BY-COVID

The BeYond COVID project:
STORIES OF SUCCESS

Funded by
the European Union

Integrated data management solutions to support European pandemic preparedness

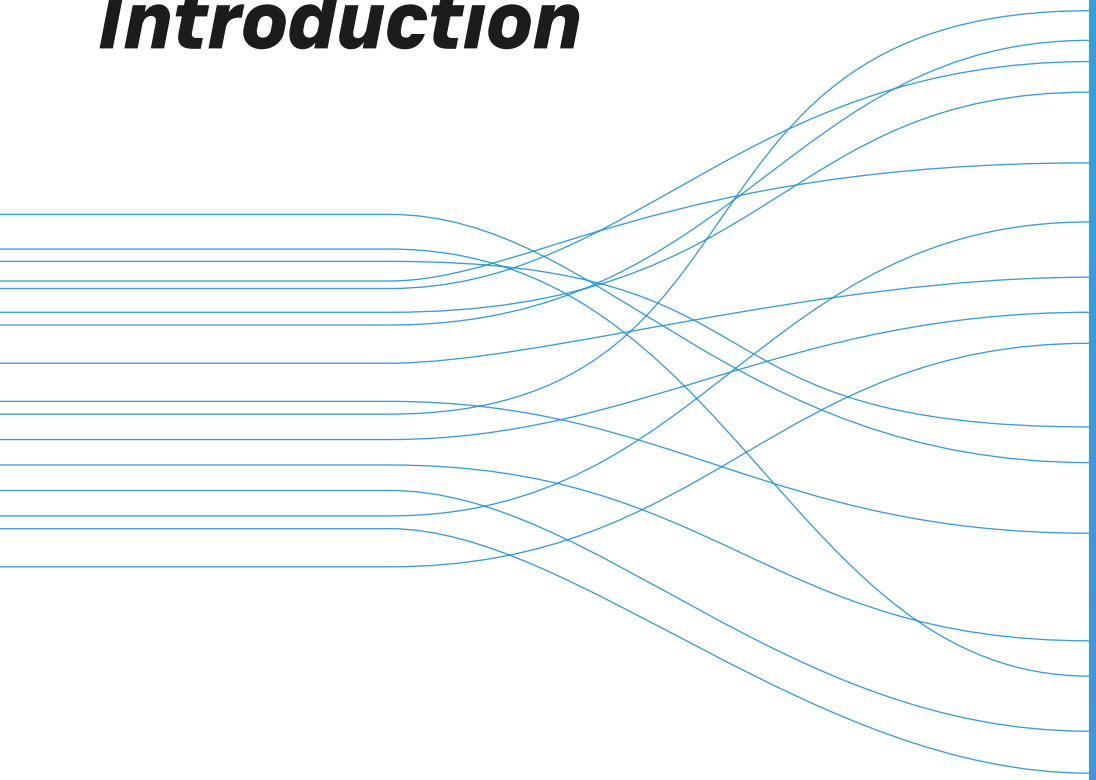
With grateful thanks to all 450 members of the BY-COVID consortium.

BY-COVID receives funding from the European Union's Horizon Europe Research and Innovation Programme under grant agreement number 101046203. Some activities have received additional funding from other sources.

September 2024
ELIXIR Hub
info@elixir-europe.org

Table of contents

Introduction



The Horizon Europe-funded BeYond-COVID (BY-COVID) project was launched in autumn 2021, as part of the European Commission's HERA incubator plan 'Anticipating together the threat of COVID-19 variants'.

The project brought together a consortium of 53 partners from 20 European countries covering clinical, public health, social, and biomolecular sciences. The aim was to consolidate solutions that had been rapidly assembled during the COVID-19 pandemic to support the continuing response to COVID-19 and preparedness for future infectious disease outbreaks.

Three years later, we present a selection of the project's many success stories, framed by the concept of a FAIR data journey. Data are first mobilised from individual centres, then connected to other data sets, standardised and made discoverable through an integrated portal, and finally made available for analysis via workflow systems.

The commitment to common, open standards and integration methodologies, such as high-level indexing across a broad range of domains, along with collaborations across scientific disciplines, are equally applicable to other major challenges requiring access to integrated data, such as food security and biodiversity.

We hope organisations outside the consortium will make use of, contribute to, and sustain this data ecosystem, in line with BY-COVID's ambition to create a data backbone.

www.by-covid.org

Mobilising data

Data Hubs: making infectious disease genomic data available to all

Challenge

Pathogen genome sequencing is a key tool in the **public health response** to disease outbreaks. However, support is required to **facilitate submission and analysis** of consensus sequences and ensure all data, including raw reads, are made **publicly available**.

