Deliverable 8.2.3 Project Data Management Plan initial release and periodic updates

Data Management Plan

Project Title (grant agreement No)	Beyond COVID Grant Agreement 101046203		
Project Acronym (EC Call)	BY-COVID		
WP No & Title	WP8: Coordination, Project Management and Ethical, Legal and Social implications		
WP Leaders	Peter Maccallum (ELIXIR hub)		
Deliverable Lead Beneficiary	7 BBMRI		
Contractual delivery date	31/07/2024	**Actual Delivery date**	31/07/2024
Delayed			
Partner(s) contributing to this deliverable	EBI, ELIXIR Hub, ELIXIR-UK Sciensano, IACS, BBMRI, ECRIN		
Authors	Eva Garcia Alvarez - BBMRI, Petr Holub - BBMRI		
Contributors	Ilaria Colussi - BBMRI, Maria Panagiotopoulou - ECRIN		
Acknowledgements (not grant participants)			
Reviewers	Enrique Bernal Delgado - IACS Ramon Launa Garces - IACS Henning Hermjakob - EMBL-EBI Peter Maccallum - ELIXIR		

Funded by the European Union

Log of changes

Date	Mvm	Who	Description
28/06/2024		Eva García Álvarez Petr Holub Ilaria Colussi Maria Panagiotopoulou	First draft
19/07/2024		Authors, contributors and reviewers	Final version

Table of contents

1. Executive Summary

This document builds on the previous versions of the Data Management Plan (hereinafter "the project DMP)[1,2]. It gathers the steps that were made within the project since its last review in M15.

The DMP is a mandatory deliverable by the European Commission (EC) that has to be tailored to the specific project activities and should document the relevant project repositories and how they are managed. The Data Management Plan for BY-COVID is developed by WP8 as an integral part of the governance structure, and also oversees the quality assurance of the project outcomes. BY-COVID will not generate new sensitive patient level data but integrate and link pre-existing resources. Hence, responsibility remains with the original data providers, data access committees and based on processes specifically established for each dataset. The project DMP considers:

- Overview of the project data flow
- Compliance with Open Science and FAIR data sharing
- Ethics Requirements[3], in accordance with the Grant Agreement (WP9)

This document is the third version of the BY-COVID DMP delivered in month 34, being an update of the second version delivered in month 15. This deliverable is based on the template and the guidelines provided by the European Commission for the Horizon Europe Framework Programme. Of note, the project DMP has been a live document that was reviewed and updated periodically during the whole project duration.

[1] García Álvarez, Eva, Mayrhofer, Michaela, & Holub, Peter. (2021). BY-COVID - D8.2 - Data Management Plan (v1.0). Zenodo. https://doi.org/10.5281/zenodo.6884816
[2] Garcia Alvarez, E., & Holub, P. (2022). BY-COVID D8.2.2 Project Data Management Plan initial release and periodic updates M15 (V1.0). Zenodo. https://doi.org/10.5281/zenodo.7476926
[3] García Álvarez, Eva, Mayrhofer, Michaela, & Holub, Peter. (2021). BY-COVID - D8.2 - Data Management Plan (v1.0). Zenodo. https://doi.org/10.5281/zenodo.6884816 (Annex 2)

2. Contribution towards project objectives

The project handbook[4] defines the project process that provides the framework to accomplish all project's objectives within the scope, budget and the required level of quality. This deliverable contributes to all objectives as listed below:

	Key Result No and description	Contributed
Objective 1 Enable storage, sharing, access, analysis and processing of research data and other digital research objects from outbreak research.	1. A research data management practice in European research infrastructures practice that drives discovery, access and reuse of outbreak data and directly links experimental data from HORIZON-INFRA-2021-EMERGENCY-02 transnational access projects into the COVID-19 Data Portal.	Yes
	2. Workflows and processing pipelines that integrate transparent quality management and provenance and are openly shared.	Yes
	3. Research infrastructures on-target training so that users can exploit the platform.	Yes
	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	Yes

