

## Deliverable D2.4

*Report on data sources discovery and integration for enabling data use and re-use in response to future outbreaks*

<b>Project Title</b> (grant agreement No)	BeYond-COVID Grant Agreement 101046203		
<b>Project Acronym</b> (EC Call)	BY-COVID		
<b>WP No &amp; Title</b>	WP2: Accessing heterogeneous data across domains and jurisdictions for enabling the downstream processing of COVID-19 and future pandemic episodes data		
<b>WP Leaders</b>	Alfonso Valencia [BSC] Salvador Capella-Gutierrez [BSC] Antje Keppler [EuroBioImaging] Aastha Mathur [EuroBioImaging]		
<b>Deliverable Lead Beneficiary</b>	Barcelona Supercomputing Center [BSC]		
<b>Contractual delivery date</b>	30/09/2024	<b>Actual Delivery date</b>	30/09/2024
<b>Delayed</b>	No		
<b>Partner(s)</b> contributing to this deliverable			
<b>Authors</b>	Laura Portell-Silva (BSC)		
<b>Contributors</b>	Vasso Kalaitzi (KNAW-DANS) Simon Saldner (KNAW-DANS) Markus Tuominen (TAU-FSD) Matti Heinonen (TAU-FSD) Maria Panagiotopoulou (ECRIN) Sergio Contrino (ECRIN) Kostis Alexandrakis (EKKE) Dimitra Kondyli (EKKE) Simone Sacchi (EUI) Robin Navest (Lygature) Jeroen Belien (VUmc) Isabel Kemmer (Euro-BioImaging)		





<b>Acknowledgements</b> (not grant participants)	Enrique Bernal-Delgado [IACS] Nina Van Goethem [Sciensano] Reagon Karki (Fraunhofer ITMP/EU-OS)  N/A
<b>Reviewers</b>	Nadim Rahman (EMBL-EBI) Dorothea Dörr (Euro-BioImaging) Ilaria Colussi (BBMRI-ERIC) Louiza Kalokairinou (ELIXIR)



## Log of changes

Date	Mvm	Who	Description
09/07/2024	v0.1	Laura Portell-Silva	First draft sent to WP2 members to add content
26/08/2024	v0.2	All	End of adding content
09/09/2024	v1.0	Laura Portell-Silva	Sent to WP2 leads and reviewers
16/09/2024	v1.1	Laura Portell-Silva	Comments from reviewers addressed and sent to review from coordination
30/09/2024	v1.2	Laura Portell-Silva	Comments from coordination and ELSI team addressed



### 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*

<b>1. Executive Summary</b>	<b>5</b>
<b>2. Contribution towards project objectives</b>	<b>6</b>
Objective 1	6
Objective 2	6
Objective 3	6
Objective 4	7
Objective 5	7
<b>3. Introduction</b>	<b>8</b>
<b>4. Methods</b>	<b>8</b>
<b>5. Description of work accomplished</b>	<b>9</b>
5.1. Non-patient related data	9
5.1.1. Data resources	9
5.1.2. Data submission guidelines	9
5.1.3. Guided data mobilisation	10
5.1.4. Knowledge Graph Generator (KGG): A fully automated workflow for creating disease-specific KGs	11
5.2. Human clinical and health data	13
5.2.1. Bespoke CDM + CDM builder	13
5.2.2. Clinical research metadata	14
5.2.3. Harmonisation and mobilisation of COVID-NL data	18
5.3. Socioeconomic data	20
5.3.1. Harvesting tool and XML transformation	20
5.3.2. Data sources	21
5.3.2.1 Primary data sources	22
5.3.2.2 Secondary data sources	22
5.3.3 Cross-domain collaboration	23
5.3.4 Awareness raising	23
5.5. Network of Covid-19 Beacons	24
<b>6. Results and discussion</b>	<b>24</b>
6.1. Connection with other WP	25
6.1.1. Connection with WP3: FAIRsharing and the COVID-19 Data Portal	25
6.1.2. Connection with WP4: IDTk pages created	29
6.1.3. Connection with WP5: Lessons learned from the use cases	30
6.1.4. Connection with WP6: Training activities	31
<b>7. Conclusions</b>	<b>32</b>
<b>8. Next steps</b>	<b>33</b>
<b>9. Impact</b>	<b>33</b>



# 1. Executive Summary

In the face of rapid technological advancements and the challenges posed by global health crises like the COVID-19 pandemic, the ability to efficiently discover and integrate diverse data sources is crucial. This report explores innovative mechanisms for data discovery and integration, focusing on strategies beyond discovery at source methods (which are part of deliverable 2.3), such as the Beacon. While the Beacon provides foundational insights into direct data discovery, building a more robust and flexible data infrastructure requires exploring alternative approaches that enhance data use and re-use across various platforms and disciplines. These strategies aim to facilitate seamless data integration, enabling more effective and timely responses to future health emergencies.

The report concentrates on the discovery and integration of three primary categories of data: non-patient data, human clinical and health data, and socioeconomic data. Each data type presents unique challenges and opportunities for comprehensive outbreak responses. The report incorporates lessons learned from BY-COVID use cases, providing practical insights into the complexities of data discovery and integration. By integrating these real-world examples, the report ensures that its recommendations are not only theoretically sound but also practically viable and effective in real-world scenarios.

The connection between WP2 and other WPs within the BY-COVID project is also highlighted in the report. WP2 worked closely with WP3 to maintain the FAIRsharing collection for BY-COVID, which includes diverse data types such as non-patient data, human clinical data, and socioeconomic data. The collaboration extended to enhancing the COVID-19 Data Portal, integrating various datasets, including host sequences, imaging, and social sciences. These efforts involved developing tools for data mobilisation, ensuring that datasets are easily discoverable and usable within the portal.

Additionally, WP2 contributed to the Infectious Diseases Toolkit (IDTk) by developing dedicated pages for data sources and analysis, aligning closely with WP4's objectives. Lessons learned from the baseline use case in WP5 emphasised the importance of seamless data integration and mobilisation of sensitive health data, ensuring compliance with legal and ethical standards. WP2 also collaborated with WP6 on training activities, including workshops on using the Beacon tool for sensitive data discovery and FAIR bioimage data management. These collaborations across work packages underscore the project's commitment to enhancing data accessibility, interoperability, and reusability, ultimately supporting global health preparedness and response efforts.



## 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
<b>Objective 1</b> 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
<b>Objective 2</b> 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
<b>Objective 3</b> Link FAIR data and metadata on	1. A platform that links normative pathogen genomes and variant representations to research cohorts and mechanistic studies to understand the biomolecular determinants of variant response on patient susceptibility, and disease pathways.	No



SARS-CoV-2 and COVID-19	2. An open and extensible metadata framework adopted cross-domain that supports comprehensive indexing of the infectious disease sources based on mappings across sources 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 sources.	No
<b>Objective 4</b> Develop digital tools and data analytics for pandemic and outbreak preparedness, including tracking genomics variations of SARS-CoV-2 and identifying new variants of concern	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
<b>Objective 5</b> 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



### 3. Introduction

In the era of rapid technological advancements and an ever-increasing volume of data, the ability to efficiently discover and integrate diverse data sources is paramount, especially in the context of responding to health crises such as infectious disease outbreaks. The recent COVID-19 pandemic underscored the critical need for robust data systems that can support not only immediate response efforts but also long-term preparedness and resilience.

This report delves into innovative mechanisms for the discovery and integration of data sources, emphasising approaches beyond the "discovery at source", such as the Beacon. The latter, thoroughly examined in our previous deliverable D2.3, provides foundational insights and methodologies for data discovery directly at the data source. However, to build a more comprehensive and flexible data infrastructure capable of handling future outbreaks, it is essential to explore alternative discovery strategies that enhance data use and re-use across various platforms and disciplines.

By investigating these alternative mechanisms, we aim to identify and implement strategies that facilitate seamless data integration, thereby enabling more effective and timely responses to future health emergencies. This report will discuss the strengths and limitations of various discovery techniques, present the connections with the BY-COVID use cases, and offer recommendations to better support data use and re-use in response to future outbreaks.

### 4. Methods

This report has been collaboratively developed by the BY-COVID WP2 partners, with a specific focus on the discovery and integration of various data types included in BY-COVID. Our efforts concentrate on three primary categories of data: non-patient data, human clinical and health data, and socioeconomic data. Each of these data types presents unique challenges and opportunities for discovery and integration, which are critical for a comprehensive outbreak response.

In addition, in this report, we have added the lessons learned through the BY-COVID use cases. These real-world examples provide valuable insights into the practicalities and complexities of data discovery and integration, ensuring that our recommendations are grounded in actual experience. By integrating these lessons, we aim to ensure that our strategies are not only theoretically sound but also practically viable and effective in real-world scenarios.





## 5. Description of work accomplished

### 5.1. Non-patient related data

#### 5.1.1. Data resources

By their very nature, non-patient-related data can and should be made openly and publicly available for unrestricted access. This type of data is critical for basic and applied research, and the open availability of data allows researchers to easily share data and thus collaborate to better understand infectious diseases and develop preventive measures.

Task 2.1 identified several open repositories where different types of non-patient related research data can be shared (already described in D2.2 and in the BY-COVID WP2 List of Resources<sup>1</sup>). As non-patient-related data encompasses data from different scientific disciplines such as structural analysis, bioimaging and bioactivity studies, these data resources are mainly distinguished by discipline, as different scientific methods require specialised repositories to ensure that the appropriate metadata and controlled vocabulary are associated with the data. Given this variety of repositories, it is important that the resources are well connected and coordinated, thus interoperable, to provide some level of consistency and thereby work towards creating a complete landscape of repositories. To enable easy discovery of data from different open repositories, T2.1 has implemented several of these in the COVID-19 Data Portal<sup>2</sup>.