Approach

Supported by various projects, including BY-COVID, the COVID-19 Data Portal and Pathogens Portal **Data Hubs** have been developed to encourage and support the submission and further analysis of pathogen sequence data by the wider research community. The tools enable users to share data quickly, both privately and publicly, and add value through rapid systematic analysis and visualisations of data. The **Contextual Data Clearinghouse** (CDCH) enables extension, correction and improvement of publicly available metadata in sample and sequence records to ensure high quality and FAIR data records.

Impact

An infrastructure has been created to **enable open sharing of pathogen genomic data**, particularly to support data submitters who may not have strong bioinformatic backgrounds or low local compute capacity. These tools and components were **repurposed in the MPox outbreak** in May 2022. For example, generating an outbreaks page and systematic analysis of MPox runs.

LEAD CONTRIBUTOR

EMBL-EBI

LINKS

DOI.ORG/10.1099/MGEN.0.001188

WWW.COVID19DATAPORTAL.ORG/DATA-HUBS

WWW.PATHOGENSPORTAL.ORG/PATHOGEN-DATA-HUBS

Recognising data stewardship

Challenge

Data stewards are experts in **making data FAIR (findable, accessible, interoperable and reusable)**, a process that is specialist and time consuming, especially for image data. However, the role of data stewardship in facilitating research is normally not recognised, making it **hard to fund staff and contributing to the profession's low visibility**.

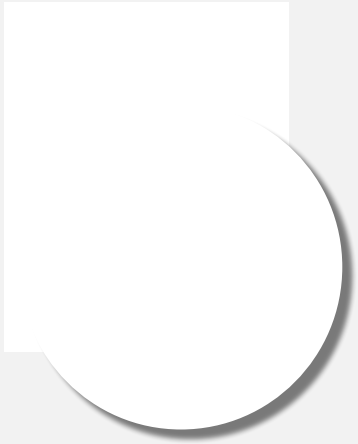
Approach

The **BioImage Archive** supports the **attribution of image datasets to collaborators and facilitators with different roles**, including that of data stewards. These contributions are linked to the researcher's ORCiDs for easy tracking.

Impact

Recognising data stewardship is a step towards sustainable funding for data steward roles, which will improve data quality in research and innovation.

Crediting data stewards shows their importance in improving data quality



LEAD CONTRIBUTOR

EURO BIOIMAGING
Image life, discover the future

LINK

WWW.EUROBIOIMAGING.EU/NEWS/RECOGNITION-FOR-THE-ROLE-OF-IMAGE-DATA-STEWARDSHIP

Connecting, standardising and exposing data

FAIRsharing: enhancing the discoverability of infectious disease data sources and their standards

Challenge

Pathogen research is accelerated by the availability of data from **clinical trials, biobanks, behavioural and socioeconomic studies, particularly if the data are combined with host and pathogen omics** information. However these data are scattered across many databases, partly or fully structured according to different standards, and tends to be **difficult to discover, access and integrate**.

Approach

FAIRsharing is a **cross-disciplinary resource that maps and interlinks** databases, standards and policies and supports their discoverability in the **European Open Science Cloud (EOSC)**. The BY-COVID FAIRsharing Collection contains over **70 data sources** including socioeconomic data, health and clinical data, images, genomic and phenotypic data and chemical biology, along with the standards used to describe the data. The Collection provides an at-a-glance view of the scope and the characteristics of each database (e.g. what type of digital objects and for which discipline; the data access restrictions), their relationships (e.g. if they exchange data), and the type of standards implemented (e.g. types of identifiers, models and terminology).

Impact

FAIRsharing has **enabled discoverability of and facilitated access to heterogeneous, yet interlinked and organised data, across domains by national data portals, building a digital space for infectious disease data**.

Over 70 data sources linked

LEAD CONTRIBUTOR

UNIVERSITY OF OXFORD

LINK

WWW.FAIRSHARING.ORG/3773

The European COVID-19 Data Platform and the Pathogens Portal: integrated data platforms to facilitate coronavirus data sharing and analysis

Challenge

In an **infectious disease outbreak**, researchers, clinicians, public health officials and policymakers need access to the **latest and most comprehensive datasets on pathogenic agents** from fields such as bioscience, social science and the biomedical sciences.

Approach

The **European COVID-19 Data Platform** provides access to open data and literature from a range of academic disciplines. The **Pathogens Portal** broadens the scope to include data on key pathogens causing diseases in humans and animals, and data on vectors and hosts. Both are augmented by a **network of Data Hubs** and **country-level Data Portals**.

Impact

As a result of this work, which is supported by many projects including BY-COVID, there is a trusted source of openly available and linked data and literature on pathogens of global importance.