[4] Arenas Marquez, Juan, & Troncoso Quilaqueo, Andrea. (2021). BY-COVID - D8.1 - Project Management Handbook (v1.0). Zenodo. https://doi.org/10.5281/zenodo.6884734

	to serve the research needs of epidemiology and public health. 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.	Yes
Objective 3 Link FAIR data and metadata on SARS-CoV-2 and COVID-19	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.	Yes
	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, including tracking genomics variations of SARS-CoV-2 and identifying new variants of concern.	1. Broad uptake of viral *Data Hubs* across Europe deliver an order-of-magnitude increase in open viral variant detection and sharing.	Yes
	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.	Yes

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.	Yes
	4. To empower national centres to build capacity and train platform users and data providers (e.g., from life, social or health sciences), and with experts from across partner institutions collaborating to create training materials for the identified gaps, and to exchange experiences and knowledge.	Yes

3. Methods

This third version of the Data Management Plan was generated based on the previous ones[5,6]. Following their structure and topics, the new steps done within the project are included in this new version.

In addition, the DMP for the baseline use case is referenced here. It is based on the Horizon Europe template[7]. Thus, we take this template as a guideline to briefly go through the three remaining use cases. Of note, as stated in the previous version of this document, the use cases or their data providers should have their own DMP or data best practises guidance, therefore here a general overview of each of them is included, not going into fine-grained details.

Similarly, the respective data providers, data access committees and processes for each dataset remain accountable for project operations. Data security remains the responsibility of the databases, repositories and archives holding data (e.g., the EGA has a defined security process[8] and follows best practice guidelines aligned with the GA4GH Security Working Group, the BBMRI-ERIC Policy for Access to and Sharing of Biological Samples and Data[9] and follows OECD Council Recommendations on Health Data Governance[10]). IP background and existing data licences for the project might be diverse. IP rights are managed in the consortium agreement. Where code, software or ontologies are developed, a permissive licence is applied and an open approach is taken by partners in line with their institutional policies. While access rights and licences predating the BY-COVID project are respected, participating organisations are encouraged to review existing licences to comply with Open Access requirements. Data sharing and reuse follow the FAIR principles[11] (Findability, Accessibility, Interoperability, and Reusability) and make use of licences such as the Creative Commons or Open Data Commons. Release on publication, as part of the open access policies, is envisioned and BY-COVID follows established embargo principles (e.g., in the European Nucleotide Archive, ENA; release after 6 or 12 months).

---

[5] García Álvarez, Eva, Mayrhofer, Michaela, & Holub, Peter. (2021). BY-COVID - D8.2 - Data Management Plan (v1.0). Zenodo. https://doi.org/10.5281/zenodo.6884816 (Annex 2)
[6] Garcia Alvarez, E., & Holub, P. (2022). BY-COVID D8.2.2 Project Data Management Plan initial release and periodic updates M15 (V1.0). Zenodo. https://doi.org/10.5281/zenodo.7476926
[7] https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/temp-form/report/data-management-plan_he_en.docx
[8] European Genome-phenome Archive: Security Overview https://ega-archive.org/files/European_Genome_phenome_Archive_Security_Overview.pdf [accessed 16.11.2022]
[9] BBMRI-ERIC Policy for Access to and Sharing of Biological Samples and Data: https://www.bbmri-eric.eu/wp-content/uploads/AoM_10_8_Access-Policy_FINAL_EU.pdf [accessed 16.11.2022]
[10] OECD Council Recommendations on Health Data Governance [accessed 16.11.2022]
[11] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

4. Description of work accomplished

4.1 Project overview

BY-COVID aims to identify, connect and integrate data for the effective study of the COVID-19 disease and causative agent as well as other infectious diseases. Infectious diseases are complex and their analysis requires data from different disciplines. Thus, BY-COVID links established and emerging research infrastructures and data resources from biomolecular research, public health, clinical research and social sciences and humanities. This is done using standards fully aligned with the European Open Science Cloud (EOSC[12]), such as "Guidance and policy on standards and tools to facilitate sharing and reuse of multimodal data (including imaging), cohort integration, and biosamples"[13] or "Report on data standards for observational and interventional studies, and interoperability between healthcare and research data"[14] and with the European Health Data Space (EHDS) legislative proposal[15], mainly through the outcomes from the currently ongoing pilot project[16].