Currently, the COVID-19 Data Portal only displays COVID-19 related data and has some utility in future outbreaks. Additionally the aggregation of data resources developed within BY-COVID can be quickly used for outbreaks of other diseases, giving researchers a head start by readily providing valuable past research data to build on. Building on the success of the COVID-19 Data Portal, the Pathogens Portal expands its scope to include data on various pathogens beyond COVID-19. Its primary goal is to support responses to emerging disease outbreaks, with a strong focus on pandemic preparedness.

#### 5.1.2. Data submission guidelines

It is essential that these data platforms align their deposition requirements, procedures and infrastructures with the current state of data production; they must be ready to handle data that meet today's as well as tomorrow's quality standards. It is therefore crucial to initiate and maintain direct contact between depositors and archives, and this contact is currently established through the efforts of data stewards. The requirements and procedures for data submission vary between repositories and are currently, in most cases,

---

<sup>1</sup> <https://zenodo.org/records/6939376>

<sup>2</sup> <https://www.covid19dataportal.org/>



not straightforward. In this way, repositories currently inadvertently discourage researchers from submitting their data routinely.

Therefore, in order to actively encourage researchers to submit data and thus increase the number of datasets available for reuse, Task 2.1 aims to clarify the data submission process for several sub-areas. Therefore, we first identified gaps and shortcomings in the current data submission pipelines in several disciplines and actively worked to reduce these gaps.

For example, thanks to the work of the FAIR Image Data Steward, the internal data submission process in Euro-BioImaging to the open repository BioImage Archive<sup>3</sup> has been thoroughly established and tested with various image datasets submitted. In addition, several open training sessions on the FAIR BioImage Data showcasing repositories and their submission procedures have been conducted (see Training section). In addition, openly available material such as the recordings of these workshops and stand-alone material such as how to select the right repository per technology type have been published for the users and wider community to benefit from (link still missing, it is work in progress). We are confident that this proactive approach will lead to an increase in the volume and diversity of research data, which in turn will facilitate greater collaboration, knowledge sharing and scientific progress.

### 5.1.3. Guided data mobilisation

It is indisputable that even the most exhaustive deposition ecosystem is worthless if no data is deposited. It is therefore imperative to raise awareness and thus support data uptake in those repositories. This can be achieved by engaging directly with data producers, facilitated by data stewards. It is becoming increasingly evident and recognised that data stewardship is needed to advise and assist on metadata types and standards.

Euro-BioImaging is at the forefront of these efforts by currently appointing a data steward through BY-COVID to serve as a bridge between bioimaging researchers, communities, and repositories. This enables guided data mobilisation from the data generators directly to the open repositories. Figure 1 showcases successful data mobilisation of COVID-19 related bioimage data to the BioImage Archive. This was only made possible by the close collaboration between data generators and the FAIR Image Data Steward working at Euro-BioImaging under BY-COVID.

---

<sup>3</sup> <https://www.ebi.ac.uk/bioimage-archive/>



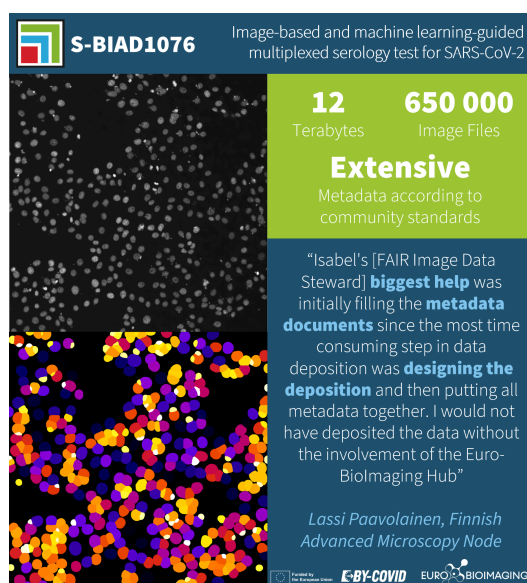


Figure 1: Example of successful data mobilisation in the bioimaging domain facilitated and made possible by the FAIR Image Data Steward. The full dataset is openly available on the BioImage Archive<sup>4</sup>

Meanwhile, Instruct-ERIC is making significant efforts to address one of the current challenges in the data cycle of structural biology: the public unavailability of raw data. It is essential that the data collected, regardless of whether it leads to a final publication or not, is made publicly available at least when the embargo period ends.

Instruct-ERIC provides scientists access to research infrastructures and facilities for various structural biology techniques (such as Electron Microscopy, Nuclear Magnetic Resonance, and X-Ray) through a web portal called ARIA (Access to Research Infrastructure Administration)<sup>5</sup>. The tool being developed aims to achieve the FAIRification of raw data by facilitating the transfer of data to internal data resources and defining an appropriate metadata schema to make it truly useful. This will be accomplished in two phases: first, by integrating it into ARIA, and second, by connecting ARIA with external data repositories.

#### 5.1.4. Knowledge Graph Generator (KGG): A fully automated workflow for creating disease-specific KGs

Aligning with the scope and objectives of BY-COVID WP2, Fraunhofer ITMP has developed several FAIR data compliant workflows to facilitate pre-clinical drug discovery through KG-based applications. This includes identification of bio-active analogues for fragments

<sup>4</sup> <https://www.ebi.ac.uk/biostudies/BioImages/studies/S-BIAD1076>

<sup>5</sup> <https://instruct-eric.org/help/about-aria>



identified through COVID-NMR<sup>6</sup> studies and a MPox KG representing its chemotype-phenotype<sup>7</sup>, which are already published as scientific publications. With these workflows in place, we took a step forward to consolidate and harmonise them in order to develop KGG for generation of KGs for any disease of interest (Manuscript in preparation).

Our fully automated workflow enables fast and rapid generation of KGs compared to conventional methods that are time consuming and require a lot of manual effort. We have embedded underlying schema of curated databases (such as OpenTargets, Uniprot, ChEMBL, etc.) to fetch relevant biochemical knowledge of diseases (Figure 2). The output is a comprehensive and rational assembly of disease-associated entities such as proteins, protein-related pathways, biological processes and functions, chemicals, mechanism of actions, assays and adverse effects, SNPs and mutations. Such a KG in place provides the foundation for answering complex scientific queries and even be a subject of Machine Learning/Artificial Intelligence (ML/AI) algorithms for applications such as link prediction between drug and disease, drug repurposing and drug safety. Lastly, we have also developed several methods for mapping entities across databases, identifying PDB structures of proteins, exploring drug-likeness of chemicals. The programmatic scripts and codes are available at <https://github.com/Fraunhofer-ITMP/kgg>.

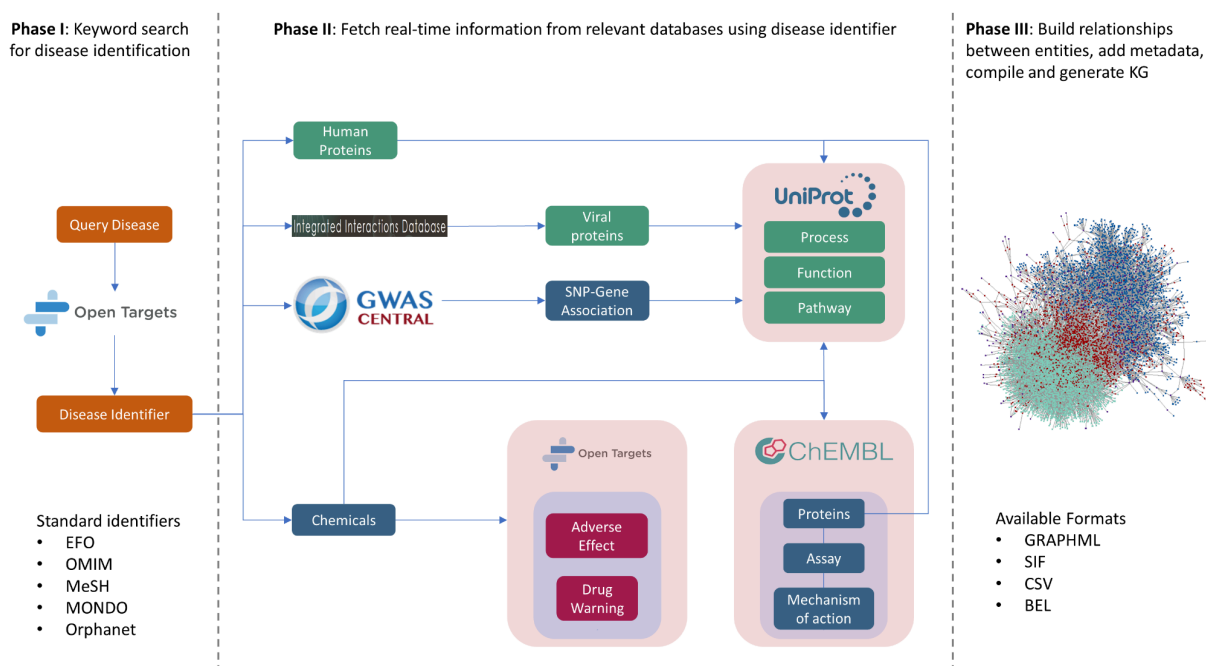


Figure 2: A schematic representation of KGG workflow with different phases involved

<sup>6</sup> Berg, Hannes, et al. "Comprehensive fragment screening of the SARS-CoV-2 proteome explores novel chemical space for drug development." *Angewandte Chemie International Edition* 61.46 (2022): e202205858.

<sup>7</sup> Karki, Reagon, et al. "Mpx Knowledge Graph: a comprehensive representation embedding chemical entities and associated biology of Mpx." *Bioinformatics Advances* 3.1 (2023): vbad045.



## 5.2. Human clinical and health data

In the context of human clinical and health data, discoverability and harmonisation are critical components enabling researchers to identify, access and share relevant datasets across various platforms. While Beacon has been considered a standard for the discovery of human and clinical data<sup>8</sup> by enabling data discovery at source, as shown in deliverable 2.3<sup>9</sup>, there are other platforms using alternative discoverability mechanisms and harmonisation approaches. For example, PHIRI<sup>10</sup> (Population Health Information Research Infrastructure) implements its own discoverable mechanisms tuned to population health data. The following subsections will explore some of these alternative platforms and their methods, showcasing the diversity and innovation within the field.

