we have developed tools (example Validator), workflows (Semi-automated transfer of resource metadata from FAIRsharing to Covid-19 portal), and documentation and training (Training: Discoverability hackathon) to support the efficient integration of additional resources from a broad range of domains into the COVID-19 Data Portal as well as improved the end user facing COVID-19 Data Portal itself (Global search).

This work establishes the basis for the further development of the COVID-19 Data Portal metadata discovery, and provides a path for integration of metadata from multi-domain partners in BY-COVID, as well as our ISIDORe sibling project[1], and other relevant external resources. 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.

[1] https://isidore-project.eu/

Funded by the European Union

BY-COVID is funded by the European Union’s Horizon Europe research and innovation programme under grant agreement number 101046203.

2. Contribution towards project objectives

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

Table 1: Contribution towards project objectives.

	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.	No
	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.	No
	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 Link FAIR data and metadata	1. A platform that links normative pathogen genomes and variant representations to research cohorts and mechanistic	Yes

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

on SARS-CoV-2 and COVID-19	studies to understand the biomolecular determinants of 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 *Data Hubs* 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.	No

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

3. Introduction

BeYond-COVID (BY-COVID) aims to provide comprehensive open data on SARS-CoV-2 and other infectious diseases across scientific, medical, public health and policy domains. The project will mobilise existing data resources (i.e catalogues) and marshal the resources for research, connect and expose the data and resources via the COVID-19 Data Portal, and drive use and analysis by connecting workflows, national portals and analysis environments. Running for three years, the project brings together 53 partners from 19 countries.

This deliverable is part of BY-COVID Work Package 3 that is focussed 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 will enable 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. This has the potential to accelerate infectious disease research, surveillance and outbreak investigation.

Our inter-domain metadata mapping (Task 3.1, D3.1 Metadata standards[2]) supports data discovery, access, and analysis across fields from molecular biology to social sciences. The harmonised metadata will be discoverable through the central index (Task 3.2 - this deliverable) and web portal (Task 3.3), but also openly accessible for third party applications through web services. The project also addresses analysis transparency, sharing and trusted exchange to support reproducibility (Task 4.3) and attribution and credit for data submitters and workflow developers (Task 3.5). Here, we describe the implementation of the discoverability concept for the COVID-19 Data Portal (https://www.covid19dataportal.org/), based on the EBI Search system, FAIRsharing, and a three-tiered indexing concept.

3.1 EBI Search

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). As of August 2022, it indexes more than 5.5 billion data objects from 172 datasets, mostly from within EMBL-EBI, but also from international

[2] Hermjakob, Hennig, Kleemola, Mari, Moilanen, Katja, Sansone, Susanna-Assunta, Lister, Allyson, David, Romain, Panagiotopoulou, Maria, Ohmann, Christian, Bellen, Jeroen, Lischke, Julia, Juty, Nick, & Soiland-Reyes, Stian. (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

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

partners. EBI Search responds to more than 50 million queries/month. Indexed data resources include nucleotide and protein sequences at both the genomic and proteomic levels; structures ranging from chemicals to macro-molecular complexes; gene-expression experiments; binary level molecular interactions as well as reaction maps and pathway models; functional classifications; biological ontologies; diseases; and comprehensive literature libraries covering the biomedical sciences and related intellectual property.

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 pipe-lines.

Conceptually, EBI Search offers "Search as a Service"; data sources provide their metadata in a file in a specific file format and notify the EBI Search team of the file location. In a nightly check, if the file is found to be modified, it will be re-indexed. The search platform is built on the Lucene Core library with an API that provides access to data sets which, combined, contain more than 5.5 billion entries. It is built on the SpringBoot framework, and runs on a Kubernetes cluster to enable optimal scalability. The API-based query interface is available both within and outside EMBL-EBI, and can be used to develop website-specific search/results pages. The API provides simple and advanced search options, result summaries, faceting support, and a configurable display of search results. The full API documentation is available at https://www.ebi.ac.uk/ebisearch/apidoc.ebi. EBI Search has been used to power search across the resources in the COVID-19 Data Portal since its inception; experience and source code are also used in the context of pandemic preparedness to continuously update a "Baseline Portal" library, which can be used to rapidly instantiate new portal sites, both for potential new pathogens of interest, or other contexts requiring cross-resource data integration. Both the Early Cause portal and Pathogens Portal have been developed with the Baseline Portal as a starting point.

