

Deliverable D2.2

Data Access and Transfer across research domains and jurisdictions

Project Title (grant agreement No)	BeYond-COVID Grant Agreement 101046203		
Project Acronym (EC Call)	BY-COVID		
WP No & Title	WP2: Accessing heterogeneous data across domains and jurisdictions for enabling the downstream processing of COVID-19 and future pandemic episodes data		
WP Leaders	Alfonso Valencia [BSC] Salvador Capella-Gutierrez [BSC] Antje Keppler [EuroBioImaging] Aastha Mathur [EuroBioImaging]		
Deliverable Lead Beneficiary	Fundació Centre de Regulació Genòmica [CRG]		
Contractual delivery date	30/06/2023	Actual Delivery date	30/06/2023
Delayed	No		
Partner(s) contributing to this deliverable	CRG, INFRAFRONTIER, Sciensano, BSC, ELIXIR-NL/VUmc, ERINHA, Lygature, CESSDA, EMBL-EBI, BBMRI-ERIC		
Authors	Aina Jené Cortada (CRG)		
Contributors	Philipp Gormanns (INFRAFRONTIER) Andrea Furlani (INFRAFRONTIER) Iris Van Dam (Sciensano) Laura Portell-Silva (BSC) Jeroen Belien (ELIXIR-NL/VUmc) Romain David (ERINHA) Robin Navest (Lygature) Matti Heinonen (CESSDA/TAU-FSD) Markus Tuominen (CESSDA/TAU-FSD) Dimitra Kondyli (CESSDA-EKKE) Mallory Freeberg (EMBL-EBI)		
Acknowledgements (not grant participants)	N/A		
Reviewers	Salvador Capella-Gutierrez [BSC] Aastha Mathur [EuroBioImaging]		



Mari Kleemola [CESSDA/TAU-FSD]
 Ilaria Colussi [BBMRI-ERIC]
 Carla Giuffre [ELIXIR Hub]
 Niklas Blomberg [ELIXIR Hub]

Log of changes

Date	Mvm	Who	Description
13/04/2023		Aina Jené (CRG)	Initial structure of the deliverable from TOC discussed with WP2 April meeting
23/05/2023		WP2	First draft sent to WP2 where all contributors suggested improvements
19/06/2023		Aina Jené (CRG)	Second round of writing for deliverable 2.2
22/06/2023		WP2	Second draft sent to WP2 where all contributors suggested improvements
25/05/2023		Aina Jené (CRG)	Final writing and harmonisation before submit
26/06/2023		WP Leads	Quality check
30/06/2023		Aina Jené (CRG)	Deliverable submitted



Disclaimer

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



Table of contents

1. Executive Summary	4
2. Contribution towards project objectives	5
Objective 1	5
Objective 2	5
Objective 3	6
Objective 4	6
Objective 5	6
3. Introduction	7
4. Methods	8
5. Description of work accomplished	9
5.1 Type of data sources and their characteristics	9
5.1.1 Non-patient Related data	12
5.1.2 Human/Patient Bio-Molecular data	13
5.1.3 Human/Patient Clinical and Health data	14
5.1.4 Socio-Economic data sources	14
5.2 Barriers for research data access and transfer	15
5.2.1 Data Access	15
5.2.1.1 Legal barriers	17
5.2.1.2 Organisational barriers	19
5.2.1.3 Technical barriers	20
5.2.2 Data Transfer	21
5.2.2.1 Legal barriers	22
5.2.2.2 Organisational barriers	22
5.2.2.3 Technical barriers	23
6. Results	25
6.1 Identified legal barriers to data access and transfer	25
6.2 Identified organisational barriers to data access and transfer	25
6.3 Identified technical barriers to data access and transfer	26
7. Discussion	26
8. Conclusions	29
9. Next steps	29
10. Impact	30
11. Deviation from Description of Action	31
12. Annex1 - Survey	31



1. Executive Summary

BY-COVID Work Package (WP) 2 brings together datasets and catalogues of metadata across domains, assesses data governance and access procedures, and contributes to increase the FAIRness level of different data sources. It aims to align metadata descriptions and other semantic information, first within domains (e.g., biomolecular and imaging, clinical and health, survey, etc), and in a second stage (in alignment with WP3 developments) across domains to propose a reference catalogue with harmonised metadata descriptions, which can facilitate data access and eventually enable data mobilisation for the analysis workflows specified by WP4 .

Deliverable 2.2 (D2.2) focuses on data access and transfer across research domains and jurisdictions. Originally intended to address technical, legal, and organisational barriers in data sources, its scope expanded to include all types of data that also link to pathogen variants via SARS-CoV-2 DataHubs (WP1). The objectives of D2.2 are two-fold: ensuring compliance with data privacy principles for controlled access data and proposing strategies to enhance data interoperability. Efforts included developing a survey, conducting ad-hoc meetings with key stakeholders, and aligning with the FAIR data principles. The deliverable contributes to ongoing efforts to facilitate efficient and responsible data sharing, which should enable faster and coordinated responses in future pandemics.

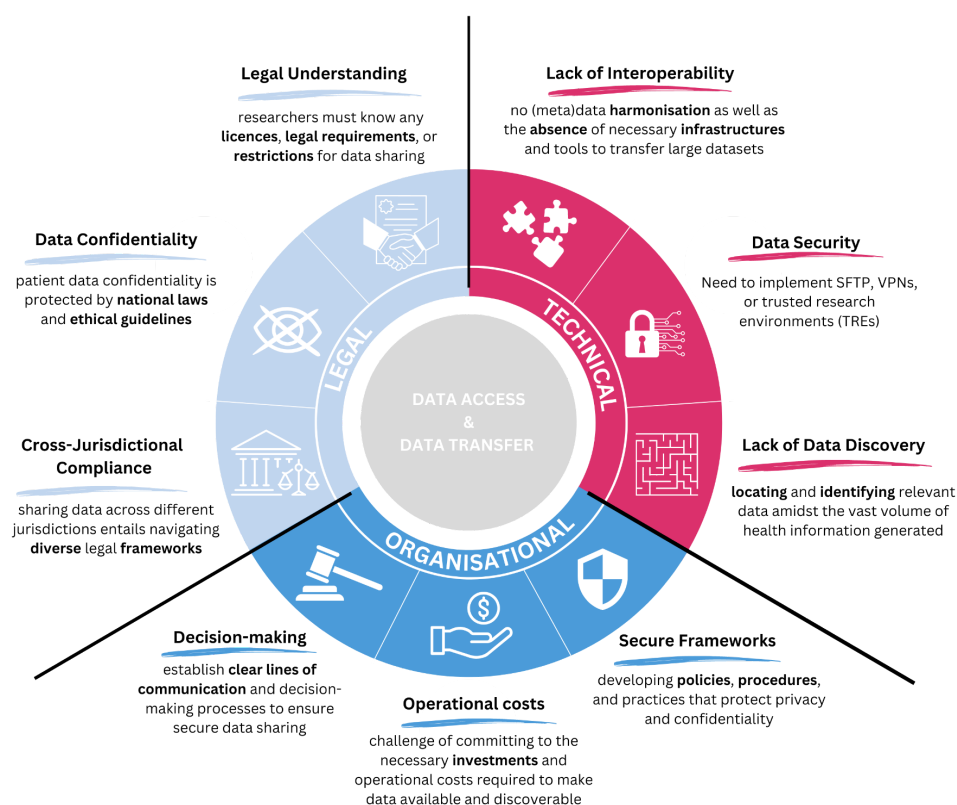


Figure 1: Summary of the barriers identified for Deliverable 2.2 in the context of the BY-COVID project.