### 5.2.1. Bespoke CDM + CDM builder

Most of the work done in the baseline use case is based on the mobilisation of real-world data. Data that is primarily collected for other purposes (e.g., treating patients, clinical or epidemiological monitoring and surveillance, statistics) and is reused for research purposes. As opposed to the implementation of general models, such as Observational Medical Outcomes Partnership (OMOP), that require large investments to get all potentially required datasets standardised beforehand, the BY-COVID option has been to develop a bespoke common data model (CDM) approach<sup>11</sup> to RWD access and reuse. BY-COVID has successfully tested this approach in the baseline use case.

The BY-COVID CDM approach facilitates the mobilisation of human clinical and health data by privileging compliance with the data minimisation General Data Protection Regulation (GDPR) principle<sup>12</sup>, thus facilitating the exchange with Data Protection Officers and/or other professionals responsible for ethical and legal compliance within the institution. On the other hand, the CDM acts as the cornerstone of semantic interoperability by defining the data entities and their relationships. The level of specification of the CDM allows data managers (when the access permission is provided) to easily implement the ETL (Extract, Transform, and Load) processes, while fostering the exchange with those able to provide relevant information on data provenance.

A question may remain though about the scalability of this approach as it is driven by a particular research question: Could new research questions, new attributes from other data sources or more data holders be included? The construction of the baseline use case CDM demonstrates that if entities remain, new attributes can be easily included whatever the

---

<sup>8</sup> <https://tehdas.eu/tehdas1/results/tehdas-assesses-data-interoperability-standards/>

<sup>9</sup> <https://zenodo.org/records/13255875>

<sup>10</sup> <https://www.phiri.eu/>

<sup>11</sup> <https://zenodo.org/records/7572373>

<sup>12</sup>Data minimisation as in GDPR article 5.c: Personal data shall be “adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed”



data source. A question comes up about how to automate the CDM construction and upgrade. CDM building tools such as the CDMbuilder<sup>13</sup> may well serve in the automation process, making the BY-COVID bespoke CDM more scalable.

### 5.2.2. Clinical research metadata

The clinical research MetaData Repository (crMDR) operated by the European Clinical Research Infrastructure Network (ECRIN) is a large database of metadata describing clinical research, covering both the *research studies* themselves and the various *data objects* linked to them (e.g., journal articles, study protocols, datasets, patient information sheets, statistical analysis plans, data collection forms, clinical study reports, available biosamples etc.). All information is searchable using a web portal available at <https://crmdr.ecriin.org/>. The goal of the crMDR is to serve as a 'one-stop shop' for researchers and stakeholders. It allows them to identify clinical studies related to specific search terms (such as a diagnosis, study methodology, or geographical location), view associated materials, and locate where those materials are stored. The crMDR code is available in open access on ECRIN's GitHub repository at <https://github.com/ecrin-github>.

Data from the crMDR, a resource of global scope and all trial registries providing data to the WHO<sup>14</sup> for study information, have been incorporated, along with PubMed<sup>15</sup> for bibliographic data. The specialised data repositories, YODA<sup>16</sup> and the NIH BioLINCC<sup>17</sup> have also been included as data sources. More recently, the BBMRI-ERIC directory of biomedical samples<sup>18</sup> in Europe was added (detailed report in a peer-review publication in Open Research Europe<sup>19</sup>). Table 1 lists all the crMDR data sources and the data acquisition method applied by the ECRIN team.

Table 1: Data sources for the crMDR, and the data acquisition method.

Data Source	Data acquisition method
Australian New Zealand Clinical Trials Registry (ANZCT)	File (from WHO)
BBMRI Directory	File / API *
BioLINCC	Web scraping
Brazilian Clinical Trials Registry (ReBec)	File (from WHO)

<sup>13</sup> <https://github.com/cienciadedatosysalud/cdmb>

<sup>14</sup> <https://www.who.int/clinical-trials-registry-platform/network/primary-registries>

<sup>15</sup> <https://pubmed.ncbi.nlm.nih.gov/>

<sup>16</sup> <https://yoda.yale.edu/>

<sup>17</sup> <https://biolincc.nhlbi.nih.gov/home/>

<sup>18</sup> <https://www.bbmri-eric.eu/bbmri-sample-and-data-portal/>

<sup>19</sup> Ohmann C, Canham S, Majcen K et al. Linking the ECRIN Metadata Repository with the BBMRI-ERIC Directory to connect clinical studies with related biobanks and collections. Open Res Europe 2024, 4:50. <https://doi.org/10.12688/openreseurope.17131.1>



Chinese Clinical Trials Registry (ChiCTR)	<i>File (from WHO)</i>
Clinical Research Information Service (CRIS), Korea	<i>File (from WHO)</i>
Clinical Trials Registry – India (CTRI)	<i>File (from WHO)</i>
Clinical Trials Information Service (Europe)	<i>File (from WHO)</i>
ClinicalTrials.gov	<i>API</i>
Cuban Public Registry of Clinical Trials (RPCEC)	<i>File (from WHO)</i>
EU Clinical Trials Register (EU-CTR)	<i>File (from EMA)</i>
German Clinical Trials Register (DRKS)	<i>File (from WHO)</i>
Iranian Registry of Clinical Trials (IRCT)	<i>File (from WHO)</i>
ISRCTN	<i>API</i>
International Traditional Medicine Clinical Trials Registry	<i>File (from WHO)</i>
Japan Registry of Clinical Trials (jRCT)	<i>File (from WHO)</i>
Lebanese Clinical Trials Registry (LBCTR)	<i>File (from WHO)</i>
Pan African Clinical Trials Registry (PACTR)	<i>File (from WHO)</i>
Peruvian Clinical Trials Registry (REPEC)	<i>File (from WHO)</i>
PubMed	<i>API</i>
Sri Lanka Clinical Trials Registry (SLCTR)	<i>File (from WHO)</i>
Thai Clinical Trials Registry (TCTR)	<i>File (from WHO)</i>
Yoda	<i>Web scraping</i>

Study data from the different sources is structured differently - even though all trial registries meet the WHO core data requirements<sup>20</sup>, they do so in slightly different ways. Additionally, the level of detail about linked data objects varies significantly, ranging from just an object type and name, to the rich metadata available from PubMed for journal articles. As a result, the ECRIN Metadata Schema<sup>21</sup> applied is a composite of two existing

<sup>20</sup> <https://www.who.int/clinical-trials-registry-platform/network/who-data-set>

<sup>21</sup> Canham S, Ohmann C. A metadata schema for data objects in clinical research. *Trials*. 2016;17(1):557. doi:10.1186/s13063-016-1686-5.



schemas. For study data, a subset of the data structure used by *ClinicalTrials.gov*<sup>22</sup> was adopted. For data objects, *DataCite (version 3.1)*<sup>23</sup> was used, augmented with additional data points for sensitive and managed data. The ECRIN metadata schema has undergone substantial revisions since its inception, and the current version (v8) was released in September 2023<sup>24</sup>.

During BY-COVID, the crMDR was continuously updated on regular intervals, FAIRifying clinical studies and associated data objects on COVID-19 or other infectious diseases. Figure 3 indicates the total number of COVID-19 studies in the crMDR portal per year for the period 2019 to 2024 (so far).

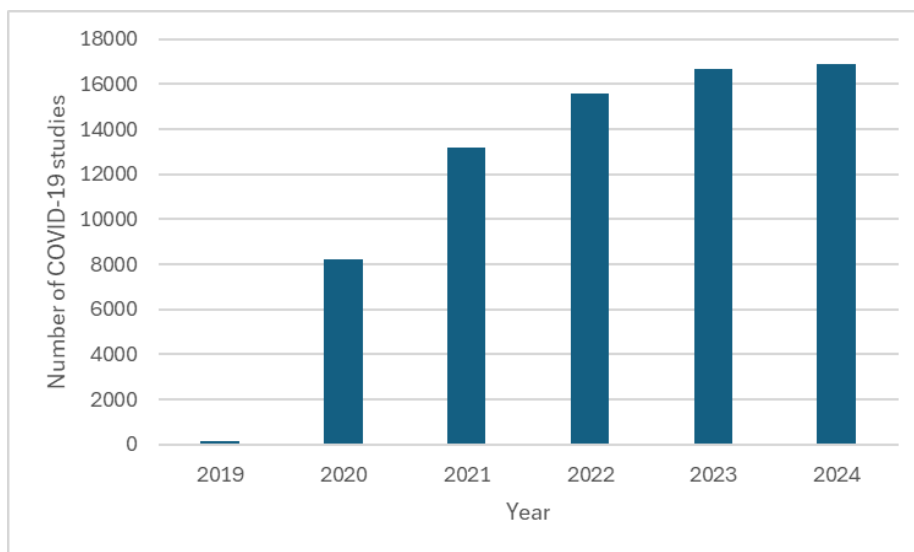


Figure 3: Total number of COVID-19 studies in the crMDR per year for the period 2019 to 2024 (so far).

An API has been created to enable querying of the crMDR and extraction of information on COVID-19 clinical studies. Currently it is applied by the COVID-19 Data Portal and information on 17,239 clinical studies is listed<sup>25</sup>. More information on API calls can be found at: <https://crmdr.ecrin.org/About>.