3.2 FAIRsharing

FAIRsharing 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 2022, it serves 1606 standards, 1902 databases and 157 policies, citable via DOI; many records are also maintained by the individuals and/or organisations (behind the resource) that are identified via their ORCID and ROR, respectively. FAIRsharing defines the indicators necessary to monitor the development, evolution and integration of standards, as well as their implementation and use in databases, and adoption in data policies by funders, journals and all concerned

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

organisations. Adopted by major scholarly publishers, funders and other stakeholders, and an endorsed output of the RDA, FAIRsharing guides consumers to discover, select and use these resources with confidence and enables producers to make their resources more findable, adopted and cited. Recently, a RDA/EOSC-Future ambassadorship grant has been awarded to Allyson Lister, FAIRsharing Content and Community Coordinator, to launch the FAIRsharing Community Curation Programme to build a network of community curators that will be sustained for the long term.

3.3 Three-tiered approach

Exposing and effective connection and linking of different data types requires indexing based on cross-mapped metadata. Indexing and incorporation of metadata into the COVID-19 Data Portal proceeds via a flexible, tiered system for metadata integration (Table 2).

- For tier 1, a limited number of key resources will be deeply indexed through EBI Search, capturing granular, record level identifiers and detailed metadata. For these resources, the indexing strategy will support complex interoperability tasks implemented in WP2 and WP5.
- In tier 2, a broader range of resources will be indexed through EBI Search with a focus on record level discoverability, with coarse-grained metadata, and limited but iteratively refined metadata harmonisation. This tier will support deep discovery of relevant datasets from large resources in the multi-disciplinary COVID-19 space.
- In tier 3, additional resources will be included in the COVID-19 Data Portal at the resource level only, supporting high level discovery of relevant resources, but delegating record level searches to the relevant resources themselves. Tier 3 will allow us to support discovery of a broad range of relevant resources, while avoiding complex metadata harmonisation challenges for a number of resources beyond the feasible scope of the project.

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.

To minimise repeated requests to collaborators and “Federation Fatigue”, we are co-ordinating requests for metadata with WP2, and centralise requests for resource level (tier 3) metadata in FAIRsharing. FAIRsharing will complement 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 will be captured in FAIRsharing records and selectively imported to the COVID-19 Data Portal through the API.

Funded by the European Union

BY-COVID is funded by the European Union’s Horizon Europe research and innovation programme under grant agreement number 101046203.

Table 2: 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	
Tier 3	Focus on resource level discoverability.	

4. Description of Work

Tier 3: Resource level discoverability

FAIRsharing collection

Whilst the COVID-19 Data Portal indexes the data sources and metadata about datasets; FAIRsharing is the *de facto* BY-COVID data source catalogue as it will provide the description of and relationships among the data sources (databases, knowledge bases, repositories) and the standards they use. A FAIRsharing BY-COVID Collection has been created in collaboration with WP2; this will link the description of the data sources to their datasets indexed in the COVID-19 Data Portal.

FAIRsharing will store all descriptors that the COVID-19 Data Portal requires of its data sources prior to indexing. When a provider submits data sources to be indexed by the COVID-19 Data Portal, the data portal directs them to FAIRsharing where they will describe their record and ensure all required metadata is provided. This metadata will be pulled, as required, from FAIRsharing when the COVID-19 Data Portal is ready to index it.