2. Contribution towards project objectives

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

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



<p>Objective 3</p> <p>Link FAIR data and metadata on SARS-CoV-2 and COVID-19</p>	<p>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.</p>	No
	<p>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.</p>	Yes
	<p>3. A provenance framework for researchers and policy-makers that enables trust in results and credit to data submitters, workflow contributors and participant sources.</p>	No
<p>Objective 4</p> <p>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</p>	<p>1. Broad uptake of viral <i>Data Hubs</i> across Europe deliver an order-of-magnitude increase in open viral variant detection and sharing.</p>	No
	<p>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.</p>	No
<p>Objective 5</p> <p>Contribute to the Horizon Europe European Open Science Cloud (EOSC) Partnership and European Health Data Space (EHDS)</p>	<p>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).</p>	Yes
	<p>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.</p>	Yes
	<p>3. Alignment (both policy and implementation routes) will have been achieved between the data governance strategies for routinely collected health</p>	No



	data in the EHDS initiative, including the TEHDAS Joint Action and future EHDS Pilot Actions.	
	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

The BY-COVID project is an international effort aimed at developing a coordinated framework for data sharing and analysis to better understand and take action against the COVID-19. Work Package 2 (WP2) is focused on consolidating and standardising data sources and catalogues across domains, including biomolecular and imaging, clinical and health, survey, and other relevant areas. This work builds on previous efforts within WP2, including Deliverable 2.1¹ (D2.1), Milestone 2.1² (M2.1), and Milestone 2.2³ (M2.2).

- D2.1 (*Initial data and metadata harmonisation at domain level to enable fast responses to COVID-19*) explained the infrastructure required to facilitate access to a broad portfolio of data sources relevant to COVID-19 in preparation for future outbreak responses. This deliverable enabled the alignment of metadata descriptions and semantic information within domains.
- M2.1 (*Identified data sources have been registered in the BY-COVID reference catalogue*) focused on identifying data sources to be registered in the COVID-19 Data Portal. This milestone was a critical step in consolidating and standardising data sources across domains, enabling more efficient access to data and facilitating interoperability.
- M2.2 (*Identified the preferred mechanisms for data access and use of Real-World Data (RDW)*) was another key component of WP2's efforts to enable efficient and secure data sharing. This milestone focused on identifying the preferred

¹ 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>

² Salvador Capella-Gutierrez, Alfonso Valencia, Aastha Mathur, Antje Keppler, & Laura Portell Silva. (2022). BY-COVID Work Package 2 List of Resources (V1.0) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.6939376>

³ Estupiñán-Romero, Francisco, Launa-Garcés, Ramón, Van Dam, Iris, Cosgrove, Shona, Van Goethem, Nina, & Bernal-Delgado, Enrique. (2023). BY-COVID Milestone 2.2 Mechanism for data access and use of routine real-world data. Zenodo. <https://doi.org/10.5281/zenodo.7944195>



mechanisms for accessing RWD while ensuring compliance with local, national, and European regulations and participant consent agreements.

Beyond WP2, the BY-COVID project also includes Work Packages 3 (WP3), which respectively focus on the development of the COVID-19 Data Portal, including the establishment of a metadata catalogue for promoting data sharing. These efforts, along with WP2 work, contribute to enable effective and responsible data sharing across jurisdictions.

Deliverable 2.2 (D2.2) (*Data Access and Transfer across Research Domains and Jurisdictions*) was originally intended to describe the technical, legal, and organisational barriers encountered across data sources. Its purpose was to identify solutions for data access, including procedures for managing user permissions for controlled access data. However, the scope of the deliverable expanded to encompass all types of data. As a result, the primary objective of D2.2 is now two-fold. First, it aims to ensure compliance with data privacy and data protection principles for controlled access data, which all European researchers are obligated to follow. Second, the deliverable proposes strategies for enhancing data interoperability - a key component of the FAIR principles⁴ - by implementing community-driven standards applicable to various types of data, such as those developed by the Global Alliance for Genomics and Health⁵ (GA4GH), for all types of data.

Overall, this deliverable represents an important step forward in the BY-COVID project's efforts to establish a coordinated framework for data access and sharing. By consolidating data governance procedures, promoting interoperability, and addressing access challenges, this deliverable contributes to enable more efficient and responsible sharing of critical research data.

4. Methods

The project contributes towards the mobilisation of existing data sources by putting the associated metadata together in a catalogue. It then utilises the COVID-19 Data Portal⁶ to connect and expose these data sources for research purposes.

D2.2 focuses on data access and transfer across research domains and jurisdictions. For that, we focused on the sources described in the List of sources⁷ identified in M2.1, and

⁴ FAIR Principles: <https://www.go-fair.org/fair-principles/> [accessed 30.06.2023]

⁵ "GA4GH: International policies and standards for data sharing across genomic research and healthcare.", [https://www.cell.com/cell-genomics/fulltext/S2666-979X\(21\)00036-7](https://www.cell.com/cell-genomics/fulltext/S2666-979X(21)00036-7)

⁶ COVID-19 Data portal: <https://www.covid19dataportal.org/> [accessed 30.06.2023]

⁷ Salvador Capella-Gutierrez, Alfonso Valencia, Aastha Mathur, Antje Keppler, & Laura Portell Silva. (2022). BY-COVID Work Package 2 List of Resources (V1.0) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.6939376>



took into account the data and metadata available in D2.1, as well as the mechanisms identified in M2.2.

To gather comprehensive information about the mentioned data sources enlisted in the List of sources⁸, the members of Task 2.2 undertook an extensive survey development process. The survey was carefully crafted (Annex 1), encompassing targeted questions that specifically addressed the nature of the data resources and the specific types of SARS-CoV-2 data housed within their respective repositories. In addition to gathering information about the data sources, we took a proactive approach by seeking feedback from the participants regarding any barriers they encountered in accessing and sharing data. We encouraged them to identify any challenges that might have been overlooked during our initial assessment.

In addition, scheduled ad-hoc meetings were conducted with three distinct data providers to further examine the challenges associated with accessing and mobilising data from their respective data sources. These meetings served as valuable opportunities to engage in in-depth discussions with key stakeholders and gain insights into the specific obstacles encountered during the data sharing process. The valuable feedback received from these interactions contributed to the development of informed strategies and recommendations outlined in this deliverable.

The insights collected through this survey, which have been discussed in the [Results](#) section of this deliverable, have allowed us, in collaboration with the project partners, to gain a deeper understanding of the current landscape surrounding SARS-CoV-2 data sources. Moreover, it has shed light on the significant challenges faced by researchers when it comes to accessing and sharing critical data, and allows us to propose specific actions in preparation for future pandemic episodes.

5. Description of work accomplished

5.1 Type of data sources and their characteristics

Data sources play a critical role in various fields, including healthcare and research. Different types of data sources exist, and each type has its characteristics that impact how data is reused.

In the BY-COVID project, we have extensively analysed a wide range of data sources, comprising a total of 92. These sources have been classified into four distinct categories

⁸ Salvador Capella-Gutierrez, Alfonso Valencia, Aastha Mathur, Antje Keppler, & Laura Portell Silva. (2022). BY-COVID Work Package 2 List of Resources (V1.0) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.6939376>



based on their data types (Figure 2A). However, it is important to note that not all of these data sources are represented in this specific deliverable. The reason behind this is that the responsible teams for those data sources are not actively involved in the work package. For this reason, we have focused on a total of 26 data sources, identified and active participants of WP2 tasks. The breakdown of data sources by data type categories is shown in Figure 2B, and each data type is described in more detail in the following sections.