<sup>22</sup> <https://clinicaltrials.gov/>

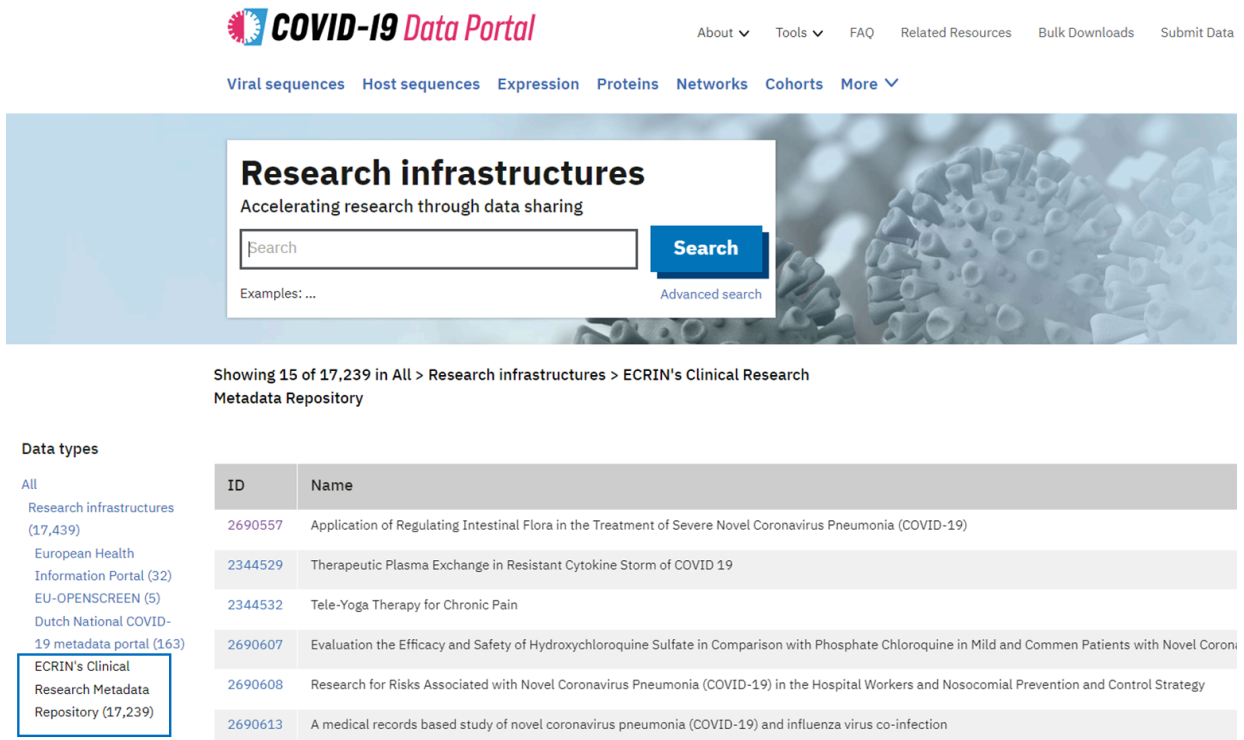
<sup>23</sup> <https://schema.datacite.org/meta/kernel-3.1/index.html>

<sup>24</sup> ECRIN Metadata Schemas for Clinical Research, Version 8 (September 2023), available at <https://zenodo.org/records/8368709>.

<sup>25</sup> <https://www.covid19dataportal.org/search/research-infrastructures?crossReferencesOption=all&overrideDefaultDomain=true&db=ecrin-mdr-covid&size=15>







**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 15 of 17,239 in All > Research infrastructures > ECRIN's Clinical Research Metadata Repository

**Data types**

- All
- Research infrastructures (17,439)
- European Health Information Portal (32)
- EU-OPENSREEN (5)
- Dutch National COVID-19 metadata portal (163)
- ECRIN's Clinical Research Metadata Repository (17,239)**

ID	Name
2690557	Application of Regulating Intestinal Flora in the Treatment of Severe Novel Coronavirus Pneumonia (COVID-19)
2344529	Therapeutic Plasma Exchange in Resistant Cytokine Storm of COVID 19
2344532	Tele-Yoga Therapy for Chronic Pain
2690607	Evaluation the Efficacy and Safety of Hydroxychloroquine Sulfate in Comparison with Phosphate Chloroquine in Mild and Common Patients with Novel Coronavirus
2690608	Research for Risks Associated with Novel Coronavirus Pneumonia (COVID-19) in the Hospital Workers and Nosocomial Prevention and Control Strategy
2690613	A medical records based study of novel coronavirus pneumonia (COVID-19) and influenza virus co-infection

Figure 4: Research Infrastructures including ECRIN metadata entries from the crMDR in the COVID-19 Data Portal

The crMDR allows searching by different study types. Figure 5 demonstrates the percentage of the “interventional”, “observational”, “patient registry” and “other” study types for all the COVID-19 studies listed in the crMDR.

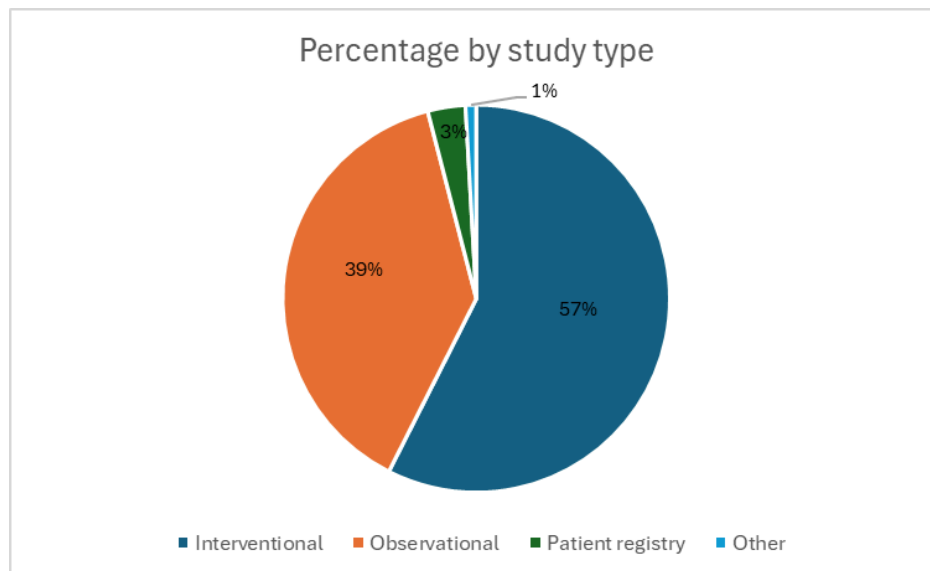


Figure 5: Percentage of “interventional”, “observational”, “patient registry” and “other” COVID-19 studies listed in the crMDR.

Beyond COVID-19, the crMDR can provide information on all registered studies and thus serve as a valuable findability tool in pandemic preparedness scenarios. For example, Figure 6 provides quantitative information on the total number of studies in the crMDR



portal for other infectious diseases such as malaria, tuberculosis and monkeypox for the period 2019 to 2024 (so far).

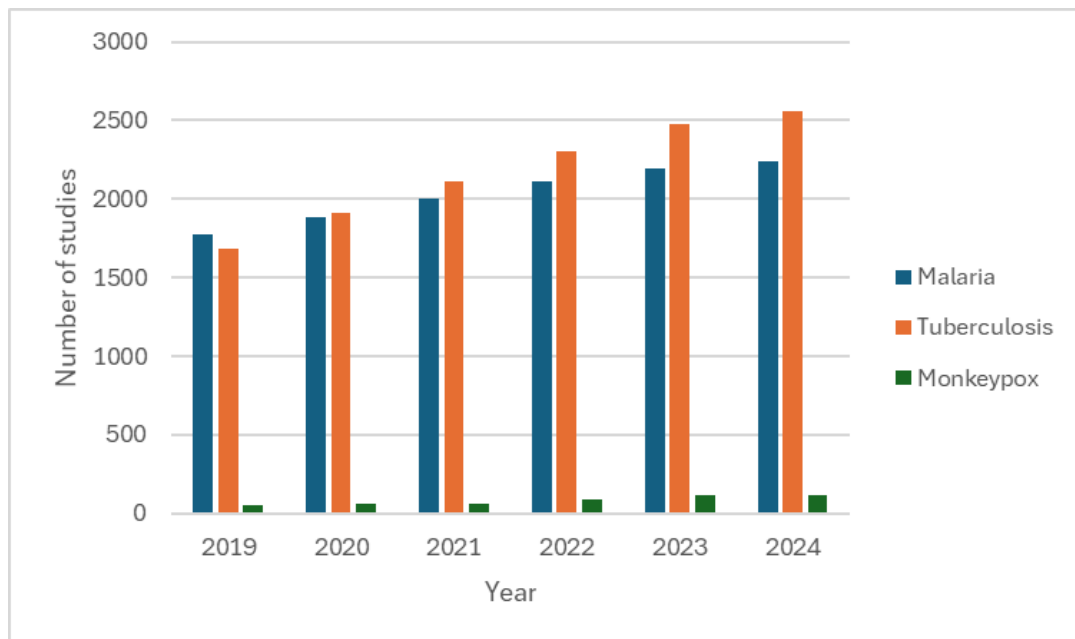


Figure 6: Total number of studies on malaria, tuberculosis and monkeypox in the crMDR per year for the period 2019 to 2024 (so far).

As next step, a collaboration with the Pathogens Portal<sup>26</sup> can be envisaged in order to exchange information on other diseases with “pandemic potential” via for example API calls in a similar way as done for the COVID-19 Data Portal.

### 5.2.3. Harmonisation and mobilisation of COVID-NL data

When the COVID-19 outbreak was declared a global pandemic in March 2020, many local research actions took off. Since COVID-19 was a newly emerging disease, no established data registration or collection of COVID-19 specific clinical data was available. In response, the WHO published a Case Report Form (CRF) with an implementation guideline to arrive at a harmonised collection of data through examination, interview and review of hospital notes. The WHO at that time also provided an option to enter data on the central electronic WHO OpenClinica database<sup>27</sup>, while the International Severe Acute Respiratory and emerging Infection Consortium (ISARIC<sup>28</sup>) provided a similar platform based on REDcap and endorsed use of both the RAPID<sup>29</sup> and the CORE<sup>30</sup> CRF version. Where, the RAPID CRF includes fewer variables than the CORE form and collected data are focused on specific risks, treatments and variables identified as priorities. Around the same time Castor

<sup>26</sup> <https://www.pathogensportal.org/>

<sup>27</sup> <https://who.eclinicalhosting.com/OpenClinica/>

<sup>28</sup> <https://isaric.org/>

<sup>29</sup> [https://isaricdev.wpenginepowered.com/wp-content/uploads/2020/10/ISARIC-WHO-COVID-19-RAPID-CRF\\_EN.pdf](https://isaricdev.wpenginepowered.com/wp-content/uploads/2020/10/ISARIC-WHO-COVID-19-RAPID-CRF_EN.pdf)

<sup>30</sup> [https://isaricdev.wpenginepowered.com/wp-content/uploads/2021/02/ISARIC-WHO-COVID-19-CORE-CRF\\_EN.pdf](https://isaricdev.wpenginepowered.com/wp-content/uploads/2021/02/ISARIC-WHO-COVID-19-CORE-CRF_EN.pdf)



launched their WHO CRF implementation<sup>31</sup> for free to all COVID research. With all these available options, every organisation could decide which one would best fit their organisations' purpose.