Based on the BY-COVID Collection, FAIRsharing works to ensure these resources are surfaced in the EOSC ecosystem via its collaboration with OpenAIRE. FAIRsharing records from the OpenAIRE Research Graph will help OpenAIRE users who wish to discover

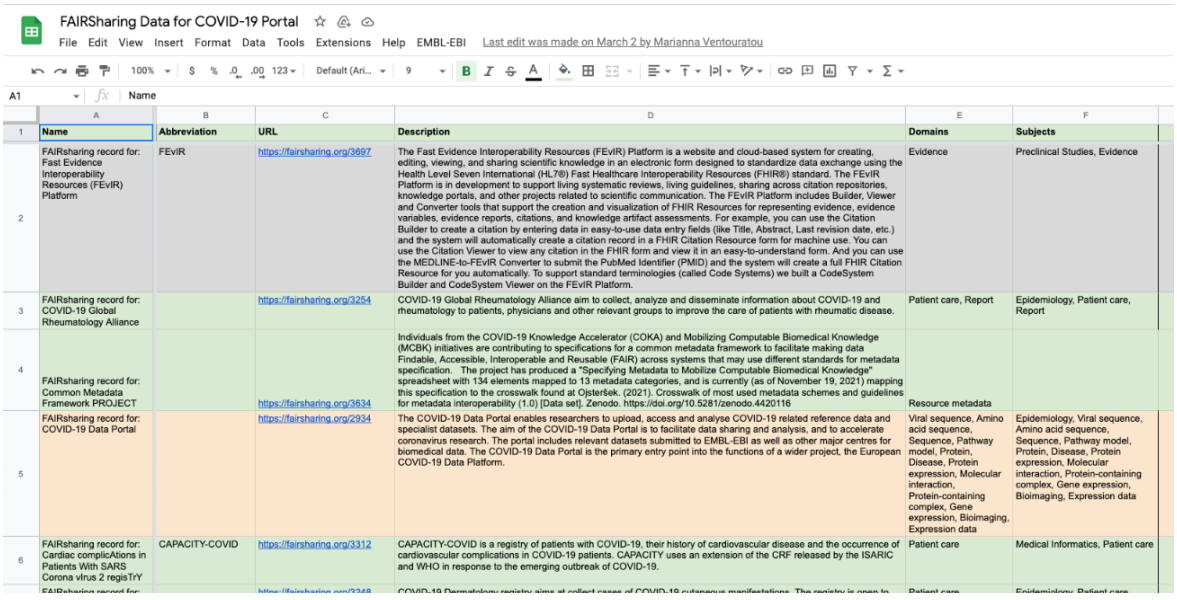
Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

relationships to their standard of interest, alongside the repositories that implement that standard and the data policies (from journal publishers, funders and other organisations) that endorse their use.

Semi-automated transfer of resource metadata from FAIRsharing to the COVID-19 Data portal

The COVID-19 Data Portal links to the FAIRsharing collection for detailed exploration of relevant standards, formats, etc. However, we also integrate a relevant subset of those resources for resource level discoverability directly into the portal. To enable an efficient workflow and avoid requesting the same data from resources more than once, we have developed a Google spreadsheet with integrated calls to the FAIRsharing API, allowing rapid, but curated import of FAIRsharing collection data into the COVID-19 Data Portal (Fig 1). The spreadsheet allows automated updates from FAIRsharing, followed by manual curation, for example to identify duplicates with existing resources listed in the portal. After preprocessing, resources are currently manually transferred into the Content Management System (CMS) of the COVID-19 Data Portal. The next step will be to automate this workflow to enable a direct export from the spreadsheet into the CMS. From the CMS, the records are automatically harvested for indexing, and are thus available through the global search of the COVID-19 Data Portal.


FAIRSharing Data for COVID-19 Portal

File Edit View Insert Format Data Tools Extensions Help EMBL-EBI Last edit was made on March 2 by Marianna Ventouratou