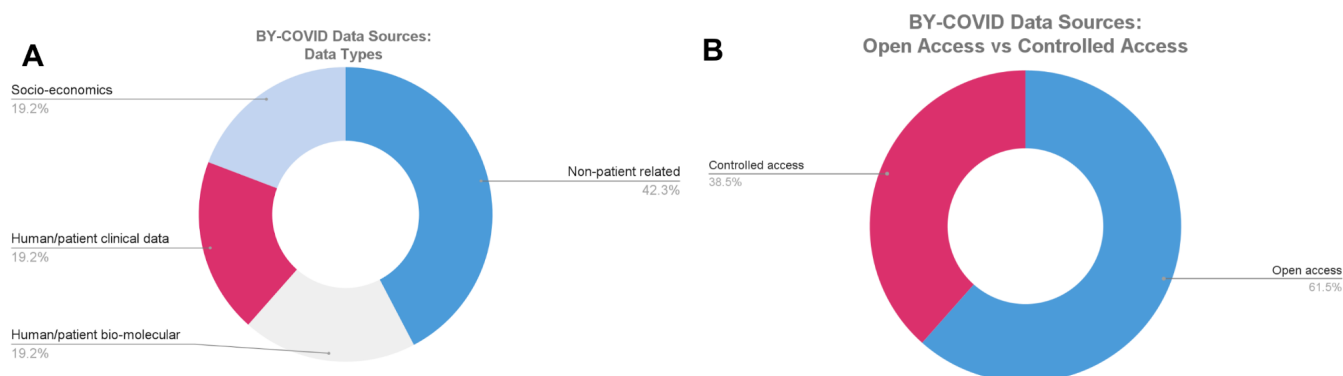


Figure 2: A) Proportion of data sources available in the BY-COVID Project divided by its data types: Non-patient related data (11 data sources (DS)), Human/patient bio-molecular data (5 DS), Human/patient clinical data (5 DS), and Socio-economics data (5 DS). B) Proportion of data sources available in the BY-COVID Project divided by access type: Open access (16 DS) vs Controlled access (10 DS).

Figure 2A provides an overview of the data sources available for the BY-COVID project. It reveals that the majority of the data sources (42.3%) pertain to non-patient related data. Human/patient bio-molecular data, human/patient clinical and health data, and socio-economic data each constitute 19.2% of the sources, with an equal distribution.








Furthermore, Figure 2B illustrates that over 60% of the data sources host open access data. However, Table 1, which presents the findings from the survey conducted as described in the methods section, demonstrates a balanced representation with 50% of the data sources offering controlled access and the remaining 50% providing open access. This is due to the fact that not all data sources completed the survey.

Table 1: Visualisation of the results obtained through the survey used to gather information for D2.2.

From the 26 data sources mentioned, 11 completed the survey.

Representation of data source by data type

Non-patient related data sources	●●●●●●●●●●●●
Human/patient bio-molecular data sources	●●●●●●●●●●●●

Human/patient clinical and health data sources	
Socio-economics data sources	
<i>Representation of data sources by access type</i>	
Open access	
Controlled access	
Both	
<i>Representation of data sources indexed in the COVID-19 Data Portal</i>	
Yes	
In progress	

Open data refers to data and metadata that is publicly available and can be accessed without restrictions. This type of data plays a crucial role in advancing research efforts. By making data openly available, researchers can collaborate, share findings, and work towards developing effective treatments and preventive measures. This type of data is mainly represented by BY-COVID tasks 2.1 and 2.4, non-patient related data⁹ and socio-economic data¹⁰, respectively.

However, it is important to note that certain data, currently regulated and under access control, also contribute towards these objectives but cannot be openly shared due to existing regulations, this type of data is controlled access data. Controlled access data, in its broader context, encompasses both the underlying closed data and the associated metadata. The closed data refers to the actual information that is subject to specific access and usage conditions, such as data use agreements and restricted access policies. This closed data may include sensitive and valuable information like pseudonymised patient data, genetic data, and other confidential information that necessitates stringent privacy protections.

On the other hand, the associated metadata, which provides general information about the controlled access data, can be published openly, as long as it is not sensitive information, allowing anonymous users to access it at a higher level. Then, with this public information researchers can decide to apply access to the data. If approved, when users identify themselves, they may gain access to more detailed or deeper levels of metadata.

⁹ Non-patient related data:

https://docs.google.com/document/d/1OUZwyAK_8JkWOO32HjIa6PI9FvSr9ApSXRSYk2yr_28/edit#heading=h.y9goh3be2nq0

¹⁰ Socio-economic data:

https://docs.google.com/document/d/1OUZwyAK_8JkWOO32HjIa6PI9FvSr9ApSXRSYk2yr_28/edit#heading=h.6thpb9avbbfb



In the context of BY-COVID, this type of data is mainly represented by tasks 2.2 and 2.3, human/patient biomolecular data¹¹ and human/patient clinical and health data¹², respectively. Accessing and utilising these data types for research purposes requires careful adherence to legal and ethical requirements, as well as implementing appropriate security measures to safeguard data confidentiality.

5.1.1 Non-patient Related data

Within WP2, a substantial portion of the data relates to non-patient related information. This category encompasses diverse data arising from structural analyses, bioimaging studies, and bioactivity investigations. As this data does not involve sensitive patient information, it is openly available through dedicated repositories that utilise various metadata standards and controlled vocabularies. In Table 2, we can see the work accomplished by Task 2.1 (T2.1) for listing the data sources with non-patient related data.

Table 2: Overview of ontologies and identifiers used by data resources associated with T2.1 as of June 2023.

Entity	Ontology/Dictionary Metadata
Human Proteins	HUGO Uniprot, ChEMBL
Assays	ChEMBL IC50 vals, Type, Organism, pChEMBL, FBbi ¹³ , EDAM BioImaging ¹⁴
Chemicals	ChEMBL, PubChem CAS, SMILES, Physicochemical props.
Pathways	Reactome, WikiPathway Pathway IDs
Mechanism of action / Process	Gene Ontology GO components
Disease Indications	MeSH, OMIM EFO ids, synonyms

The majority of the non-patient data is accessible through open access channels, and there are already established programmatic access mechanisms in place, such as [REST APIs](#), to retrieve the data. However, the current setup inadvertently disincentives researchers from

¹¹ Human/patient biomolecular data:

https://docs.google.com/document/d/1OUZwyAK_8JkWOQ32HjIa6PI9FvSr9ApSXRSYk2yr_28/edit#heading=h.otrrtj49excj

¹² human/patient clinical and health data:

https://docs.google.com/document/d/1OUZwyAK_8JkWOQ32HjIa6PI9FvSr9ApSXRSYk2yr_28/edit#heading=h.71u8nndd4s88j

¹³ FBbi: <https://www.ebi.ac.uk/ols4/ontologies/fbbi> [accessed 30.06.2023]

¹⁴ EDAM BioImaging Github instance: <https://github.com/edamontology/edam-bioimaging> [accessed 30.06.2023]



actively submitting their data, limiting the amount and diversity of research data in the Open sphere. To address this issue, efforts are underway to harmonise and encourage data submission, particularly focusing on data from various subdomains like structural, bioimaging, and bioactivity data sources. By clarifying the data submission process and making it simpler and more user-friendly, the aim is to actively encourage researchers to contribute their data to the Open sphere. This proactive approach will foster a greater volume and variety of research data, ultimately enhancing collaboration, knowledge sharing, and scientific advancement.

The data accumulation and harmonisation process outlined in D2.1 (section 4.1.1)¹⁵ was successfully implemented within the Mpox¹⁶ use-case. This reproducible workflow proved highly efficient, enabling us to quickly achieve the desired progress while saving significant time and effort. All programming scripts and data used in this process are securely maintained in the GitHub repository¹⁷.

5.1.2 Human/Patient Bio-Molecular data

Human/patient bio-molecular data include genetic, genomic, proteomic, and metabolomic data. These data provide information about the molecular processes and pathways that underlie biological functions and diseases. Bio-molecular data is typically collected from blood, tissue, or saliva samples.

The data sources employed in this project are subject to stringent control measures, as they are collected from individuals who have given their informed consent and are used for specific research objectives. Within this deliverable, we have taken into account a total of five data sources, which are comprehensively described in D2.1 (section 4.1.2)¹⁸. Nevertheless, subsequent to that, the members of task 2.2 have actively worked on gathering further information pertaining to data access and transfer, resulting in the production of this deliverable.