In the following months the NFU, the Dutch Federation of University Medical Centres (UMCs), took the initiative to create a joint harmonised COVID-19 data collection based on the WHO CRFs. It was decided to take the ISARIC codebook (REDCap implementation) as the “gold standard” towards which each of the participating hospitals would map and transform their own data. The Castor and ISARIC REDCap implementations unfortunately were not directly interchangeable as Castor CRF implementation, while following the WHO CRF structure, use their own variable names and value/option group lists. This codebook harmonisation was the first step towards a FAIR Dutch COVID-19 WHO/ISARIC CRF data collection. This harmonised codebook also benefited (in part) large Dutch COVID studies like CovidPredict<sup>32</sup> and Eracore.

The next step took local variations due to specific data model/structure of each Electronic Data Capture (EDC) tool in use within the UMCs into account. Apart from variable mapping, re-coding and programmatic mapping towards the agreed-upon (ISARIC) codebook, these local variations had to be accounted for before combining all data into the COVID-NL data. To facilitate this need for data remodelling, it was decided to transform all institute data towards a flat wide data format that could be used to transform the collective data towards the desired data model/structure of choice.

As broad COVID-19 data sharing was one of the objectives from the start, it was acknowledged that data governance would be an important aspect. For that reason, Health-RI and the UMCs initiated an additional collaboration to arrive at common guidelines for COVID-19 studies to allow data reuse. This resulted in a document describing best practices and practical guidelines<sup>33</sup>. One of the main gaps identified was the lack of a standard Data and Material Transfer Agreement (DTA/MTA). Templates for these agreements created by the NFU are available through the Health-RI ELSI servicedesk<sup>34</sup> (page in Dutch, but templates in English). Additionally, a common data governance was implemented by a working group with representatives of most UMCs under coordination of Health-RI. This resulted in a common NFU COVID-19 policy document<sup>35</sup> that was formally approved by the Deans of the UMCs.

---

<sup>31</sup> <https://www.castoredc.com/castor-covid-19-study-database/>

<sup>32</sup> <https://www.covidpredict.org/>

<sup>33</sup> <https://www.health-ri.nl/sites/healthri/files/2020-06/Handreiking%20%20Data%20Governance%20COVID-19.pdf> (in Dutch)

<sup>34</sup> <https://elsi.health-ri.nl/categorieen/verzamelen-en-uitgeven-van-data-en-lichaamsmateriaal/waar-vind-ik-een-voorbeeld-van-0>

<sup>35</sup> [https://elsi.health-ri.nl/sites/elsi/files/2022-03/HRI\\_COVID-NL%20policy%20doc%20v2.0%20final.pdf](https://elsi.health-ri.nl/sites/elsi/files/2022-03/HRI_COVID-NL%20policy%20doc%20v2.0%20final.pdf)



## 5.3. Socioeconomic data

T2.4 on “Relevant socio-economics data sources for infectious disease outbreaks” placed its focus on socioeconomic data sources that could be integrated into the systems and platforms developed across BY-COVID, in order to enhance and allow discoverability of socioeconomic metadata. It did so by a) identifying primary and secondary data sources that were candidate for integration into the COVID-19 Data Portal, b) creating a methodology, tools and process for integration through harvesting and XML transformation, c) participating in discussions and activities in relation to enhancement, cross-fertilisation and interoperability in collaboration with other WP2 and WP5 tasks, and d) contributing to activities that raise awareness on the importance of socioeconomic data in the process of addressing challenges deriving from COVID19 and other infectious diseases.

### 5.3.1. Harvesting tool and XML transformation

An advanced harvesting tool<sup>36</sup> has been implemented with the use and extension of the open-source software DSpace. The concept of the harvesting tool is to collect, process, retrieve, search and browse the harvested metadata records of the selected socioeconomic data sources. The tool can be reused and potentially be customised to support different metadata schemas and content providers. Figure 7 provides a schematic presentation of the tooling.

The basic protocol used for the harvesting process is OAI-PMH<sup>37</sup>, which is supported by most repositories world-wide. Communities can maintain an unlimited number of collections in DSpace. Collections can be organised around a topic or by type of information (e.g. working papers or datasets) or by any other sorting method a community finds useful in organising its digital objects. For the purpose of the BY-COVID project, each collection corresponds to a metadata provider. Using this resource, millions of metadata records can be aggregated. The metadata processes can be run periodically or triggered manually.

Users can browse the harvested metadata records using keywords, topics or any other metadata fields, as well as query available metadata using advanced search functionalities. SOLR is used as the search engine for the repository, which is a highly reliable, scalable and fault-tolerant search engine, that provides distributed indexing, replication and load-balanced querying, automated failover and recovery, centralised configuration and more.

The harvesting tool has been extended with a set of endpoints that allow for the search and retrieval of the metadata records programmatically based on specific business needs, including the consumption of the metadata records to the COVID-19 Data Portal. The API is based on Open API principles (RESTful) and provides JSON responses.

---

<sup>36</sup> The UI of the harvesting tool can be accessed here: <https://t2-4.by-covid.bsc.es/jspui/>. The source code of the harvesting tool is hosted here: [https://github.com/elixir-europe/BY-COVID\\_WP2\\_T2.4\\_Socioeconomics\\_Metadata\\_Harvester](https://github.com/elixir-europe/BY-COVID_WP2_T2.4_Socioeconomics_Metadata_Harvester)

<sup>37</sup> <https://www.openarchives.org/pmh/>



The XML transformation is done with a Python app (by-covid-xml-transformer) that uses the previously mentioned extended DSpace endpoints with query 'covid' to get a list of records to harvest and link to where they can be harvested from. Those records are then harvested, processed (e.g. HTML tags removed) and transformed into extended OmicsDI format supported by the COVID-19 Data Portal. Transformation supports metadata that is in Dublin Core, DDI-Codebook 2.5 or DataCite format. The result is one XML file for every source. Those files are validated against the OmicsDI schema and also validated to be valid XML before being publicly available for the portal to retrieve and use.

Future use cases can be accommodated by configuring DSpace to harvest new data sources and configuring by-covid-xml-traformer queries as long as the new data sources implement an OAI-PMH endpoint, that supports either Dublin Core, DDI-Codebook 2.5 or DataCite formats.

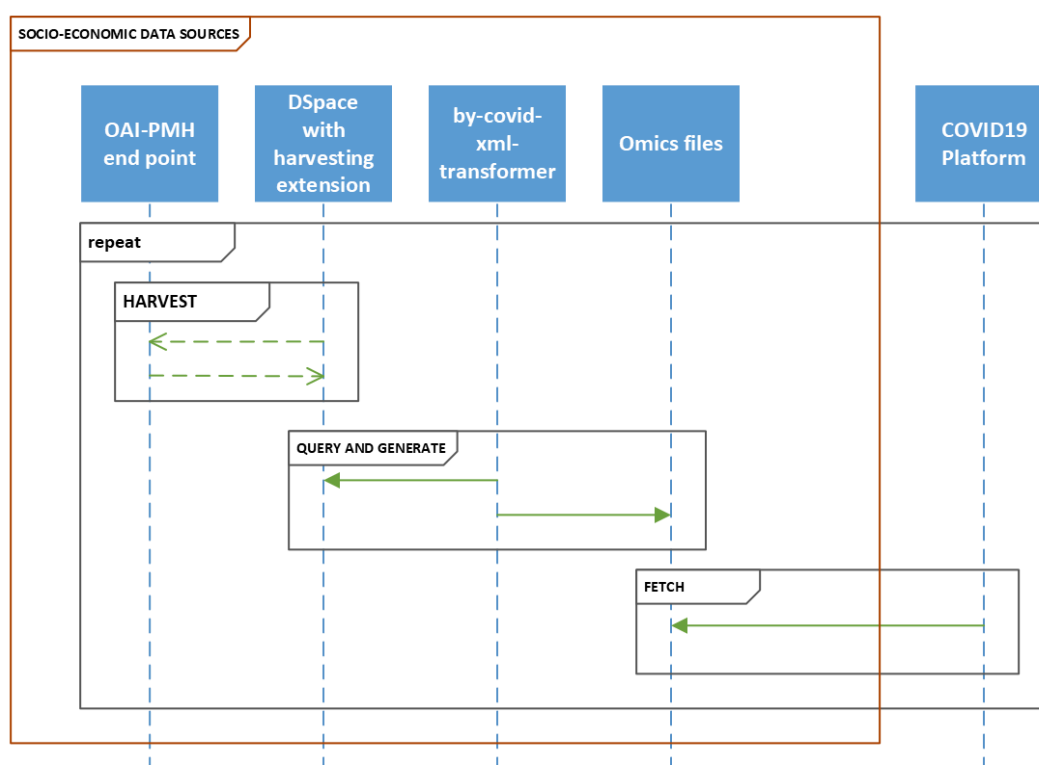


Figure 7: Schematic presentation of the technical solution.

### 5.3.2. Data sources

In order to set realistic and achievable goals within the possibilities offered in the duration of the project, BY-COVID and the T2.4 have identified two socioeconomic data sources as the main goal for integration at the project proposal phase. The primary data sources are curated, enhanced and hosted by project partners. Furthermore, secondary socioeconomic data sources were explored. The results of this work are depicted below.