	Name	Abbreviation	URL	Description	Domains	Subjects
2	FAIRsharing record for: Fast Evidence Interoperability Resources (FEvIR) Platform	FEvIR	https://fairsharing.org/3697	The Fast Evidence Interoperability Resources (FEvIR) Platform is a website and cloud-based system for creating, editing, viewing, and sharing scientific knowledge in an electronic form designed to standardize data exchange using the Health Level Seven International (HL7®) Fast Healthcare Interoperability Resources (FHIR®) standard. The FEvIR Platform is in development to support living systematic reviews, living guidelines, sharing across citation repositories, knowledge portals, and other projects related to scientific communication. The FEvIR Platform includes Builder, Viewer and Converter tools that support the creation and visualization of FHIR Resources for representing evidence, evidence variables, evidence reports, citations, and knowledge artifact assessments. For example, you can use the Citation Builder to create a citation by entering data in easy-to-use data entry fields (like Title, Abstract, Last revision date, etc.) and the system will automatically create a citation record in a FHIR Citation Resource form for machine use. You can use the Citation Viewer to view any citation in the FHIR form and view it in an easy-to-understand form. And you can use the MEDLINE-to-FEvIR Converter to submit the PubMed Identifier (PMID) and the system will create a full FHIR Citation Resource for you automatically. To support standard terminologies (called Code Systems) we built a CodeSystem Builder and CodeSystem Viewer on the FEvIR Platform.	Evidence	Preclinical Studies, Evidence
3	FAIRsharing record for: COVID-19 Global Rheumatology Alliance		https://fairsharing.org/3254	COVID-19 Global Rheumatology Alliance aim to collect, analyze and disseminate information about COVID-19 and rheumatology to patients, physicians and other relevant groups to improve the care of patients with rheumatic disease.	Patient care, Report	Epidemiology, Patient care, Report
4	FAIRsharing record for: Common Metadata Framework PROJECT		https://fairsharing.org/3634	Individuals from the COVID-19 Knowledge Accelerator (COKA) and Mobilizing Computable Biomedical Knowledge (MCBK) initiatives are contributing to specifications for a common metadata framework to facilitate making data Findable, Accessible, Interoperable and Reusable (FAIR) across systems that may use different standards for metadata specification. The project has produced a "Specifying Metadata to Mobilize Computable Biomedical Knowledge" spreadsheet with 134 elements mapped to 13 metadata categories, and is currently (as of November 19, 2021) mapping this specification to the crosswalk found at Ojsteršek. (2021). Crosswalk of most used metadata schemes and guidelines for metadata interoperability (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.4420116	Resource metadata	
5	FAIRsharing record for: COVID-19 Data Portal		https://fairsharing.org/2934	The COVID-19 Data Portal enables researchers to upload, access and analyse COVID-19 related reference data and specialist datasets. The aim of the COVID-19 Data Portal is to facilitate data sharing and analysis, and to accelerate coronavirus research. The portal includes relevant datasets submitted to EMBL-EBI as well as other major centres for biomedical data. The COVID-19 Data Portal is the primary entry point into the functions of a wider project, the European COVID-19 Data Platform.	Viral sequence, Amino acid sequence, Sequence, Pathway model, Protein, Disease, Protein expression, Molecular interaction, Protein-containing complex, Gene expression, Bioimaging, Expression data	Epidemiology, Viral sequence, Amino acid sequence, Sequence, Pathway model, Protein, Disease, Protein expression, Molecular interaction, Protein-containing complex, Gene expression, Bioimaging, Expression data
6	FAIRsharing record for: Cardiac complicAtions in Patients With SARS Corona virus 2 regisTrY	CAPACITY-COVID	https://fairsharing.org/3312	CAPACITY-COVID is a registry of patients with COVID-19, their history of cardiovascular disease and the occurrence of cardiovascular complications in COVID-19 patients. CAPACITY uses an extension of the CRF released by the ISARIC and WHO in response to the emerging outbreak of COVID-19.	Patient care	Medical Informatics, Patient care

Figure 1: Processing spreadsheet for integration of "Related resources" records into the COVID-19 portal. Colour codings are manual classifications of redundant records.

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

Tier 2: Record-level discoverability

For Tier 2 resources, we aim to implement record level discovery, meaning that for each resource, multiple data objects, from a few to potentially tens of thousands, are discoverable via their metadata. This requires that the resource provides an XML file with metadata for both the overall resource, and the individual records. The file format follows the established EBI Search metadata format, with the attributes defined in the context of T3.1. To support the efficient implementation of the metadata file by data sources, we have developed documentation, a simple validator, training, and a case study based on CESSDA ERIC data.