¹⁵ 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>

¹⁶ "Mpox Knowledge Graph: a comprehensive representation embedding chemical entities and associated biology of Mpox": <https://doi.org/10.1093/bioadv/vbad045>

¹⁷ Source code and data repository for paper titled "Monkeypox Knowledge Graph: A comprehensive representation embedding chemical entities and associated biology of Monkeypox " on Github <https://github.com/Fraunhofer-ITMP/mpox-kg> [accessed 30.06.2023]

¹⁸ 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>



Moreover, it is worth highlighting that the European Genome-phenome Archive¹⁹ (EGA) is currently collaborating closely with members of WP3 to utilise EGA's SARS-CoV-2 released studies as a compelling use case for demonstrating access attributes within the COVID-19 Data Portal. The integration of EGA studies in the portal serves as an important step towards enhancing the functionality and utility of the platform for researchers and stakeholders involved in combating COVID-19.

5.1.3 Human/Patient Clinical and Health data

Human/patient clinical and health data include any type of observational data source that collects sensitive data. These data sources provide information about a patient's health status, medical history, symptoms, diagnoses, treatments, place of residence, place of treatment, and outcomes. Similar to bio-molecular data, clinical and health data are highly controlled, and access is restricted to authorised personnel and require anonymisation or pseudonymisation.

Task 2.3 has made significant progress in achieving milestone M2.2, which focuses on patient clinical and health data and serves as a crucial foundation for Deliverable 2.2. This milestone offers insights into the challenges surrounding the access and utilisation of sensitive Real World Data (RWD). It specifically addresses the reuse of routine health data, rather than data primarily collected for research purposes.

Through this milestone, we have established the definition of key concepts and clarified the legal framework governing access and reuse for research purposes. Additionally, task 2.3 has developed a comprehensive template for describing and evaluating policies and procedures related to the access and use of RWD. These achievements lay a strong foundation for promoting responsible and effective data access and reuse in the research community.

5.1.4 Socio-Economic data sources

Socio-economic data include demographic, social, political, geographic, economic, and environmental data. These data sources provide information about social and economic factors that influence health and healthcare outcomes. Socio-economic data can provide important insights for infectious diseases research, as well as policy for disease preparedness and mitigation. For example, such data can help understand how infectious diseases spread among and impact different segments of society such as different socioeconomic groups or economic sectors; how effective different government strategies are at combating the spread of a virus; or how political and psychological factors can affect vaccine uptake; provide the socioeconomic elements on which policies and communication actions should focus and/or adapt in order to be more efficient.

¹⁹ EGA <https://ega-archive.org/> [accessed 30.06.2023]








Socio-economic data sources are generally public and accessible, but authorisation or licensing agreements may be required to access some datasets²⁰. While the complete datasets may not always be openly available, the metadata is usually freely accessible. Socio-economic data sources in the BY-COVID project (for example, the CESSDA Data Catalogue) also offer programmatic access to their metadata using standards like the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). This enables the automated harvesting, harmonisation, and transformation of metadata into the OmicsDI format, which can then be utilised by the COVID-19 Data Portal.

5.2 Barriers for research data access and transfer

As previously mentioned, the initial scope of this deliverable solely encompassed controlled access data. However, following a subsequent discussion, it was collectively decided to expand its coverage to include all types of data. In light of this decision, it was agreed to divide this section of the deliverable into two parts: data access and data transfer. This division allows for a comprehensive discussion of the barriers identified in each part, ensuring that specific barriers related to open access data sources are also addressed.

Table 3: Visualisation of the barriers reported through the survey used to gather information for D2.2.

From the 26 data sources mentioned, 11 completed the survey.

<i>Representation of barriers reported by data sources through the survey</i>	
Legal barriers	
Organisational barriers	
Technical barriers	
<i>Representation of data sources by data accessibility</i>	
Data source already sharing data	
Data source not sharing data	

5.2.1 Data Access

In the context of biomedical research for COVID-19, data access plays a crucial role in facilitating scientific advancements and combating the ongoing pandemic. This data can be provided by a variety of sources, including biobanks, clinical trials, and other research studies. Researchers require access to this data to answer research questions, identify potential areas of investigation and advance in biomedical research, like developing new

²⁰ Bolton, Sharon, Jakobsen, Morten, & Storviken, Silje. (2022). Specification for interoperable access conditions in CDC. <https://doi.org/10.5281/zenodo.7252394> [accessed 30.06.23]

vaccines such as the SARS-CoV-2 vaccines created during the COVID-19 outbreak in 2020 and the following years.

In recent years, there has been a growing recognition of the importance of availability, data sharing, and transparency in the scientific community. Many funding agencies and scientific journals now require researchers to make their datasets reusable to other researchers for further analysis and replication. The aforementioned evolutions contributed to the growing number of repositories²¹, either institutional or thematic ones.

There are several steps involved in the data access process for biomedical research (Figure 3).

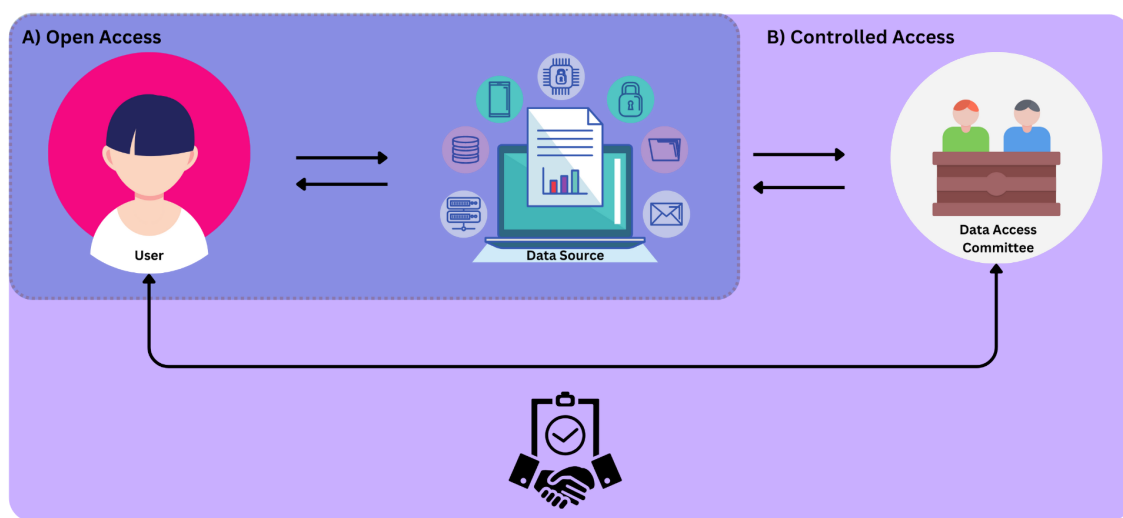


Figure 3: Steps involved during the data access and transfer procedures taking into account the different types of data access mentioned in this deliverable.

The first step is identifying the dataset or resource that is needed for the research study. This may involve searching electronic databases, contacting data repositories or biobanks, or collaborating with other researchers who have access to the required data.

Once the dataset has been identified, the researcher should verify whether the access is open or controlled. If an open access is provided, thus data is downloadable for the researchers without any constraint; and they do not need to go for an access request and associated access decision. This usually occurs when research does not entail personal data (thus, anonymous, or anonymised, or aggregated).

On the other hand, if a controlled access is provided, access is subject to precise procedures and measures. It is the standard approach where personal data is concerned, given the need to maintain data protection compliance. Access is granted through policies

²¹ List of BY-COVID data sources in FAIRsharing: <https://fairsharing.org/3773> [accessed 30.06.23]






establishing who accesses data, how, according to which protocols, and it is controlled by oversight bodies, in order to avoid risks of unauthorised access. An important document of the controlled-access model is the Use and Access Policy: it sets rules for data access and limits access to the database to the approved users and under specific terms and conditions. Furthermore, this type of policy contains rules as to how the access control is implemented, and this occurs most of the time through a Data Access Committee (DAC) that reviews data access requests from potential users and approves or refuses them. After approval, parties sign a Data Transfer Agreement, which is a legally binding document indicating terms and conditions governing the researcher's data access are defining responsibilities.