Three main pillars can be identified in the project methodology:

Mobilisation - Mobilised data are indexed and organised in the COVID-19 Data Portal (https://www.covid19dataportal.org/), where (meta)data are incorporated through a flexible, tiered system for data integration. Data incorporated into the COVID-19 Data Portal will be embedded in the wider EOSC data ecosystem,

[12] EOSC Association: https://eosc.eu/ [accessed 28.06.2024]

[13] Boiten, Jan-Willem, Ohmann, Christian, Adeniran, Ayodeji, Canham, Steve, Cano Abadia, Monica, Chassang, Gauthier, Chiusano, Maria Luisa, David, Romain, Fratelli, Maddalena, Gribbon, Phil, Holub, Petr, Ludwig, Rebecca, Th. Mayrhofer, Michaela, Matei, Mihaela, Merchant, Arshiya, Panagiotopoulou, Maria, Pireddu, Luca, Richard, Audrey, Sanchez Pla, Alex, ... Gorianin, Sergei. (2021). EOSC-Life Guidance and policy on standards and tools to facilitate sharing and reuse of multimodal data (including imaging), cohort integration, and biosamples. Zenodo. https://doi.org/10.5281/zenodo.4591011

[14] Canham, Steve, Ohmann, Christian, Boiten, Jan-Willem, Panagiotopoulou, Maria, Hughes, Nigel, David, Romain, Sanchez Pla, Alex, Maxwell, Lauren, Aerts, Jozef, Facile, Rhonda, Griffon, Nicolas, Saunders, Gary, van Bochove, Kees, & Ewbank, Jonathan. (2021). EOSC-Life Report on data standards for observational and interventional studies, and interoperability between healthcare and research data. Zenodo. https://doi.org/10.5281/zenodo.5810612

[15] Proposal for a Regulation on the European Health Data Space - Analysis of the final compromise text with a view to agreement, https://www.consilium.europa.eu/media/70909/st07553-en24.pdf [accessed 18.07.2024]

[16] European Health Data Space - EHDS HealthDat@EU Pilot, https://www.ehds2pilot.eu/ [accessed 16.11.2022]

establishing guidelines and procedures for FAIR data management and ensuring long term, rapid open access.

BY-COVID allows mobilisation - meaning access and transfer - of data by using a trialled system of SARS-CoV-2 Data Hubs and other existing infrastructures such as the European Nucleotide Archive (ENA), CESSDA social science archives, and biobank catalogues in a "federation of federations", following community practised standards. Hence, the ultimate data responsibility belongs to the data providers, reflected in their own DMP or data best practises guidance.

Connect and expose - A tiered indexing system was developed, protecting truth and privacy, specially in the case of sensitive data. Guidelines for ensuring data interoperability are established by implementing community-driven standards, together with offering support to the discovery of metadata related to sensitive data and the integration of COVID-19 data resources in the COVID-19 Data Portal[17].

Use & analyse data - BY-COVID integrates standardised data management and analysis methods and protocols to ensure FAIR and FAIR-Health[18] are an integral part of the process.

4.2 Data flow

As stated in the previous versions of this DMP, the data flow is distributed among the different technical Work Packages (WP1-WP2-WP3-WP4-WP5), flowing in a domain specific manner. During these 34 months of the project, this flow followed the plan foreseen in month 06, and a brief description on how it was developed can be found in this section.