Validator

We have modified an existing validation tool to provide a simple validator for BY-COVID formatted files (https://www.omicsdi.org/validate), supporting interactive validation of files in the development phase. The validator allows uploading a file, and supports two reporting levels, either errors only, or errors and warnings. In "errors" mode, only mandatory fields, as defined by the WP3 metadata format spreadsheet will be checked. In "errors and warnings" mode, recommended fields will be checked in addition.

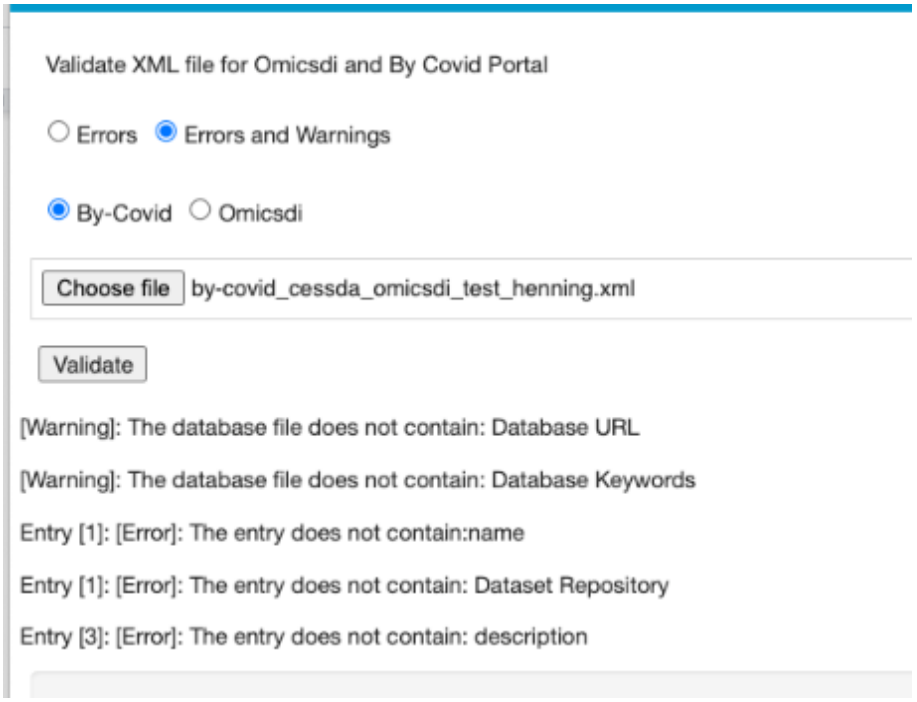


Figure 2: BY-COVID validator test output.

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

Training: Discoverability hackathon

We anticipate addition of metadata to the COVID-19 Data Portal from a significant number of additional resources; while the required metadata format is quite straightforward, it requires data providers to develop new or modified metadata exporters. To support efficient development of the required exporters, WP3 and WP6 jointly organised a well-attended "Discoverability Hackathon" with 37 participants on August 2nd, 2022. The hackathon provided practical training on FAIRsharing registration and the implementation of the BY-COVID metadata file, as well as an opportunity for in-depth discussion on the anticipated workflow for discoverability of external data sources. We expect that this hackathon will be re-run, potentially multiple times, in updated form, to support the needs of data providers and their schedules.

Slides, notes, and recording are at https://drive.google.com/drive/u/1/folders/1YfCyIhqhJWpVC-kx_mKKz1CtSiDAkEOB

CESSDA case study

The Consortium of European Social Science Data Archives (CESSDA) ERIC provides large-scale, integrated and sustainable data services to the social sciences. It brings together social science data archives across Europe, with the aim of promoting the results of social science research and supporting national and international research and cooperation. The CESSDA Data Catalogue (CDC) contains metadata of 40,000+ data collections held by the European social science data archives. The data may be quantitative, qualitative or mixed-modes data, cross-sectional or longitudinal, recently collected or historical data. The study descriptions follow the CESSDA Data Catalogue DDI profiles which are based on the CESSDA Metadata Model, a subset of the Data Documentation Initiative (DDI) metadata standard.