5.2.1.1 Legal barriers

With the survey and the ad-hoc discussions mentioned in the methods, we were able to recognise several legal barriers that must be navigated to protect individual privacy, ensure data integrity, and comply with relevant regulations.

Table 4: Visualisation of the legal barriers reported through the survey used to gather information for D2.2.

From the 26 data sources mentioned, 11 completed the survey.

GDPR implementation	
Specific legislations from European countries	
Data Anonymisation or Pseudonymisation	
Patient's Consent Agreement or similar	
N/A	

As we can see in Table 4, these barriers include regulatory compliance, variations in interpreting data protection regulations like the General Data Protection Regulation (GDPR), disparities in national laws, and the process of data anonymisation or Pseudonymisation. Next, we will include some of the key legal considerations for ensuring the security and privacy of data access:

- The confidentiality of patient data is protected by a number of national laws and regulations, the General Data Protection Regulation (GDPR) in Europe and Ethical

Guidelines such as the Helsinki Declaration²². Moreover, it's worth mentioning the European Network of Research Ethics Committees - EUREC²³, can be a valuable source of information in order to conduct responsible research and to meet new challenges and emerging ethical issues.

Researchers must adhere to these regulations as well as must obtain the necessary approvals and permissions to access the data. For this reason, regulatory compliance is the first legal barrier identified when accessing controlled access data.

- However, this brings us to the following barrier, the lack of common European interpretation of the GDPR, as well as the different national laws and regulations, such as what constitutes “sufficient anonymisation”, “pseudonymisation” or “secondary use of data”. These disagreements have been extensively discussed in other European projects such as the deliverable 5.1 of the Towards European Health Data Space (TEHDAS) project²⁴.
- To access controlled access data, researchers must obtain appropriate permissions and comply with data use agreements that specify how the data can be accessed, used, and shared. These agreements are necessary to ensure that the privacy and confidentiality of patient data are protected. It is important to note that controlled access data can encompass both unprocessed data from patients, as well as processed data for secondary use, which is structured, harmonised, normalised, and often subjected to techniques such as pseudo-anonymisation or anonymisation to protect patient privacy. In addition, researchers must also consider intellectual property rights and ownership of data, particularly in cases where the data is generated through collaborations or partnerships between multiple institutions.

To tackle legal barriers to data access, researchers must be familiar with the regulations and guidelines that apply to their specific area of research. They must obtain appropriate permissions and comply with data use agreements, as well as ensure that they are in compliance with all relevant regulations and guidelines. In addition, data owners and their institutions must work closely with legal experts and Data Protection Officers (DPO) to develop appropriate data sharing agreements and ensure that all parties are aware of their rights and obligations. For example, in the Dutch National Covid-19 Data Portal²⁵ data request is a multistep approach. The data request will undergo evaluation by a Data Access Committee (DAC) in the first instance. Upon approval, the individual university medical centre may share the data, provided that the request aligns with the requirements of their

²² Helsinki Declaration: <https://www.wma.net/wp-content/uploads/2016/11/DoH-Oct2008.pdf> [accessed 30.06.23]

²³ EUREC: <http://www.eurecnet.org> [accessed 30.06.23]

²⁴ Report on secondary use of health data through European case studies: <https://tehdas.eu/app/uploads/2022/08/tehdas-report-on-secondary-use-of-health-data-through-european-case-studies-.pdf> [accessed 30.06.23]

²⁵ Dutch National Covid-19 Data Portal: <https://www.health-ri.nl/covid-19-data-portal> [accessed 30.06.23]



local informed consent process. This process includes obtaining consent from the patient for both participation in the research and the secondary use of research data.

In accordance with the recommendations outlined in the Independent Ethics Advisor Report, the infrastructure will establish robust ethical compliance procedures. These procedures ensure that any further processing of data remains within the boundaries of the original informed consent for research participation and ethics approval, particularly in cases where the data originates from research projects. By adhering to these procedures, the infrastructure guarantees that data usage and processing adhere to the ethical guidelines and consents obtained during the research process.

It is important to note that open data, which is data that is freely available to the public, does not contain personal data and is therefore not subject to privacy laws like GDPR. However, researchers should still be aware of any licences, legal requirements or restrictions that may apply to the use and sharing of open data and should always ensure that they are in compliance with relevant regulations and guidelines. Hence, ethics are an integral part of the research across scientific domains. For example, for socio-economic research, a combination of gaining consent (of research subjects), anonymising data, defining ownership of copyright to the data as well as defining access control to the data can ensure the ethical and legal sharing of data.

5.2.1.2 Organisational barriers

Organisational barriers can also present challenges to data access in biomedical research, whether the data is open or controlled.

- The first barrier identified is establishing effective data governance frameworks that ensure that data is managed in a way that is secure, ethical, and compliant with relevant laws and regulations. This can be particularly challenging in cases where data is collected and managed by multiple institutions or stakeholders.

Improved coordination and collaboration between different stakeholders involved in data sharing and research is essential to overcome organisational barriers to data access. This includes collaboration between researchers, data owners, data custodians, legal experts, and institutional leaders. It is important to establish clear lines of communication and decision-making processes to ensure that data is shared in a way that is ethical, secure, and compliant with relevant regulations and guidelines. The stakeholders need to agree on standards and protocols to ensure interoperability. Failure to agree on these will result in technical barriers that could have been avoided.

- Additionally, but specially for open access data, the main barrier is to make open data available and discoverable. Organisational barriers may include a lack of resources and infrastructure to support data sharing and collaboration. Open data



initiatives often require significant investments in technology, data management, service management, support activities, and governance to ensure that data is managed in a way that is accessible and useful to researchers. Organisations need to commit to the investments and operational costs required for the implementation and operations. An essential component to overcome organisational barriers can be cooperation, knowledge transfer and services among organisations in accordance with Open Science principles.

To overcome organisational barriers to data access, it is important to establish effective data governance frameworks that balance the need for data security and privacy with the need for data access and sharing. In addition, stakeholders involved in data sharing and research should work together to improve coordination and communication, and to develop infrastructure and sources that support open data initiatives and controlled access data access²⁶.

5.2.1.3 Technical barriers

The SARS-CoV-2 pandemic highlighted the importance of data access and sharing in biomedical research. As the global scientific community raced to understand the virus and develop effective treatments and vaccines, access to high-quality data was essential. Sadly, the pandemic also exposed several technical barriers that hindered these essential data access and sharing efforts.

- One of the challenges is to locate and identify data that is relevant to a particular research question or area of interest. With the vast amount of health information and research data generated during the pandemic, finding and accessing the right data could be a significant challenge. Hence, the first identified technical barrier is the lack of tools and sources to facilitate data discovery. This underscores the requirement for tools and platforms that empower researchers from public and private institutions involved in COVID-19 monitoring and surveillance, who were at the forefront of generating a vast amount of health and research information during the pandemic. These tools and platforms should enable effortless data sharing, foster collaboration, and facilitate effective data analysis among these researchers.
- Another existing challenge lies in the insufficiency of high-quality metadata, especially for controlled access data, which hampers its discoverability and the ability to search and query it effectively through a public portal or endpoint.

To tackle this barrier within BY-COVID, WP2 and WP3 members are collaborating in order to maintain and improve the COVID-19 Data Portal²⁷. The platform is designed to facilitate

²⁶ European Interoperability Framework (eEIF): <https://joinup.ec.europa.eu/collection/nifo-national-interoperability-framework-observatory/3-interoperability-layers#3.4> [accessed 30.06.23]

²⁷ COVID-19 Data Portal: <https://www.covid19dataportal.org/> [accessed 30.06.23]

data discovery and collaboration across different institutions and research groups in the context of the many European projects. Additionally, WP2 is also working towards increasing awareness about these tools and standards in different domains to promote adoption of the tools and standards.