### 5.3.2.1 Primary data sources

The two primary socioeconomic data sources identified were the CESSDA Data Catalogue (CDC)<sup>38</sup> and European University Institute (EUI) COVID-19 Social Sciences Data Portal.<sup>39</sup>

The CDC holds descriptions of the more than 40,000 data collections held by CESSDA's Service Providers (SPs), originating from over 20 European countries. The CDC is a one-stop shop for searching and discovering European social science data. The presented data includes a range of data types including quantitative, qualitative or mixed-modes data, covering both cross-sectional and longitudinal studies, as well as recently collected and historical data. CESSDA is committed to supporting researchers working on COVID-19 related research, hence a dedicated page has been created with further information specific to COVID-19.<sup>40</sup> Relevant CDC datasets are available<sup>41</sup> through the COVID-19 Data Portal as a result of the T2.4 work.

The EUI COVID-19 Social Sciences Data Portal provides a “one stop-shop” to discover diverse data on the social, economic and political effects of the COVID-19 pandemic, with a goal to further add data on income, social status, well-being, legal responses and related topics from across Europe and beyond, for the benefit of current and future social science research. Relevant EUI COVID-19 Social Sciences Data Portal datasets are available<sup>42</sup> through the COVID-19 Data Portal as a result of the T2.4 work.

### 5.3.2.2 Secondary data sources

In the duration of the project several secondary socioeconomic data sources were considered. Of these, only one was made possible to be integrated into the COVID-19 Data Platform in the duration of the BY-COVID project, while others have significant potential for future actions. Below are the results of this investigation.

Considerable effort has been dedicated into the exploration and harvesting of the Dutch COVID-19 Data Support Programme. In the end a decision has been made for the data source to not be primarily included under the socioeconomic section of the COVID-19 Data Portal, however, relevant metadata from this data source are also exposed and discoverable<sup>43</sup> under that section.

The DANS data station SSH<sup>44</sup> was launched in 2023. The development and timing did not allow for this data source to be integrated directly into the COVID-19 Data Portal, however

<sup>38</sup> <https://datacatalogue.cessda.eu/>

<sup>39</sup> <https://covid19data.eui.eu/>

<sup>40</sup> <https://www.cessda.eu/Covid-19>

<sup>41</sup> <https://www.covid19dataportal.org/search/social-sciences?crossReferencesOption=all&overrideDefaultDomain=true&db=cessda-covid19&sortignorenull=true&size=15>

<sup>42</sup> <https://www.covid19dataportal.org/search/social-sciences?crossReferencesOption=all&overrideDefaultDomain=true&db=eui-covid19&sortignorenull=true&size=15>

<sup>43</sup> <https://www.covid19dataportal.org/search/social-sciences?crossReferencesOption=all&overrideDefaultDomain=true&db=covid-nl-ssh&sortignorenull=true&size=15>

<sup>44</sup> <https://dans.knaw.nl/en/social-sciences-and-humanities/>



DANS is working together with CESSDA towards a potential integration under the CDC and discoverability of the relevant datasets in the COVID-19 Data Portal through the CDC.

The ODISSEI portal<sup>45</sup> was another secondary data source that was explored, however the timing of development of the portal, sustainability challenges and upcoming activities of this aggregator led to the decision by the team not to pursue its integration at this time.

The SHARE ERIC Corona survey<sup>46</sup> was also investigated for the possibility of integration, however, it was not viable due to not having any programmatic access to the metadata in the duration of the BY-COVID project.

Lastly, the European Social Survey (ESS) was investigated, but the data format could not be supported at this time.

It should also be noted that one possible path for any data actor to get their data included in the COVID-19 Data Platform is to deposit it at their local CESSDA Service Provider. Deposited data will then be included in the CESSDA Data Catalogue, thus enabling it to be harvested for the Platform.

### 5.3.3 Cross-domain collaboration

T2.4 played a key role in fostering cross-domain collaboration within the BY-COVID project. This was achieved through activities such as regularly sharing updates on the integration of socioeconomic data sources with other WP2 tasks, contributing to interoperability, harmonisation, and discoverability efforts, and, most importantly, co-organizing collaborative sessions aimed at improving dataset queries and raising awareness. As awareness raising activities have been extremely important for cross-domain collaboration and one of the main goals of the participating organisations in the BY-COVID project, hence they are briefly described under section 5.3.4.

### 5.3.4 Awareness raising

The T2.4 team through their effort in WP2, their individual organisations' role and in line with their participation in other activities have consistently worked towards raising awareness on the importance of socioeconomic data in the process of addressing challenges deriving from COVID-19 and other infectious diseases. Some examples of such awareness raising activities included but are not limited to the organisation of sessions for knowledge exchange during General Assembly meetings, co-organisation and active participation in workshops, engagement and training sessions in collaboration with WP5 and WP6, as well as contributions on the socioeconomics section of the Infectious Disease

---

<sup>45</sup> <https://portal.odissei.nl/>

<sup>46</sup> <https://share-eric.eu/data/data-set-details/share-corona-survey-1>



Toolkit (IDTK).<sup>47</sup> Relevant activities are described in detail under the BY-COVID deliverables D6.1 Stakeholder engagement report and D6.2 Training efforts report.

## 5.5. Network of Covid-19 Beacons

Previous deliverable 2.3<sup>48</sup> encompasses a collection of Beacons that form part of the BY-COVID initiative, covering a diverse spectrum of data types. These Beacons include but are not limited to:

- Viral genomes
- Basic and rich patient clinical data
- Patient genomic data
- Epidemiological data

The integration of these Beacons could be done with a Beacon Network, allowing for seamless aggregation and querying of similar data types across different Beacons. This network would enhance data accessibility and interoperability, significantly benefiting researchers and healthcare professionals. However, the establishment of a dedicated Beacon Network requires substantial resources and ongoing maintenance to ensure its effectiveness and reliability.

Given these considerations, we recommend that the BY-COVID initiative integrates its Beacons into an existing Beacon Network to leverage already established resources and infrastructure. The ELIXIR Beacon Network<sup>49</sup>, maintained by ELIXIR, presents an ideal opportunity. This network is not only open and well-maintained but also offers a mature platform for data sharing and collaboration.

## 6. Results and discussion

The concerted efforts within WP2 of the BY-COVID project have significantly influenced the outcomes of other work packages, highlighting the collaborative nature of our work. This approach ensures that our outputs are both user-centric and sustainable.

The collaborative efforts across these work packages have not only advanced the BY-COVID goals but have also established a robust infrastructure. This infrastructure is critical for the rapid deployment and dissemination of information in response to future pandemics, ensuring readiness and resilience.

---

<sup>47</sup> <https://www.infectious-diseases-toolkit.org/data-sources/socioeconomic-data>

<sup>48</sup> <https://zenodo.org/records/13255875>

<sup>49</sup> <https://beacon-network.elixir-europe.org/>





## 6.1. Connection with other WP

### 6.1.1. Connection with WP3: FAIRsharing and the COVID-19 Data Portal

One of the focuses of WP2 was to maintain and populate the FAIRsharing collection for BY-COVID. This collection includes 72 records, encompassing the four domains of data that are part of WP2: non-patient data, human/patient biomolecular data, human/patient clinical and health data and socioeconomics data. Among these resources, 20 are distinct data sources, while the remaining entries include links to these sources, such as metadata standards or ontologies that they use.

As shown in Figure 8, FAIRsharing also offers a functionality to view a relationship graph, which visually depicts how the various resources within the collection are interconnected.

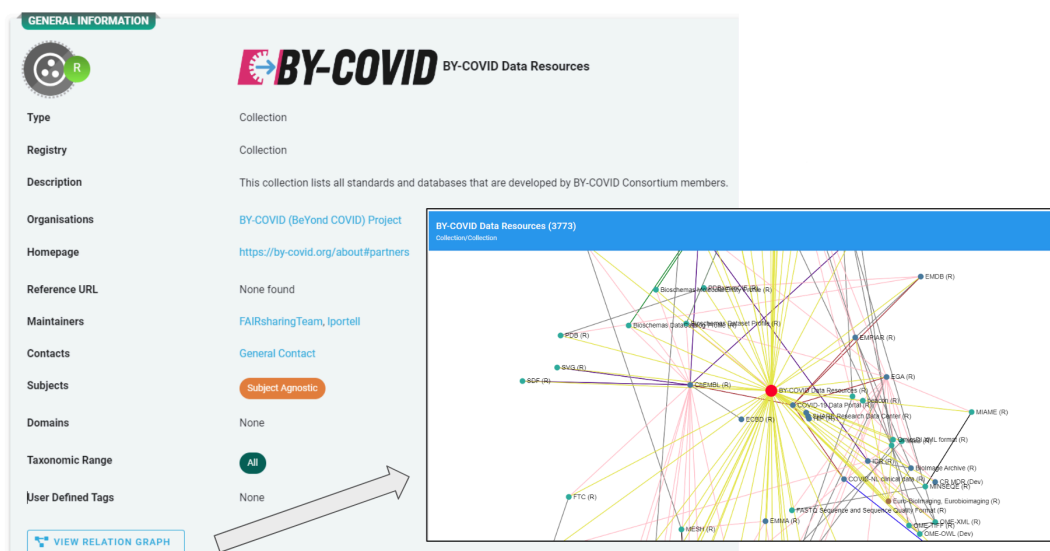


Figure 8: FAIRshaing BY-COVID collection.

Another significant activity in WP2 was enhancing and expanding the resources in the COVID-19 Data Portal. We added information across various sections, including host sequences, imaging, social sciences and humanities, and research infrastructures (Figure 9).