Onboarding the CESSDA metadata in the COVID-19 Data Portal has been done in collaboration with Task 2.4. The process has been iterative and parallel with the development of the portal and tooling (like the validator), which has provided synergies on both sides. In the process, WP3 mapped the CESSDA metadata with the OmicsDI format required by the COVID-19 Data Portal, and also some additional metadata fields were added for social sciences. Task 2.4 automated the creation of the OmicsDI xml file from CESSDA, utilising CESSDA's OAI-PMH endpoint to first harvest part of the metadata of all studies into DSpace to query for relevant studies and then harvesting full metadata of those specific studies. Transforming metadata using an XSLT file was fairly easy when harvesting from CESSDA's OAI-PMH endpoint in DDI-Codebook 2.5 format since it contains high-quality metadata for all the mandatory fields and also many of the recommended fields of BY-COVID OmicsDI format.

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

At the time of writing, 450+ social science studies from the CDC are included in the COVID-19 Data Portal. The next steps will include improving the query for metadata harvested from CESSDA to include more studies and adding metadata from EUI's (European University Institute) COVID-19 SSH Data Portal. EUI also has an OAI-PMH endpoint so the same OmicsDI xml file creation process can be used but a new XSLT file will have to be created for DataCite as DDI-Codebook 2.5 is not available.

Tier 1: Fine-grained support for interoperability use cases

In Tier 1, we intend to index data sources providing rich metadata records to support complex use cases driven by WP4 (Workflows) and WP5 (Use cases). Complex metadata models and the required interoperability between the metadata model from different resources will not be feasible for all resources, we expect a limited number or resources to be classified as Tier 1 based on use cases. While these use cases are still being developed by partner WPs, the flexible and extensible metadata framework we have developed in T3.1/D3.1 already provides support for the anticipated rich metadata sets, and the EBI Search system has a wealth of existing resources with complex metadata sets, for example for COVID-19 Sequences (embl-covid19), with 6 million data objects, each with up to 47 indexed metadata attributes (including 13 cross-references to other data sets). EBI Search also provides both a streaming download interface and deep pagination, supporting API-based access to very large result sets of more than 1 million entries. Thus the infrastructure is in place for efficient discovery, both interactively and workflow-based, of relevant data objects based on both large and complex metadata sets.

Global search

At the start of BY-COVID, the COVID-19 Data Portal provided access to mainly biomolecular data across nine categories, from "Viral Sequences" to "Literature", see figure 3 for a screenshot from June 2022. BY-COVID aims to increase the scope and thematic diversity of the portal through integration of additional data sources from domains like clinical and social sciences. This will extend the range of categories in the portal, and provide the opportunity for serendipitous discoveries of relevant data from other domains. However, to enable such discoveries, the portal needs to provide a global search interface, in addition to the existing category-specific searches. This will allow researchers to discover potentially relevant search results from outside their field, and allow them to efficiently access these new, potentially unfamiliar resources.

We have implemented the new global search feature based on the existing EBI Search API, adding a new global search box prominently on the page, and providing faceted search results to support efficient navigation, see figure 4.

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

Global search requires the merging of search results across multiple EBI Search domains.

These are queried in parallel to achieve a faster state of interactivity. An initial concern was raised with regards to the amount of requests to retrieve these results; hence we have performance tested two distinct implementations, one merging results on the client side and one on the server side. These tests did not detect major differences. Therefore the production implementation is based on the client-side merging of results, achieving interactive response times of ca. 2.4 seconds from query to the first interactive page.

The global search feature was designed taking into account accessibility and general best practices for search design. For example, the search allows for both faceted and categorised search, which can help users build more specific queries[4]. Previous user experience research conducted for EBI search and Open Targets informed the design decisions for the COVID-19 Data Portal Global Search, as it performs in a similar context[5]. It has been "silently" released on 2022-08-18, followed up by a news item on 2022-08-23[3].