WP1 established SARS-CoV-2 Data Hubs that handle centralised data, which are not sensitive after de-identification. In line with the described procedure in the first version of this DMP, WP2 gathered the relevant Data Hubs focussing on data other than viral sequences and offering interoperable data as well as open-source solutions for data interoperability.

[17] COVID-19 Data Portal, https://www.covid19dataportal.org/ [accessed 16.11.2022]
[18] Holub P, Kohlmayer F, Prasser F, et al. Enhancing Reuse of Data and Biological Material in Medical Research: From FAIR to FAIR-Health. *Biopreserv Biobank*. 2018;16(2):97-105. https://doi:10.1089/bio.2017.0110

WP2 collected the existing data and metadata controllers, looking at their alignment with proposed standards and the FAIR principles. Discoverability was one of the main activities of this WP, including the integration and harmonisation of non-patient data and an open source harvesting tool for socio-economic data (both non-sensitive data types).

Regarding sensitive data, institutions have been onboarded to the COVID-19 Data Portal and some metadata resources have been harmonised following the specifications from WP3 (see below) and were included in the portal by WP1. One example of this approach can be found in the BY-COVID deliverable D3.2[19], which describes the indexing of social sciences metadata from The Consortium of European Social Science Data Archives (CESSDA[20]) and their integration in the COVID-19 Data Portal. ECRIN has also created an API that links its clinical research Metadata Repository (crMDR)[21] to the COVID-19 Data Portal, contributing to the findability of registered clinical studies on COVID-19 globally.

In addition, WP2 worked on the development of open source federated search tools and conversion tools to format data following the FAIR principles. Of note, all tools for data harvesting or data gathering developed in this WP do not mobilise data, they always stay in the source. A more in depth look at the activities described above is provided in the D2.1[22], which is included in the BY-COVID Zenodo community[23].

While WP2 focused on intradomain harmonisation, cross-domain efforts were done in WP3. The collaborative work of these two WPs resulted in the creation of a FAIRsharing BY-COVID Collection[24], linking the description of the data sources to

[19] Hermjakob, Henning, Kleemola, Mari, Moilanen, Katja, Tuominen, Markus, Sansone, Susanna-Assunta, Lister, Allyson, David, Romain, Panagiotopoulou, Maria, Ohmann, Christian, Belien, Jeroen, Lischke, Julia, Juty, Nick, & Soiland-Reyes, Stian. (2022). BY-COVID D3.2: Implementation of cloud-based, high performance, scalable indexing system. Zenodo. https://doi.org/10.5281/zenodo.7129553

[20] The Consortium of European Social Science Data Archives (CESSDA), https://www.cessda.eu/ [accessed 16.11.2022]

[21] https://crmdr.ecrin.org/

[22] Giles, Tom, Quinlan, Phil, Belien, Jeroen, Lischke, Julia, Portell-Silva, Laura, Capella-Gutierrez, Salvador, Karki, Reagon, Kalaitzi, Vasso, Bernal-Delgado, Enrique, & Keppler, Antje. (2022). BY-COVID- D2.1 - Initial data and metadata harmonisation at domain level to enable fast responses to COVID-19 (V1.0). Zenodo. https://doi.org/10.5281/zenodo.7017728

[23] BY-COVID Zenodo community, https://zenodo.org/communities/bycovid/?page=1&size=20 [accessed 16.11.2022]

[24] BY-COVID data resource collection in FAIRsharing, https://fairsharing.org/3773 [accessed 16.11.2022]

their datasets indexed in the COVID-19 Data Portal. Indeed, FAIRsharing, which was established as the tier 3 of the tiered index described below, is endorsed by the Research Data Alliance (RDA) and used to provide information about the standards, relation and description of data from different sources across all disciplines.