**COVID-19 Data Portal**

Home About Tools FAQ Related Resources Bulk Downloads Submit Data

Viral sequences Host sequences Expression Proteins Networks Cohorts More

**COVID-19 Data**  
Accelerating research through data sharing

Search [Search] Advanced search

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

More

- Samples
- Imaging
- Social Sciences & Humanities
- Literature
- Research Infrastructures

**Viral sequences** →  
Raw and assembled sequence and analysis of SARS-CoV-2 and other coronaviruses.  
21,760,426 records

**Host sequences** →  
Raw and assembled sequence and analysis of human and other hosts.  
31,164 records

**Expression** →  
Gene and protein expression data of human genes implicated in the virus infection of the host cells. Identifying cell types and genes with highest expression in SARS-CoV-2 infections.  
339 records

**Proteins** →  
Curated functional and classification data on the SARS-CoV-2 protein entries and associated protein receptors.  
6,925 records

**Share new data** →  
Contact our curator teams, who will assist you with submitting your data to EMBL-EBI repositories >

**Latest news** →

**New paper presents Data Hubs functionality for analysing pathogen sequences** →  
27 Mar 2024

Figure 9: COVID-19 Data Portal including the sections where we collaborated.

For the host sequences section, we have automated the display of EGA studies and datasets on the portal (Figure 10). As shown in the menu, you can access all controlled access studies and datasets from EGA, along with other datasets from various sources such as ENA.

The EGA Helpdesk identifies COVID-19-related studies and datasets and tags them with the COVID-19 MONDO ontology term. These entries are then identified by the indexing system - EBISearch and available in the COVID-19 Data Portal.

Additionally, we have added a field in the table that includes associated data access attributes (DUO codes) from EGA to the COVID-19 Data Portal, providing information about the consent terms for each dataset.



Data types		<a href="#">Edit table view</a>
	Accession	Associated consent
<b>All</b>		
Host sequences (31,164)	EGAS00001004391	
Human studies (controlled access) (75)	EGAS00001004412	
Human datasets (controlled access) (157)	EGAS00001004419	
Human reads (consented for full access) (24,688)	EGAS00001004481	
Other species reads (5,970)	EGAS00001004489	
Association studies (274)	EGAS00001004502	
	EGAS00001004503	
<b>Repository</b>	EGAS00001004508	health or medical or biomedical research
<input type="checkbox"/> EGA (75)	EGAS00001004571	
<b>Publication Date</b>	EGAS00001004689	
<input type="checkbox"/> 2024 (6)	EGAS00001004717	health or medical or biomedical research; not for profit, non commercial use only
<input type="checkbox"/> 2023 (11)	EGAS00001004772	
<input type="checkbox"/> 2022 (6)	EGAS00001004928	
<b>Associated consent</b>	EGAS00001004951	project specific restriction; general research use
<input type="checkbox"/> user specific restriction (9)	EGAS00001004996	health or medical or biomedical research; time limit on use; user specific restriction
<input type="checkbox"/> project specific restriction (8)		
<input type="checkbox"/> institution specific restriction (6)		

Showing 15 results [Previous](#) Page 1 of 5 [Next](#)

Figure 10: Host sequences section in the COVID-19 Data Portal displaying EGA studies.

For the imaging section, a similar approach was followed. Now, images related to COVID-19 can be found in the portal's imaging section (Figure 11).

**COVID-19 Data Portal** | About | Tools | FAQ | Related Resources | Bulk Downloads | Submit Data

Viral sequences | Host sequences | Expression | Proteins | Networks | Cohorts | More

### Imaging

Biological images from microscopy and other platforms

Search [ ] [Search](#)

Examples: polymerase , movies , complex ... [Advanced search](#)

Showing 6 of 72 in All > Imaging

**Data types** | **Images** 4 results

All

Imaging (72)

Images (4)

Electron microscopy

public image archive (68)

A high-resolution 3D atlas of the spectrum of tuberculous and COVID-19 lung lesions Source: bioimages

Postmortem high-dimensional immune profiling of severe COVID-19 patients reveals distinct patterns of immunosuppression and immunoactivation Source: bioimages

Bronchial epithelia from adults and children: SARS-CoV-2 spread via syncytia formation and type III interferon infectivity restriction Source: bioimages

[View all 4 results in Images](#)

Figure 11: Imaging section in the COVID-19 Data Portal

Before adding these datasets to the COVID-19 Data Portal, they first had to be mobilised into the appropriate data sources, such as the BioImage Archive. To facilitate this, a data mobilisation and discoverability pipeline was developed. The pipeline achieved several important milestones:

- Metadata in REMBI format: Using the community standard ensures that the data is findable, interoperable, and reusable.
- Ontology terms: Incorporated wherever possible to improve interoperability.

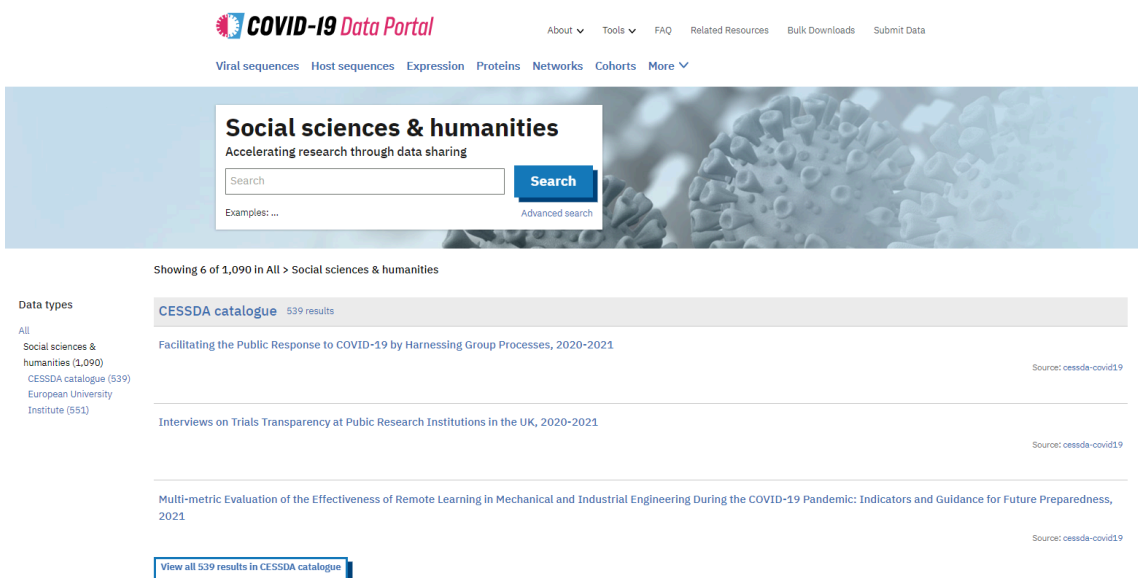


- Structured datasets: Organised into experimental units, enhancing reusability.
- New hire: A dedicated person was hired to assist with data submission.
- Image conversion to OME-Zarr: This format allows in-browser visualisation, making it more user-friendly.

The goal is to share this pipeline with the community for widespread use. We also aim to include it in the IDTk, along with all relevant information.

Moving on to social sciences and humanities, we collaborated closely with WP3 to integrate datasets from the CESSDA Data Catalogue and the EUI COVID-19 Data Portal into the COVID-19 Data Portal (Figure 12).

To achieve this, a metadata harvesting tool was developed. This tool collected metadata from these resources and transformed it into the XML format required by the COVID-19 Data Portal to display the datasets. This tool has since been used by others in WP2 for similar tasks.



The screenshot shows the COVID-19 Data Portal interface. At the top, there is a navigation menu with links for 'About', 'Tools', 'FAQ', 'Related Resources', 'Bulk Downloads', and 'Submit Data'. Below this, there are category links: 'Viral sequences', 'Host sequences', 'Expression', 'Proteins', 'Networks', 'Cohorts', and 'More'. The main heading is 'Social sciences & humanities' with the tagline 'Accelerating research through data sharing'. A search bar is present with a 'Search' button and an 'Advanced search' link. Below the search bar, it says 'Showing 6 of 1,090 in All > Social sciences & humanities'. On the left, there is a 'Data types' sidebar with 'All' selected and 'Social sciences & humanities (1,090)' highlighted. The main content area lists three datasets: 'CESSDA catalogue' (539 results), 'Facilitating the Public Response to COVID-19 by Harnessing Group Processes, 2020-2021', and 'Interviews on Trials Transparency at Public Research Institutions in the UK, 2020-2021'. A button at the bottom of the list says 'View all 539 results in CESSDA catalogue'.

Figure 12: Social sciences and humanities section in the COVID-19 Data Portal

Finally, other research infrastructures, including the European Health Information Portal, EU-OPENSREEN, the Dutch National COVID-19 Metadata Portal, and ECRIN's Clinical Research Metadata Repository, have also indexed their records in the COVID-19 Data Portal (Figure 13).



The screenshot displays the 'Research infrastructures' section of the COVID-19 Data Portal. At the top, there is a navigation menu with links for 'About', 'Tools', 'FAQ', 'Related Resources', 'Bulk Downloads', and 'Submit Data'. Below this, a secondary menu lists categories like 'Viral sequences', 'Host sequences', 'Expression', 'Proteins', 'Networks', 'Cohorts', and 'More'. The main content area features a search bar with the text 'Research infrastructures' and 'Accelerating research through data sharing'. A search button is visible next to the search bar. Below the search bar, there are three search results:

- European Health Information Portal** (32 results)
- COVID-19 Belgian Database**
- COSMO-Spain**

Each result includes cross-references to the Ontology Lookup Service (OLS) and a source link to phiri. A button at the bottom of the results section says 'View all 32 results in European Health Information Portal'.

Figure 13: Research Infrastructures section in the COVID-19 Data Portal

The COVID-19 Data Portal will remain operational for at least two years following the end of the project. During this period, automated data feeds from partners will continue to update as usual. The team will strive to maintain consistency by removing references to externally disabled resources and updating links when resource URLs change, as long as these changes are straightforward. However, more complex changes might not be accommodated, such as significant alterations in metadata formats from external resources, which might necessitate the removal of those resources if we cannot adjust our parsers.