5.2.2 Data Transfer

Data transfer can involve the movement of data from one location to another, which can present a number of challenges related to data security, privacy, and compliance with relevant laws and regulations, as well as other technical challenges.

The transfer of data in biomedical research can take many forms, including the sharing of data between researchers within the same institution or research group, the transfer of data between different institutions or organisations, and the distribution of data to third-party entities such as data repositories and cloud providers. Each of these scenarios presents unique challenges and considerations that must be addressed to ensure that data is transferred faithfully, securely and in compliance with relevant laws and regulations.

Table 5: Visualisation of the data transfer procedures reported through the survey used to gather information for D2.2.

From the 26 data sources mentioned, 11 completed the survey.

<i>Representation of data transfer with standardised data transfer mechanisms</i>	
Available	
Not available	
Under development	
<i>Representation of data sources by data transfer mechanism</i>	
Web Portal	
FTP/Aspera/Globus/API	
Trusted Research Environment	
Biobank Dependant	

Data must be transferred in a way that is respectful of the privacy and confidentiality of the individuals involved, and that appropriate measures are taken to prevent the misuse or unauthorised access to the data being transferred.

In addition, ensuring data quality and consistency is crucial in biomedical research, regardless of whether the data is being transferred or not. Researchers should always take



into account data quality and consistency, tailoring their approach to meet the requirements of their research questions. However, when data is being transferred between different institutions or organisations, additional challenges may arise due to differences in data standards, formats, and quality control procedures. These differences can pose obstacles to achieving interoperability and may require special attention to ensure accurate and reliable data exchange. By addressing data quality and consistency from the outset and considering the unique aspects of data transfer, researchers can enhance the integrity and usefulness of the data for biomedical research purposes.

5.2.2.1 Legal barriers

When it comes to data transfer, one of the most important legal barriers to overcome is ensuring the security and privacy of the data being transferred. This is particularly important when dealing with controlled access data sources, where the data may contain sensitive personal information about individuals. Some of the key legal considerations for ensuring the security and privacy of transferred data include:

- **Data use agreements:** Data use agreements (DUAs) are legal contracts that govern the use and sharing of controlled access data, and are an important tool for ensuring the security and privacy of transferred data. DUAs may include provisions related to data security and privacy, and may require researchers to implement specific technical and administrative controls to protect the data. Usually, DUAs have some granularity in it defining three data access levels, free access to the data, data upon request, and data under embargo period. This classification depends on the level of sensitivity of the data that is to be decided by the data controller.
- **Use and Access Policy:** This type of policy can complement the Data Use Agreements and, as explained above, this document sets rules for data access and limits access to the database to the approved users and under specific terms and conditions.
- **Data Processing Agreement:** this document defines who is the Data Controller and who the Data Processor under art.28 GDPR²⁸, as well as their respective duties and liabilities.
- **Risk assessments:** Conducting a risk assessment of the data being transferred can help identify potential security and privacy risks, and can inform the development of appropriate controls and safeguards to mitigate those risks. This may include assessing the sensitivity of the data, identifying potential threats and vulnerabilities, and developing a risk management plan to address any identified risks.

By paying close attention to these legal considerations, researchers can help to ensure the security and privacy of transferred data, and can promote responsible and ethical data sharing practices.

²⁸ GDPR art.28: <https://gdpr-info.eu/art-28-gdpr/> [accessed 30.06.23]

5.2.2.2 Organisational barriers

Organisational barriers can present significant challenges when it comes to data transfer in biomedical research. These barriers may include issues related to data governance, institutional policies, and the need for improved coordination and collaboration between different stakeholders involved in data sharing and research. Some specific organisational barriers to consider include:

- Collaboration between different stakeholders involved in data sharing and research is essential for ensuring that data transfer processes are efficient, effective, and compliant with relevant laws and regulations. Developing clear lines of communication, establishing formal agreements and partnerships, and promoting a culture of collaboration can all help to overcome these barriers and facilitate the transfer of data between different organisations.
- Application of procedures and the documentation required for data transfer, such as ethical review and data management plan.
- Limitation in creating and maintaining tools and services. While the development of tools and services aims to promote data sharing and research, it is essential to recognise that responsible research extends beyond scientific expertise. The long-term sustainability and maintenance of these tools and services are crucial for their effective utilisation. To ensure their continued functionality and development, regular funding or formal permanent cooperation among organisations is necessary.

To tackle this last barrier mentioned, embracing Open Science principles, such as the EOSC interoperability framework²⁹, can play a pivotal role in fostering cooperative initiatives that enable sustainable maintenance and further advancements in data sharing and research. The BY-COVID project has actively worked in this direction, emphasising the importance of creating a supportive ecosystem for long-term tool and service sustainability.

5.2.2.3 Technical barriers

The main barrier for research data transfer is the lack of FAIRness, which includes challenges such as the availability of reliable data transfer protocols, especially for large volume datasets, computational capabilities, data standardisation, and the need for improved tools and infrastructure to support data sharing and collaboration. Moreover, controlled access data sources included challenges related to data security and privacy.

- One of the main challenges associated with data transfer in biomedical research is ensuring the security of the data being transferred. Data security is critical to protecting sensitive patient information, preventing unauthorised access, and ensuring that data is not lost or corrupted during transfer. This requires the use of

²⁹ EOSC Interoperability Framework (v1.0):

<https://eoscsecretariat.eu/sites/default/files/eosc-interoperability-framework-v1.0.pdf> [accessed 30.06.23]



secure transfer protocols, such as encryption of the data or encrypted connections, secure file transfer protocols (SFTP), or virtual private networks (VPN), which help to ensure that data is transferred securely and remains protected from interception or unauthorised access.

- Additionally, due to the large number of institutions, organisations, and countries involved in SARS-CoV-2 research, data interoperability was a significant challenge. Different data systems and platforms often use different data formats, structures, and terminologies, which makes it difficult to integrate and analyse data from multiple sources.

Standardisation is critical for enabling data to be shared and integrated effectively, particularly when data is generated by different sources or in different formats. However, achieving standardisation can be challenging, particularly when dealing with large, complex datasets or data generated by different types of instruments or platforms.

- As represented in Table 5, many data source providers do not frequently transfer their data and metadata of datasets based on globally accepted protocols such as REST (REpresentational State Transfer) and OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) specifically for metadata, or FTP/SFTP for data files. This happens because either they do not have the needed infrastructures and tools at all, or they have made fully customised data transfer platforms that are unable to interoperate with third party tools.
- Data repositories and cloud-based platforms can help to facilitate data reusability providing a secure location for storing and accessing data, as well as tools for data visualisation, analysis, and sharing. However, such infrastructure requires significant resources to develop and maintain in a timely manner, and may not always be available or accessible to researchers working with controlled access data sources.
- The lack of know-how and documentation (based on Open API standards and tools such as Swagger³⁰) of data transfer endpoints is a major problem as most of the content providers do not provide details of how to use their data and endpoints.

To overcome the barriers associated with research data transfer, it is crucial to adopt solutions that promote interoperability, collaboration, and investment in infrastructure. Embracing FAIR principles, such as standardised data formats and metadata descriptions¹⁷, facilitates seamless data sharing and reuse. Establishing collaborative partnerships and networks fosters cooperation and the development of common standards. Additionally, investing in secure data transfer protocols, robust repositories, and cloud-based platforms ensures efficient and accessible storage, analysis, and sharing of research data.

³⁰ Swagger website: <https://swagger.io/> [accessed 30.06.23]



By implementing these solutions, we can enhance data transfer in both open access and controlled access settings, enabling accelerated biomedical research and improved healthcare outcomes.

6. Results

With the work accomplished we were able to identify technical, legal, and organisational barriers that can hinder the sharing and reuse of data for research purposes. These barriers have been vastly discussed in section 5 (Work accomplished), so here we will provide a summary of the barriers identified in the context of BY-COVID data sources.