WP3 has developed an infrastructure for the three tiers of the COVID-19 Data Portal discoverability schema, which is based on the metadata model described in D3.1 "Metadata standards. Documentation on metadata standards for inclusion of resources in data portal"[25]. It contains open access tools and workflows ready to integrate resources from several domains into the different tiers, depending on the sensitivity and availability of the (meta)data from each resource. Regarding legal and ethical considerations for WP3, the portal will not include personal data.

The work developed in the BY-COVID project is meant to be used beyond it. To be compliant with this statement, the code and lessons learned from part of the work in WP3 are extended to related portals Early Cause[26] and Pathogens[27] using the "Baseline Portal" software package.

Apart from related portals, the outcomes from this project should be available beyond its end. Thus, for taking care of sustainability, WP3 collaborates with EOSC and RDA initiatives, including the RDA FAIRsharing WG and the Life Science IG.

While data resources and the standards implemented are registered and described in FAIRsharing, tools and other services are included in the Infectious Disease Toolkit (IDTk)[28], developed by WP4. This toolkit provides information and references to open access analysis methods and protocols for the four different domains considered in the BY-COVID project: Pathogen characterisation, Human biomolecular data, Human clinical and health data and Socio-economics data.

---

[25] Hermjakob, Henning, Kleemola, Mari, Moilanen, Katja, Sansone, Susanna-Assunta, Lister, Allyson, David, Romain, Panagiotopoulou, Maria, Ohmann, Christian, Belien, 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

[26] Early Cause, https://portal.earlycause.eu/ [accessed 16.11.2022]

[27] Pathogens portal, https://www.ebi.ac.uk/ena/pathogens/home [accessed 16.11.2022]

[28] Infectious Disease Toolkit, https://www.infectious-diseases-toolkit.org/ [accessed 16.11.2022, under development]

This WP also focused on a semantic interoperability model, for describing datasets and their provenance through Research Objects (RO-Crate)[29] and schema.org[30] vocabularies, along with promoting the use of other relevant terminologies, via FAIRsharing. It also collected the quality measures used to deal with infectious diseases related data and prepared them to be part of the IDTk. Finally, this WP provided advice to the use cases with regard to the FAIRification of their scripts and workflows through RO-Crate and WorkflowHub[31], both aligned with EOSC and the emerging EHDS. This practical FAIRification examples will also become a recipe in the FAIR Cookbook[32,33].

4.2.1 Use cases

The BY-COVID baseline use case (WP5, Task 5.2) aims to provide answers to policy-relevant questions by leveraging observational data (individually-linked sensitive data) obtained from multiple existing data sources (e.g., clinical, administrative and socio-economic data). Aggregated analysis results have been obtained after the federated deployment of the analytic pipeline. This federated infrastructure imposes that the scripts (containerized) move towards the data, while the sensitive data is kept at the premises of each of the nodes (so far institutions that have provided the required data are Sciensano, IACS and THL). As such, all the analyses with individual-level sensitive data are performed at the nodes' premises following their own governance rules and regulatory restrictions. The digital objects constituting the pipeline that are produced, such as the causal model, Common Data Model and synthetic data sets have been published in Zenodo (BY-COVID - WP5 - Baseline Use Case: COVID-19 vaccine effectiveness assessment - Study protocol (zenodo.org)[34]). Note that a Data Management Plan specific to this use case is published in the same repository[35]. Besides, the analysis code and execution