Furthermore, adding new resources to the COVID-19 Data Portal is not anticipated. Future developments will be focused on the Pathogens Portal. The goal is to integrate relevant functionalities from the COVID-19 Data Portal into the Pathogens Portal. Ideally, users will be able to access most of the current COVID-19 features through the Pathogens Portal by selecting "COVID-19" as a filter or a similar option. The extent and speed of this integration will depend on funding constraints and may require prioritisation based on usage and other factors. Beyond the initial two-year period, the COVID-19 Data Portal may be archived or frozen, with potential redirection to the Pathogens Portal.

### 6.1.2. Connection with WP4: IDTk pages created

In the context of WP2, we have made significant contributions to various sections of the IDTk<sup>50</sup>. Specifically, we have developed three pages dedicated to Data Sources, which align closely with the WP2 focus. These pages cover:

- Human Biomolecular Data Sources

<sup>50</sup> <https://www.infectious-diseases-toolkit.org/>



- Human Clinical and Health Data Sources
- Socioeconomic Data Sources

Additionally, we have created a page focused on other areas:

- Human Biomolecular Data Analysis
- Imaging Data Analysis showcase
- Knowledge Graph Generator (KGG) showcase
- Socioeconomics Data Analysis
- Socioeconomics Provenance

### 6.1.3. Connection with WP5: Lessons learned from the use cases

In the context of WP2, the baseline use case is an exemplary illustration of seamless integration within HealthData@EU, where the wealth of health-related data could ideally be made accessible for data reuse in response to future outbreaks, under the jurisdiction of Health Data Access Bodies - those that have participated in the baseline use case as nodes in the federation.

This seamless integration aimed at setting a methodology and the required technology to obtain easier access to and mobilisation of sensitive real-world health data - which includes data from electronic health records, disease registries, surveillance and monitoring health information systems, claims and admin data. The seamless integration consists of: 1) ensuring ethical and legal compliance (using data privacy by design, implementing a solution, -based on a detailed protocol, a data management plan and then a thorough CDM- to comply with the data minimisation principle, and compliant with data holders' disclosure policies); 2) including actual HDABs and data holders, usually institutions not research-oriented, as part of the use case to test discoverability, accessibility and data integration in real conditions; 3) embedding data discovery, access and integration within the pipeline - the bespoke CDM provides a thorough specification for input data, including syntactic and semantic standards, thus, supporting the decision on what data source is the more appropriate, and helping data holders to implement the ETL processes including linkage across data sources - see in section 5.2.1 CDM builder requirements); and, 4) implementing all pipelines in a trusted research environment (Docker, Singularity) that can be deployed in the secure processing environments foreseen in the HealthData@EU legislation.

Unlike the primary collection of data for research purposes, the baseline use case has designed, implemented and deployed methods and technology to deal with sensitive health real-world data, natively collected for other purposes and usually not prepared for further reuse, embedding data discoverability, access, integration and mobilisation within the



same pipeline while complying with legal, organisational, semantic and technological interoperability.

When it comes to the FAIRification of clinical research data, ECRIN worked together with an established Trusted Research Environment (TRE) (*also known as Secure Processing Environment or Data Safe Haven*) operated by the University of Oslo (UiO) and named TSD<sup>51</sup> to enhance the *Findability, Accessibility and Reusability* of COVID-19 clinical research data. In brief, WP5 has delivered: 1) a technical solution for securely storing sensitive data from clinical studies according to applicable legal and ethical requirements (*clinical research Data Sharing Repository - crDSR*<sup>52</sup>); 2) a governance framework elaborated with potential data providers and data reusers from the clinical research community; 3) documentation of processes and ELSI templates (e.g., Data Transfer Agreement, Informed Consent Form for secondary use of clinical study data); 4) an assessment of 200 COVID-19 clinical studies with an explicit Data Sharing Statement in a clinical trial registry to quantify willingness to share COVID-19 clinical study data in a global scale and inventory studies willing to share data<sup>53</sup> (*the crMDR and the metadata schema described in section 5.2.2 of this deliverable were used*); 5) a follow-up survey sent to the studies willing to share COVID-19 data to capture data access procedures and prerequisites; 6) piloting the technologies and methods delivered in BY-COVID with the collaboration of the VACCELERATE project<sup>54</sup> and their EU-COVAT-1-AGED funded trial<sup>55</sup>. Detailed information on the above is provided in deliverable “D5.2 Secondary use of vaccine trial data”.

#### 6.1.4. Connection with WP6: Training activities

In collaboration with WP6, four trainings were carried out during the BY-COVID project related to WP2 activities.

- Using Beacon to make sensitive data discoverable: This online workshop was done the 2nd and 3rd of November of 2022. The aim of this workshop was to present the Beacon tool to discover sensitive data and how BY-COVID data could be mapped onto the Beacon Model. This training will be divided in two sessions:
  - Introduction and working principles of Beacon, oriented towards clinicians & researchers working with human data and
  - Implementation or deployment of Beacon, oriented towards software developers & bioinformaticians. Both sessions will be open to all.

<sup>51</sup> <https://www.uio.no/english/services/it/research/sensitive-data/>

<sup>52</sup> <https://crdsr.ecrin.org/login>

<sup>53</sup> Ohmann, C., Panagiotopoulou, M., Canham, S. et al. An assessment of the informative value of data sharing statements in clinical trial registries. *BMC Med Res Methodol* 24, 61 (2024).

<https://doi.org/10.1186/s12874-024-02168-8>

<sup>54</sup> <https://vaccelerate.eu/>

<sup>55</sup> Neuhann JM, Stemler J, Carcas AJ et al. Immunogenicity and reactogenicity of a first booster with BNT162b2 or full-dose mRNA-1273: A randomised VACCELERATE trial in adults ≥75 years (EU-COVAT-1). *Vaccine*. 2023 Nov;41(48):7166-7175. doi: 10.1016/j.vaccine.2023.10.029.





- Beacon Network version 2 - Prototype demonstration and updates to the Beacon v2 specification: This workshop presented the concepts and ideas behind the Beacon Network, and the implications of the Beacon network v2 on the Beacon v2 specification. In the second part of the workshop, the Beacon Network v2 prototype was presented, and a live demonstration was provided. Both sessions were followed by an extended Q&A session with all groups present.
- Euro-BioImaging's Guide to FAIR Bioimage Data: Two online Workshops in 2023 and 2024. In these interactive online workshops participants were introduced to the FAIR principles in the context of bioimaging data and were provided simple yet effective steps for a smooth start to a FAIR journey. The course covered the benefits of FAIR data for molecular, cellular as well as pre-clinical imaging and best practices for data management including an advanced lecture on cloud-compatible next-generation file formats for image data.

In addition to the workshop that WP2 organised, we also participated in the different hackathons that WP2 organised to help providers onboard resources in the COVID-19 Data Portal as well as the IDTk contentathons organised by WP4.

## 7. Conclusions

The BY-COVID project has made substantial progress in advancing data accessibility, interoperability, and usability, significantly enhancing pandemic response capabilities. By prioritising different types of data (sensitive and non-sensitive) and integrating specialised repositories into platforms like the COVID-19 Data Portal, the project has fostered a collaborative research environment. These efforts help in data sharing and improve researchers' ability to access relevant data quickly, which is essential for ongoing research and future outbreak responses.

In the realm of human clinical and health data, the project has successfully implemented a bespoke common data model (CDM) that ensures GDPR compliance while facilitating efficient data mobilisation and integration. Tools like CDMbuilder and continuous updates to clinical research metadata have improved data discoverability and usability of sensitive data, showcasing scalable and adaptable data management solutions.

The integration of socioeconomic data into the BY-COVID infrastructure has broadened the scope of information available for understanding the impacts of infectious diseases. Advanced harvesting tools and XML transformation processes have incorporated key data sources into the COVID-19 Data Portal, supporting immediate research needs and setting a precedent for future data integration efforts.





Additionally, the BY-COVID project has exemplified the power of interdisciplinary collaboration. The coordination between WPs has been instrumental in making resources accessible and supporting data-driven decision-making. Contributions to the Infectious Diseases Toolkit (IDTK) and collaborative training efforts have further strengthened the project's impact by equipping participants with essential skills and fostering continuous learning.

Overall, the BY-COVID project has set a benchmark for data management and collaboration, significantly enhancing global health resilience and preparedness for future pandemics.

## 8. Next steps

As the BY-COVID project concludes, it is essential to focus on the long-term sustainability of the resources and infrastructure developed. This includes establishing ongoing maintenance and support to ensure that tools, data repositories, and platforms remain accessible and functional for the scientific community. Additionally, maintaining and strengthening the connections established among project partners and stakeholders is vital. Fostering continued collaboration will support future research endeavours and enhance our collective ability to address emerging challenges in pandemic preparedness and response.

The next steps also include the transition of the COVID-19 Data Portal to the Pathogens Portal. As mentioned before, the COVID-19 Data Portal will remain operational for at least two years post-project, with automated data feeds from partners continuing as usual. However, new resources to the COVID-19 Data Portal will not be added during this period. Instead, future efforts will focus on the Pathogens Portal, with the goal of integrating key functionalities from the COVID-19 Data Portal into this new platform. After the initial two-year period, the COVID-19 Data Portal may be archived or frozen, with potential redirection to the Pathogens Portal.

## 9. Impact

The BY-COVID project has had a significant impact on infectious disease research and global health preparedness, particularly through its development and enhancement of data resources and repositories. By improving the accessibility, interoperability, and usability of research data, the project has empowered researchers and policymakers to collaborate more effectively and make data-driven decisions.



Specifically for WP2 work, the improvement of key platforms, such as the COVID-19 Data Portal, and the integration of automated data feeds and repositories have streamlined data sharing and ensured that valuable resources remain readily available. In addition, the introduction of FAIR data mobilisation guidelines has further improved the quality and consistency of data, fostering a more cohesive and efficient research environment.

The project's commitment to open access and interdisciplinary collaboration has set new standards for data management, laying the groundwork for future initiatives. Ultimately, the BY-COVID project has not only advanced our understanding of infectious diseases but also strengthened global health resilience, positioning the scientific community to respond more rapidly and effectively to future outbreaks.