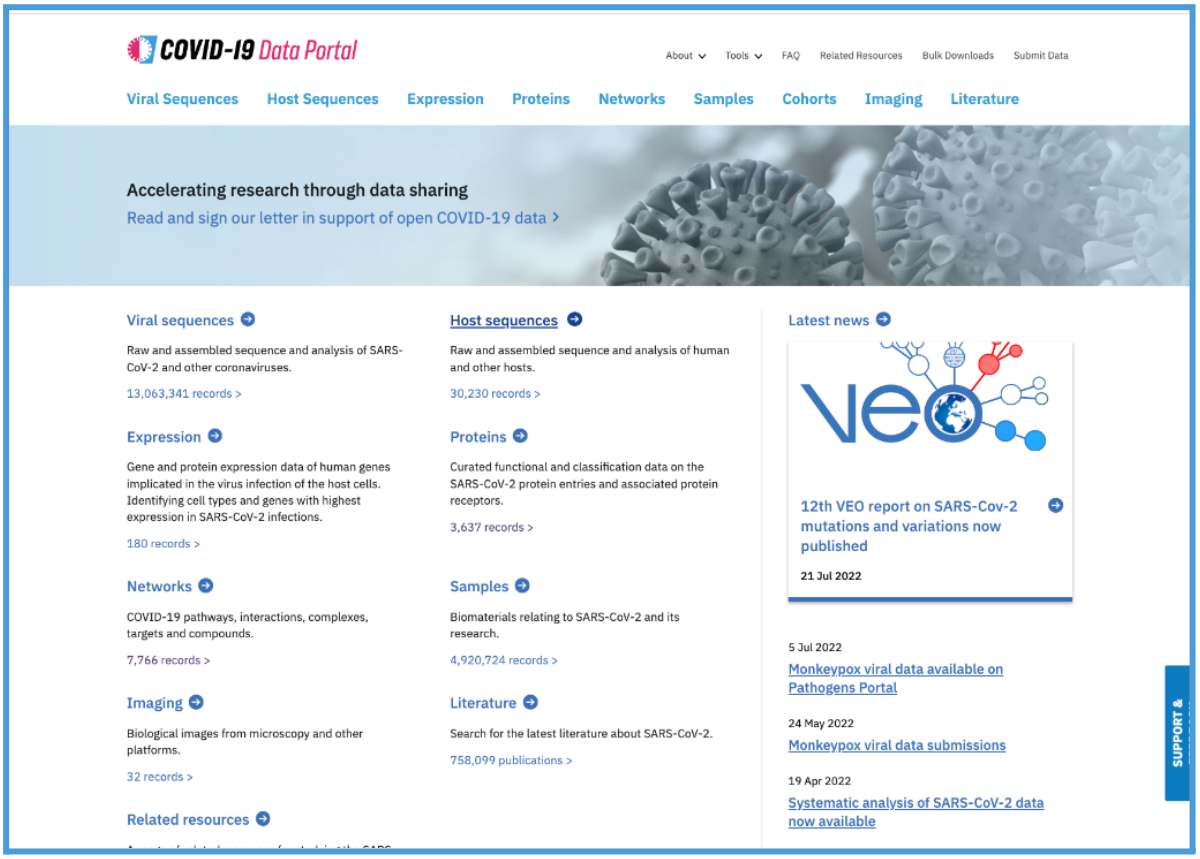


Figure 3: COVID-19 Data Portal major categories in June 2022.

3 https://www.covid19dataportal.org/news/global-search-now-available-on-the-covid-19-portal?nid=2082