[29] RO-Crate, https://www.researchobject.org/ro-crate/ [accessed 16.11.2022]
[30] schema.org [accessed 16.11.2022]
[31] WorkflowHub, https://workflowhub.eu/ [accessed 16.11.2022]
[32] FAIR Cookbook, https://faircookbook.elixir-europe.org/content/home.html [accessed 06.12.2022]
[33] Philippe Rocca-Serra, Wei Gu, Vassilios Ioannidis, Tooba Abbassi Daloii, Salvador Capella-Gutierrez, Ishwar Chandramouliswaran, Andrea Splendiani, Tony Burdett, Robert T. Giessmann, David Henderson, Dominique Batista, Allyson Lister, Ibrahim Emam, Yojana Gadiya, Lucas Giovanni, Egon Willighagen, Chris Evelo, Alasdair J. G. Gray, Philip Gribbon, ... the FAIR Cookbook Recipes' Authors. (2022). The FAIR Cookbook - the essential resource for and by FAIR doers (1.0). Zenodo. https://doi.org/10.5281/zenodo.7156792
[34] BY-COVID - WP5 - Baseline Use Case: COVID-19 vaccine effectiveness assessment - Study protocol, https://zenodo.org/records/7560731 [accessed 31.07.2024]
[35] Enrique Bernal-Delgado. (2024). BY-COVID - WP5 - Baseline Use Case: COVID-19 vaccine effectiveness assessment - Data Management Plan (2.0.0). Zenodo. https://zenodo.org/records/12636106

instructions of the reproducible analysis pipeline, including a daily matching algorithm and survival analysis, are provided within a GitHub repository https://github.com/by-covid/BY-COVID_WP5_T5.2_baseline-use-case/blob/main/vaccine_effectiveness_analytical_pipeline/documentation/BY-COVID-WP5-BaselineUseCase-VE-documentation-analytical-pipeline.pdf
Unlike foreseen, the integration of the socio-economic data and human genomic data has not been implemented. Instead, interviews and workshops have been held to figure out barriers and facilitators for the use of this type of data in an eventual new project. Deliverable 5.1 “Enriched report viral variants and health outcomes” will contain the lessons learnt (expected in September 2024). This work has not entailed mobilisation of any type of data.

From the ethical and legal perspective for WP5, Task 5.2 includes the movement of scripts and aggregated results between the orchestrator (IACS) and Data Hubs (IACS, Sciensano, THL). Data results received by the orchestrator (IACS) were either: a. aggregated, thus it is impossible to trace back to individuals whose data is processed; b. individual anonymous data in the case of the matrix with matching weights (e.g., inverse probability weights). The local outputs correspond to an excel file used as input for the comparative analysis and the interactive html reports of each main step of the analytical pipeline described in 'Methods analytical pipeline', which is already published[36]. They have been produced using real-world data from public health and the national health system in Aragon (Spain), Brussels and Wallonia (Belgium), and Finland.

As regards security measures, anonymization/pseudonymisation techniques are implemented: a. IACS uses triple anonymisation. b. In Sciensano, a link between the individual data takes place thanks to the use of a pseudonymized national reference number managed by HealthData[37,38,39].

Task 5.3 is focusing on “Vaccine Trials” and its main aim is to implement and pilot a clinical research Data Sharing Repository (crDSR)[40] to support clinical trial data

[36] Meurisse, M., Estupiñán-Romero, F., González-Galindo, J. et al. Federated causal inference based on real-world observational data sources: application to a SARS-CoV-2 vaccine effectiveness assessment. BMC Med Res Methodol 23, 248 (2023). https://doi.org/10.1186/s12874-023-02068-3
[37] healthdata.be | data we care for (sciensano.be)
[38] https://www.iacs.es/bigan/
[39] https://thl.fi/en/about-us/data-protection
[40] https://crdsr.ecrin.org/login

sharing and secondary use. The first use case of the repository is a collaboration with the VACCELERATE project[41]. The VACCELERATE EU-COVAT-1-AGED[42] trial data has entered the process of anonymisation following the completion of the trial. External expertise was needed for this step and the people behind the Amnesia anonymisation tool[43] have been engaged to develop together with the clinical trial data centre an appropriate anonymisation protocol. Next step: Signature of contractual agreements and transfer of the anonymised data to the TSD (University of Oslo) TRE in Norway. The negotiation process of the Data Transfer Agreement (DTA) has been initiated.