Over 12 million submitted or brokered viral data sets since BY-COVID began

LEAD CONTRIBUTOR

EMBL-EBI

LINKS

WWW.COVID19DATAPORTAL.ORG

WWW.PATHOGENSPORTAL.ORG

A provenance framework for trustworthy and reproducible infectious disease research

Challenge

Unreliable or poor-quality research results have an impact on the economic, public health and political **decision making**. To ensure reproducibility, the data and specimens used in research studies must be accompanied by **standardised process documentation**, as specified in a provenance model.

Approach

A provenance model was designed to support **distributed multi-organizational workflows**: e.g. samples acquired in a hospital, data generated from in a laboratory and data processed and analysed by a university or private company. **FAIRness** of a documented object is enhanced by linking provenance information in a standard way. The model enables parts of provenance traces to be kept **confidential or anonymised** to preserve confidentiality and privacy of respective personal data subjects such as patients. A set of **RO-Crate profiles** has also been created to flexibly capture the provenance of computational workflow executions.

Impact

The model forms the foundation of the **global ISO standard series:** ISO 23494 *Provenance information model for biological specimens and data*. **Workflow Run RO-Crate has been implemented in six workflow management systems**: Galaxy, COMPSs, Streamflow, WfExS, Sapporo and Autosubmit.

The model forms the foundation of the global ISO standard series: *ISO 23494 Provenance information model for biological specimens and data*

LEAD CONTRIBUTOR

LINK

DOI.ORG/10.5281/ZENODO.10927253

Connecting, standardising and exposing data

Linking data and expertise across the social science, clinical research and biodata domains

Challenge

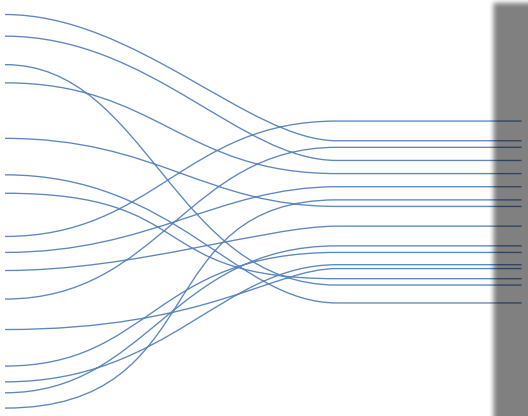
Health research is multidimensional, benefiting from linking data between biological sciences, clinical research, social sciences and public health fields. However, these data are normally **separated in silos making integrated queries challenging**.

Approach

Three approaches were taken:
- The **Pathogens Portal Cohort Browser** was developed and the linking of a multi-omics SARS-CoV-2 cohort dataset with multiple time points on participant level was demonstrated via the EMBL-EBI infrastructure.
- A flexible, **tiered metadata discovery system** was constructed, functioning across different domains, metadata standards and data maturity levels. This enables data discovery, access and analysis of disparate data through the **COVID-19 Data Portal**.
- **Metadata schemas** and related resource catalogues in **ECRIN** and **CESSDA** were compared to provide a starting point to implement 'deeper' common search mechanisms.

Impact

The Pathogens Portal Cohort Browser **brings data into context**, allowing more comprehensive and meaningful analysis. **Linked pathogen data and metadata** from omics, clinical and epidemiological research, and socioeconomic metadata is now available through the COVID-19 Data Portal.



Linked pathogen data and metadata from omics, clinical and epidemiological research, and socioeconomic metadata is now available through the COVID-19 Data Portal.

The Cohort Browser (co-funded by ReCoDID)

LEAD CONTRIBUTOR

EMBL-EBI

LINKS

WWW.PATHOGENSPORTAL.ORG/COHORTS

WWW.COVID19DATAPORTAL.ORG

DOI.ORG/10.12688/OPENRESEUROPE.16284.2

Handling sensitive data workflows

Challenge

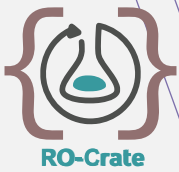
Sensitive data, such as patient data, needs to be handled carefully, for example in a **Trusted Research Environment** (TRE). **Federated analysis workflows** that are executed inside a TRE require the development of a **special profile of RO-Crate**.

Approach

Workflow Run RO-Crate has been extended to the **Five Safes RO-Crate profile**, enabling sensitive data to be used in federated workflow runs across trusted research environments, enhancing **compliance, data integrity and security**. The initial crate submits a pre-approved workflow and project details for manual and automated approval. The crate goes through multiple phases internal to the TRE, including validation, sign-off, workflow execution and disclosure control.

Impact

BY-COVID collaborations led to the spin out of **DARE UK TRE-FX** project, which will be scaled up by **HDR UK**. Five Safes Crate is also well positioned to inform the **EOSC-ENTRUST** project, which aims to create a European TRE blueprint.