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

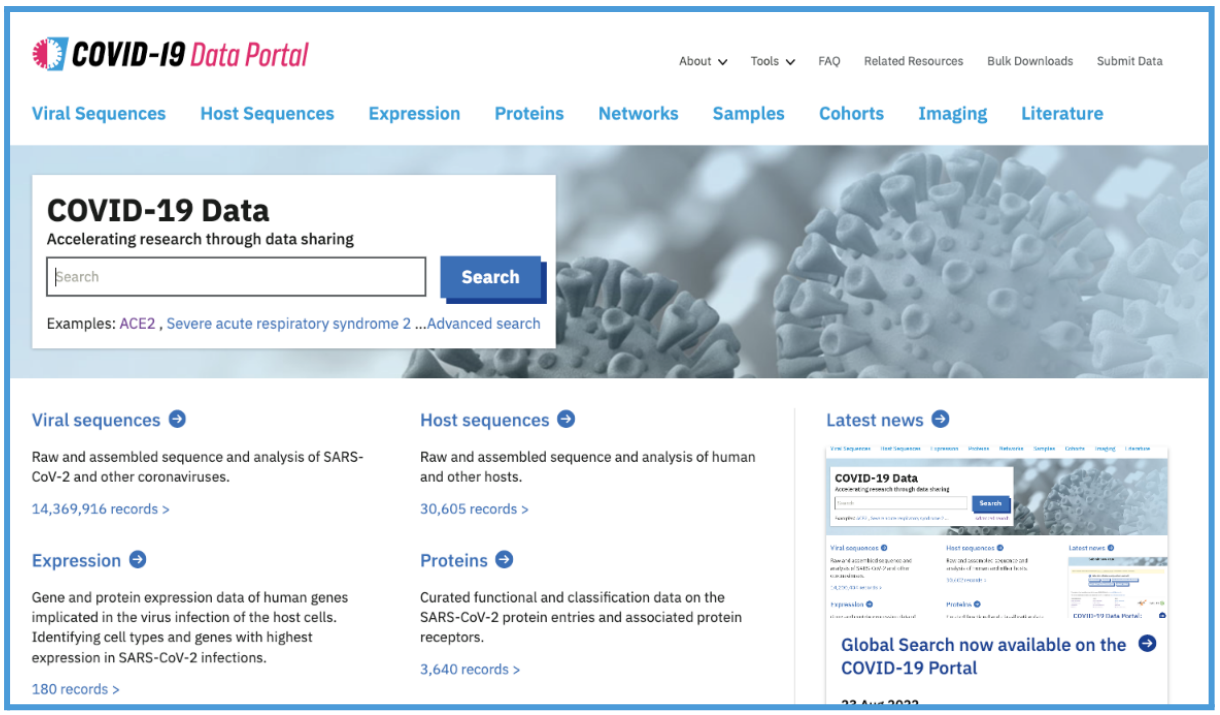


Figure 4: Global search added to COVID-19 Data Portal in August 2022.

5. Results and Discussion

Based on the metadata model developed in D3.1, we have implemented the infrastructure to support the three-tiered discoverability concept in the COVID-19 Data Portal. This infrastructure allows the indexing of data sources from a very simple, resource level only metadata (Tier 3) via coarse-grained record-level indexing (Tier 2) to complex, fine grained metadata (Tier 1). We have developed tools (example Validator), workflows (Semi-automated transfer of resource metadata from FAIRsharing to Covid-19 portal), and documentation and training (Training: Discoverability hackathon) to support the efficient integration of additional resources from a broad range of domains into the COVID-19 Data Portal as well as improved the end user facing COVID-19 Data Portal itself (Global search).

As of August 2022, we have only added CESSDA as a new Tier 2 resource, but we have the infrastructure in place to efficiently add additional resources across all Tiers, from BY-COVID partners, our sibling project ISIDORe (https://isidore-project.eu/), and third parties, as they become available over the course of the project.

Through the "Baseline Portal" software package, the source code and experience from this work is propagated back to existing related portals Early Cause and Pathogens.

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

6. Next Steps

As additional data sources will be added over the course of the project, we will continuously update and refine the discovery infrastructure to ensure stable metadata update and indexing workflows as well as a user-friendly interface for web and API access. We anticipate re-running our Discoverability hackathon in collaboration with WP6, and working with (meta)data providers to support them in practical formatting questions, as well as updates to the BY-COVID metadata model to support the evolving needs of project partners. While the developed approach for tiers 2 and 3 is expected to remain relatively stable, we anticipate significant development and metadata modelling work for the use-case driven support of Tier 1 resources, in close collaboration with WPs 2, 4 and 5.

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.

7. 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

4. Cooper, A. (2014). About Face : the essentials of interaction design. (pp. 579–590). Wiley ISBN 9781118766576.

5. Karamanis, N., *et al.* (2018). Designing an intuitive web application for drug discovery scientists. *Drug Discovery Today*, 23(6), 1169–1174. https://doi.org/10.1016/j.drudis.2018.01.032

Funded by the European Union

BY-COVID is funded by the European Union's Horizon Europe research and innovation programme under grant agreement number 101046203.