In parallel, 112 COVID-19 studies were identified as potential candidates for sharing data based on their Data Sharing Statements in clinical trial registries (see[44]). A survey was prepared in July 2024 and will be circulated to the study contact persons to evaluate their real willingness to share and under which conditions. Studies willing to collaborate with the crDSR will be approached in order to facilitate the secondary use of the datasets.

Deliverable 5.2 “Secondary use of vaccine trial data and biosamples” (due in September 2024) will describe the processes of data transfer and data use established in the crDSR and report the status of the data sharing coming from VACCELERATE. In addition, it will report on the results of the survey providing an inventory of initiatives willing to share COVID-19 data and describing the requirements for access.

Task 5.4 aims to “Use of waste water surveillance to detect epidemic hotspots”. This use case is particularly focused on data sharing for the early detection, prevention and real-time monitoring of epidemic threats and outbreaks. They are closely following other initiatives with similar objectives, such as the recently launched “Global Consortium for Wastewater and Environmental Surveillance for Public Health” (GLOWACON[45]). In collaboration with WP4, Task 5.4 leads

---

[41] https://vaccelerate.eu/

[42] 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;41(48):7166-7175. doi: 10.1016/j.vaccine.2023.10.029.

[43] https://amnesia.openaire.eu/

[44] https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-024-02168-8

[45] https://health.ec.europa.eu/latest-updates/launching-glowacon-global-initiative-wastewater-surveillance-public-health-2024-03-21_en

contributed to the publication of data quality requirements for the dataset creation, from sample collection to data analysis (see [46]).
In addition, "Mechanistic analysis via the COVID-19 Disease Mapping" is another objective of Task 5.4. Its original goal was to couple omics datasets with viral strain information, bioimaging and structural bioactivity data that are suitable to be projected on the COVID-19 disease map. The final deliverable resulted in Deliverable D5.3a "Mechanistic analyses via the COVID-19 Disease Map", mapping to the C19DMap provides context, connecting particular host and viral molecules into a bigger picture.

When it comes to DMP, the work done has implied plugging digital objects from other open access sources in a theoretical disease map. The work has consisted of:

- Connecting SARS-CoV-2 datasets on the BioImage Archive of drug repurposing analysis[47] and the Electron Microscopy Public Image Archive (EMPIAR) set of virus protein structures[48].
- Developing data overlays for a public proteomics dataset representing infection by three different SARS-CoV-2 strains in the Calu-3 cell line (Mezler et al. 2023[49]).
- Prototyping a Galaxy workflow capable of visualising transcriptomic data on the COVID-19 Disease Map.

The domain specific use cases worked with re-used data, as no new data were generated in BY-COVID. Thus, they focused on the secondary use of different data types: data objects from clinical trials (e.g., IPD datasets, clinical trial protocols, statistical analysis plans, result summaries, published papers, informed consent forms...); data from viral protein structures, and observational data from disease registries, healthcare and socioeconomic sources.

Regarding data utility within the project, re-using these data is needed in order to test one of its main objectives, which is mobilising and making data accessible to be ready for future pandemics. This objective directly links to the utility of the data outside the project.

---

[46] https://www.infectious-diseases-toolkit.org/pathogen-characterisation/quality-control
[47]BioImage Archive: https://www.ebi.ac.uk/biostudies/bioimages/studies/S-BIAD29 [accessed 24/01/2024]
[48] EMPIAR: https://www.ebi.ac.uk/emdb/search/database:EMPIAR%20AND%20SARS-CoV-2 [accessed 24/01/2024]
[49] Mezler et al: https://doi.org/10.1016/j.mcpro.2023.100537 [accessed 24/01/2024]

Data related to vaccine trials will be centralised through the COVID-19 repository developed by ECRIN and using a secure platform for sensitive data sharing (TSD) from the University of Oslo[50].