6.1 Identified legal barriers to data access and transfer

We identified various legal barriers, which differ depending on whether the data is open or under access control.

For open access data, the main barrier is the fact that researchers must be mindful of any licences, legal requirements, or restrictions governing its use and sharing, including limitations on commercial utilisation. Compliance with applicable regulations and guidelines is crucial, as ethics play a vital role in research across scientific domains. Additionally, sharing data across different jurisdictions entails navigating diverse legal frameworks, data protection regulations, and privacy laws, which can vary and necessitate adherence to multiple requirements. These considerations are particularly relevant when dealing with data under access control.

In the case of controlled access data, in addition to the points mentioned above, researchers must exercise extreme care to ensure patient data confidentiality, as it is protected by national laws, including transpositions of the GDPR into national regulations, European regulations such as the GDPR, and ethical guidelines. As well as the fact that restricted data access requests and cross-jurisdictional compliance further complicate accessing and transferring controlled data, adding complexities, potential delays, and increased costs.

6.2 Identified organisational barriers to data access and transfer

Organisational barriers for data access and sharing vary depending on the type of data involved.

For open access data, organisations face the challenge of committing to the necessary investments and operational costs required to make data available and discoverable. Limited resources and funding can also hinder the creation and maintenance of tools and services for data sharing.



In the case of controlled access data, organisations must go beyond open access challenges. They need to establish effective data governance frameworks to ensure secure, ethical, and compliant data management. This involves developing policies, procedures, and practices that protect privacy and confidentiality.

6.3 Identified technical barriers to data access and transfer

For open access data, one major challenge is locating and identifying relevant data amidst the vast volume of health data generated during the pandemic. It is important to note that health data is primarily generated for treating patients, and its secondary use for research purposes involves a series of operations that may not be straightforward. The process of data discovery becomes challenging due to the lack of tools and sources specifically designed to facilitate this task. Moreover, integrating and analysing data from diverse systems and platforms with different formats and terminologies further complicates data integration efforts. This is often due to the lack of interoperability between systems and the absence of necessary infrastructures and tools to transfer large datasets.

Controlled access data faces similar challenges as open access data, with additional emphasis on data security. Ensuring the security of transferred data becomes crucial, necessitating the implementation of SFTP, VPNs, or trusted research environments (TREs).

7. Discussion

In this deliverable, we present a comprehensive analysis of the primary barriers associated with accessing and transferring research data across diverse domains. These barriers have been carefully identified and grouped into three categories: legal, organisational, and technical. Rather than duplicating efforts, our approach recognises the valuable work accomplished in several prominent European projects, including European Open Science Cloud (EOSC), Towards the European Health Data Space (TEHDAS), and HealthyCloud. By building upon their achievements and other community-driven standards applicable to various types of data, this deliverable aims to leverage existing knowledge and foster collaborative solutions to address the identified barriers effectively.

To facilitate the access and transfer of research data, various legal solutions have been identified.

Harmonised data protection and privacy regulations across jurisdictions are crucial in promoting responsible and compliant data sharing practices. Initiatives like the FAIR cookbook³¹ offer useful insights into considerations related to data protection. For example,

³¹ FAIR Cookbook: <https://faircookbook.elixir-europe.org/content/home.html> [accessed 30.06.23]



the FAIR cookbook offers technical guidance to enable the generation of a Data Protection Impact Assessment (DPIA)³², as mandated by the GDPR in a machine-readable form.

TEHDAS has developed a set of policy options for the European Commission, including proposals for legislation to define the secondary use of health data and establish rules for its collection, use, and sharing³³. The aim is to ensure clarity and harmonisation in the handling of health data, including defining anonymization and pseudonymization practices and addressing national derogations to the GDPR within the European Health Data Space.

In addition, it is essential to follow the general recommendations outlined in the EOSC Interoperability Framework³⁴. This framework provides guidance on legal and organisational aspects of data sharing and interoperability, facilitating harmonised practices across domains and jurisdictions.

To enhance stakeholder understanding of legal requirements, it is crucial to develop a culture of collaboration and coordination among stakeholders involved in data sharing activities, such as data access and data transfer. Engaging the expertise of legal departments within institutions and leveraging the knowledge of Ethical, Legal, and Social Implications (ELSI) teams in funded projects can contribute to better compliance and understanding of licences, legal restrictions, and limitations on data use and sharing.

Moving into organisational barriers, to avoid long and complex negotiation processes, clear policies and procedures should be established for managing user permissions and access requests. By providing training and support to researchers is essential to ensure awareness of the ethical and legal implications of data sharing, mentioned above.

Allocation of sufficient resources, funding, and operational support is crucial for the successful implementation of open data initiatives. Organisations should prioritise the allocation of resources to support data management infrastructure, data curation, and data governance frameworks. These resources enable secure and responsible management of research data, ensuring compliance with privacy and security requirements for both, open and controlled access data.

To further guide organisations in addressing these barriers, recommendations from projects like EOSC-Life can be utilised³⁵, such as performing periodic audits of compliance to this

³² FAIR Cookbook recipes: <https://faircookbook.elixir-europe.org/content/recipes/introduction/dpia.html> [accessed 30.06.23]

³³ TEHDAS news release: <https://tehdas.eu/results/tehdas-suggests-options-to-overcome-data-barriers/> [accessed 30.06.23]

³⁴ EOSC Interoperability Framework: <https://eoscsecretariat.eu/sites/default/files/eosc-interoperability-framework-v1.0.pdf> [accessed 30.06.23]

³⁵ EOSC-Life Access and User Management System for Life Science – the implementation and usage report: <https://doi.org/10.5281/zenodo.4559400> [accessed 30.06.23]



Policy and making available the results of such audits to other Participants upon their request.

Finally, to overcome the technical solution we must guarantee the good level of FAIRness of the data starting by ensuring data interoperability and integration across different systems. In the context of BY-COVID project, two deliverables^{36,37} have been submitted with the goal of harmonising metadata from several BY-COVID data sources³⁸ in order to increase their searchability to the community.

Recently, data access attributes have been added to the metadata schema so these can be shown on the COVID-19 Data Portal. This step opens the possibility for data resources to implement Data Use Ontology (DUO) codes³⁹ or other informing consent attributes that could speed up the access process for data sharing. Also, it kicks off the possibility to implement machine-readable consent agreements⁴⁰ that would allow programmatic management of permissions that combined, for example, with GA4GH Passports and Visas⁴¹, could – possibly – completely streamline this process if desired by the data controllers.

A parallel solution is the implementation of technical solutions such as secure download and processing in place⁴². These environments can be provided by data hubs, researchers, or third parties. Alternatively, users can send algorithms to where the data is available for processing without accessing the data directly. The choice between these methods depends on factors such as data sensitivity and user requirements. These solutions align with the FAIR principles and offer researchers the ability to perform analyses while maintaining data security and compliance.

³⁶ 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>

³⁷ 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>

³⁸ List of BY-COVID data sources in FAIRsharing: <https://fairsharing.org/3773> [accessed 30.06.23]

³⁹ "The Data Use Ontology to streamline responsible access to human biomedical datasets": <https://www.sciencedirect.com/science/article/pii/S2666979X21000355> [accessed 30.06.23]

⁴⁰ GA4GH Machine-Readable Consent Guidance: https://www.ga4gh.org/wp-content/uploads/Machine-readable-Consent-Guidance_6JUL2020-1.pdf [accessed 30.06.23]

⁴¹ GA4GH Passport standard for digital identity and access permissions: <https://www.sciencedirect.com/science/article/pii/S2666979X21000379> [accessed 30.06.23]

⁴² HealthyCloud Deliverable D6.3 Specifications for data access: <https://docs.google.com/document/d/1HnaoB8jMgshfYgqqIuTFdHcn3yrOqrbTiSTdyOfhyxI/edit> [accessed 30.06.23]



As part of the EOSC Future project, a successful initiative has been undertaken to develop a cloud-optimised, community-driven file format called OME-Zarr⁴³ for bioimaging data. Alongside this format, relevant tooling and workflows, including a suitable transfer protocol (such as Aspera) have been developed. These efforts have been supported by data mobilisation and metadata curation activities within the BY-COVID project. As a result, a high-quality dataset (S-BIAD628⁴⁴) is now available for users to work and interact with.