As previously stated, to increase data re-use is key for this project, thus, all needed documentation for facilitating it will be provided as outcomes of BY-COVID. Cross-fertilisation with WP4 has enabled for the FAIRfication of the methodological and analytical workflows. In addition, a general provenance information model for infectious diseases is provided as part of the work from WP4[51], that was applied to the baseline and the clinical trials use cases; and it is available for the others to adopt it. Finally, to make the data used in the use case findable, the ECRIN repository will be connected to the COVID-19 Data Portal.

None of the WP5 use cases are actively recruiting/involving human participants as part of BY-COVID. They are instead re-using already existing data made available under specified conditions by data providers. The same approach applies to WP3.

4.3 Open Science and FAIR sharing

In the above summary of data flow, it is already possible to see that resources and open access tools, such as Zenodo (an OpenAIRE initiative), FAIRsharing or RO-Crate, and FAIR practices, as documented in the IDTk and FAIR Cookbook are being used throughout the project as planned, ensuring findability, discoverability and effective sharing. Furthermore, EOSC practises (e.g., EOSC Interoperability framework[52], EOSC Enhance D4.3 "Analysis of existing research data cataloguing efforts towards integrated discovery"[53]) and RDA recommendations[54] are followed, thanks to the joint work with both initiatives.

The European COVID-19 Data Platform is established as the main portal to ensure accessibility. In addition, metadata to those sensitive data that fall under the

[50] https://crdsr.ecrin.org/login
[51] Wittner, R., Soiland-Reyes, S., Leo, S., Meurisse, M., & Hermjakob, H. (2024). BY-COVID D4.3 Provenance model for infectious diseases. Zenodo. https://doi.org/10.5281/zenodo.10927253
[52] EOSC-IF: https://eosc-portal.eu/eosc-interoperability-framework [accessed 16.11.2022]
[53] Carole Goble, & Nick Juty. (2021). Analysis of existing research data cataloguing efforts towards integrated discovery. Zenodo. https://doi.org/10.5281/zenodo.4693217
[54] RDA COVID-19 Working Group. (2020). RDA COVID-19 Recommendations and Guidelines on Data Sharing (1.0). https://doi.org/10.15497/rda00052

GDPR[55] will go in FAIRsharing, and in the IDTk (currently under development) as relevant.

As already highlighted in the first version of the DMP, this project is driving data use and re-use by linking FAIR open data to workflow environments and providing access to analysis and visualisation tools building trust and reproducibility with provenance and quality assurance mechanisms. Thus, the project doesn't generate any new sensitive patient level data and each participant institution is responsible for accounting for national legislations, the administrative provisions and the implemented data access procedures.

5. Conclusions

Overall, during BY-COVID, data access committees and processes for each dataset remain accountable for the project's operations. Currently, no active participant recruitment/engagement or AI systems are used. This new version of the DMP gives a more detailed description of the tools and guidelines followed within the BY-COVID project with a particular focus on the use cases, thanks to the ongoing work and the already submitted deliverables and milestones[56].

6. Sustainability

The integrated BY-COVID ecosystem is based on established services and resources, ensuring sustainability beyond the lifetime of the project. The adoption of open source platforms on commonly used cloud-based infrastructures and connection of the components through open standards supports long-term sustainability in a distributed landscape.
In addition, a report entirely dedicated to sustainability matters is being written in parallel to this version of the DMP and will be published at the same time under the title "D8.3 Report on sustainability plans".

7. Impact

The adoption of Data Management best practices and Open Science principles collected in this and previous version of the DMP has optimised the impact and

[55] General Data Protection Regulation (GDPR), https://eur-lex.europa.eu/eli/reg/2016/679/oj [accessed 16.11.2022]
[56] BY-COVID Outcomes, https://by-covid.org/outcomes/ [accessed 28.06.2024]

reuse of the data from the project. Thus, a sound and established DMP constituted the foundation for a fruitful collaboration across WPs, with internal and external experts and stakeholders, including RDA, EOSC and FAIRsharing.