Finally, the project has identified various platforms and data sources⁴⁵ working towards the facilitation of data sharing across different systems and making a great effort to harmonise them and increase the discoverability of SARS-CoV-2 data across countries. Moreover, proposed solutions include standards developed by the GA4GH, such as the Genomic Data Toolkit (GDT)⁴⁶, which provides guidelines for the responsible sharing of genomic data, the Beacon V2 Project⁴⁷, which enables the discovery and sharing of genetic variation data, and other beacon-like mechanisms that will be discussed in the following deliverable D2.3.

8. Conclusions

Based on the proposed solutions, the following measures are recommended for improving data access and transfer for research data, specifically in the context of the BY-COVID project.

1. Establish common data standards and protocols for effective data exchange and use across systems and platforms.
2. Develop legal frameworks and mechanisms to enable responsible data sharing across jurisdictions.
3. Establish clear policies and procedures for managing user permissions and access requests.
4. Provide training and support to researchers to enhance awareness of ethical and legal implications.
5. Allocate sufficient resources, funding, and operational support for open data implementation.
6. Foster collaboration and coordination among stakeholders involved in data sharing activities.

⁴³ OME-Zarr: <https://ngff.openmicroscopy.org/latest/> [accessed 30.06.23]

⁴⁴ S-BIAD628 Dataset: <https://www.ebi.ac.uk/biostudies/BioImages/studies/S-BIAD628> [accessed 30.06.23]

⁴⁵ List of BY-COVID data sources in FAIRsharing: <https://fairsharing.org/3773> [accessed 30.06.23]

⁴⁶ Genomic Data Toolkit:

<https://www.ga4gh.org/product/framework-for-responsible-sharing-of-genomic-and-health-related-data/> [accessed 30.06.23]

⁴⁷ Beacon Project website: <https://beacon-project.io/> [accessed 30.06.23]



9. Next steps

To connect this work with the BY-COVID use-cases and address the challenges related to sensitive clinical data, the Health Information Portal will be included in the COVID-19 Data Portal. This addition will make all the datasets within it easily accessible and discoverable. This includes the data from LINK-VACC, a Belgium registry on COVID vaccination that is built on the linkage of multiple population-based registries; BIGAN, a regional data space containing routine data linked data from EHR and population-based registries; and, HEALTH RI, a national data space aiming the linkage of any type of health data, including research data.

Additionally, all tasks from WP2 will contribute to the expansion of the data sources available on the COVID-19 Data Portal and FAIRsharing Collection. Building upon the progress achieved in task 2.4, which focused on metadata harvesting and transformation, the learnings and tools developed should be applied to other tasks, starting with task 2.3.

These next steps will ensure consistency and efficiency in data management across different aspects of the project, helping to overcome, or at least ease the process, some barriers discussed in this deliverable, such as the lack of data discoverability.

Finally, to enhance data discovery capabilities, the task 2.2 leads will initiate work on assessing the implementation of Beacons V2 or other beacon-like mechanisms. This will enable the discovery of relevant data sources and promote seamless access to the data within those sources.

10. Impact

WP2 is dedicated to the development of the infrastructure that enables easy access to a diverse range of data sources pertaining to COVID-19 and future outbreak response. The key focus of WP2 is to gather, establish connections, and standardise data sources within specific domains and disciplines. This initial effort sets the foundation for more comprehensive cross-domain harmonisation endeavours in WP3 and in conjunction with pathogen genome data in WP1.

By enhancing the FAIRness of data and metadata related to COVID-19, and other infectious diseases, WP2 aims to foster collaboration and facilitate effective utilisation of the available data resources.



Furthermore, this deliverable brings significant value to the scientific community, data controllers, data processors, and data requesters alike. In collaboration with other task forces mentioned, we aim to provide clarity on the current landscape of data sharing and offer insights into various solutions tailored to different data types based on their level of sensitivity.

11. Deviation from Description of Action

In the decision-making process, we expanded our considerations to encompass all data types, moving beyond our initial focus solely on controlled access data.

12. Annex1 - Survey

BY-COVID T2.2 / D2.2

Task 2.2 will facilitate access to sensitive bio-molecular data of COVID-19 patients and healthy humans across jurisdictions and linkage to pathogen variants via SARS-CoV-2 DataHubs (WP1). Consequently, this task will consolidate the data governance procedures to identify how they can be further streamlined. Access to data and interoperability challenges will be added by implementing community-driven standards, e.g. GA4GH DUO, Research Passwords, etc.

* Indicates required question

1. Email *
2. Which BY-COVID partner organisation are you from? *
3. Write the name of your data source *
4. URL of your data resource *
5. Select the data type available for BY-COVID *

Tick all that apply.

- Non-patient related
- Human/patient bio-molecular
- Human/patient clinical and health
- Socio-economics



6. If you have sensitive data, select the sensitive data type: *

Tick all that apply.

- Personal data
- Environmental data
- Proprietary data
- DURC data
- Classified information
- Other sensitive data
- Not Applicable

7. Could you please explain in more detail the type of data you host in your data source? (e.g. Human/patient bio-molecular data -> Our data source has a total of 800 WGS from COVID-19 patients and genotyping from 100 individuals) *

Data Access

Now we know your data source and the type of data you host, let's move to data access and the legal and technical barriers.

8. Is your data... *

Mark only one oval.

- Open Access
- Controlled Access
- Both
- Other:

9. If you answered "both" or "other...", please provide more detail. Which type of data is open access, and which is controlled access? For others, please explain the types of data access you manage. (write NA if not applicable) *

10. Is your data publicly searchable? (e.g. webpage, catalogue, public API...)*

Tick all that apply.

- Public website
- Public API
- Private website



Private API

Other

11. URL of your website/catalogue/API (write NA if not applicable) *

12. If you selected any of the *private* options in the question above, please explain the requirements to access your private services. (write NA if not applicable) *

13. If you selected the "other" option in the question above, please explain your data discovery services. (write NA if not applicable) *

14. Is your data already indexed in the COVID-19 Data Portal? *

Mark only one oval.

Yes

No

No. But in progress

15. If your answer is yes, please add the link to the COVID-19 Data Portal (write NA if not applicable) *

16. Legal barriers when sharing your data... *

Tick all that apply.

GDPR/HIPAA implementation

Specific legislations from my country

Data Anonymisation or Pseudonymisation

Patient's Consent Agreement or similar

Other:

17. If you selected the "Other..." in the question above, please provide further details on your legal barriers to data sharing (write NA if not applicable). *

18. Are you already sharing your data? *

Mark only one oval.

Yes

Yes. But only for internal use / close partners

No. But I'm planning to



No, and I'm not planning to share my data

19. How do you receive data requests? *

20. Who manages your data requests? *

Mark only one oval.

Data Access Committee

Legal / Ethical Committee

Principal Investigator of my Research Group

Other:

21. If you selected the "Other..." in the question above, please provide further details on who manages data requests for your data resource (write NA if not applicable).

22. URL of a page providing details on the access procedure and requirements (write NA if not applicable) *

23. Do you provide any associated consent information to access the data? (e.g. Data Use Ontologies). If yes, please provide the list of consent information used. (write NA if not applicable) *

24. Once users have been granted access to your data, do you have a standardised way to download the data? *

Mark only one oval.

Yes

No

Yes but still under development

25. If you answered "yes" or "under development", could you explain the download process? (write NA if not applicable) *

26. Are you experiencing any other technical issue that you would like to raise in the context of the BY-COVID Project? *

Mark only one oval.

Yes

No

27. If you replied "yes", write a paragraph on other technical issues you would like to raise.