The Five Safes RO-Crate profile enables sensitive data to be used in federated workflow runs across trusted research environments, enhancing compliance, data integrity and security

LEAD CONTRIBUTORS

MANCHESTER 1824 The University of Manchester

The University of Nottingham

LINK

WWW.TREFX.UK/5S-CRATE

Analysing data

A framework for federated causal research questions reusing sensitive health data

Challenge

Existing **real-world observational data**, often routinely collected, offers great opportunities for reuse particularly to **inform policy decisions**. However there are challenges in carrying out **reuse across jurisdictional borders** when **sensitive data** is involved.

We have linked real-world data from 3 countries

Approach

A methodological framework was developed suitable for attempting **causal inference** using **federated cross-national sensitive observational data**, based on principles of legal, organisational, semantic and technical interoperability. The framework includes **step-by-step guidance**, from defining a research question, to establishing a causal model, identifying and specifying data requirements in a common data model, generating synthetic data, and developing an interoperable and reproducible analytical pipeline for distributed deployment.

Impact

This comprehensively documented methodology can be **applied to any well-defined population health research question** and is generalisable to many situations, increasing the **speed of understanding infectious diseases outbreaks**.

LEAD CONTRIBUTOR

IACS Instituto Aragonés de Ciencias de la Salud

LINK

DOI.ORG/10.1186/S12874-023-02068-3

Promoting standardisation in wastewater analysis for early detection of epidemic hotspots

Challenge

Wastewater is an excellent source of samples to **monitor disease outbreaks**, however the data collected can only be **meaningfully analysed** if they are **standardised**.

Approach

Guidance on ensuring quality control in pathogen characterisation was developed and is available on the Infectious Diseases Toolkit. Strong links were built with other groups and consortia, including the EU Wastewater Integrated Surveillance for Public Health (**EU-WISH**) and the Global Consortium for Wastewater and Environmental Surveillance for Public Health (**GLOWACON**) to support **standardisation of protocols, implementation of benchmarks and harmonisation of approaches for data collection and integration**.

Impact

European and global wastewater surveillance networks are **better prepared to systematically collect, monitor and analyse wastewater to detect, quantify and track the presence of SARS-CoV2 and its variants**, and apply these methods to other threats such as **antimicrobial resistance**.

LEAD CONTRIBUTOR

LINK

WWW.INFECTIOUS-DISEASES-TOOLKIT.ORG/QUALITY-CONTROL/PATHOGEN-CHARACTERISATION

3

Secondary use of COVID-19 vaccine trial data and biosamples to understand new variants

Challenge

Vaccine trial data and patient plasma samples stored in biobanks contain a wealth of information useful for understanding new variants of SARS-CoV-2. However, **accessing and reusing this information for further research is often challenging due to ethical, legal, and institutional constraints**.

Approach

A secure **Clinical Research Data Sharing Repository** was created to enable ethically and legally compliant data sharing and reuse. The governance, data transfer and data use processes have been implemented and in collaboration with the VACCELERATE project (HTTPS://VACCELERATE.EU/) the repository is currently being piloted. A methodology was also developed **to assess the informative value of data sharing statements** in clinical trial registries, allowing for the quick identification of studies that permit data sharing for research purposes.

Impact

The new repository provides a secure, ethically and legally compliant platform for sharing clinical trial data. It facilitates transparent and efficient research through data reanalysis, meta-analysis and secondary analysis. By sharing COVID-19 vaccine trial data and biosamples, the repository supports the **rapid testing of existing vaccines against emerging SARS-CoV-2 variants**.

The clinical trial data repository supports the rapid testing of existing vaccines against emerging SARS-CoV-2 variants

LEAD CONTRIBUTOR

ecrin

LINKS

WWW.CRDSR.ECRIN.ORG

DOI.ORG/10.1186/S12874-024-02168-8

Analysing data

Using molecular data to understand disease mechanisms

Challenge

Data must not only be available but also **readily interpretable**. Molecular data are complex and difficult to analyse, so **reproducible computational pipelines are needed to project the data onto resources to support interpretation**.

Approach

Three activities were carried out using the Covid-19 Disease Map resource:

- A visualisation plugin to **map SARS-CoV-2 protein structures** to proteins in the Covid-19 Disease Map. HTTPS://GITLAB.LCSB.UNI.LU/MINERVA/PLUGINS/BY-COVID-EMPIAR
- A visualisation plugin to **map microscopy data from a drug repurposing experiment**, displaying the screened compounds and their predicted targets in the Covid19 Disease Map. HTTPS://WWW.EBI.AC.UK/BIOSTUDIES/BIOIMAGES/STUDIES/S-BIAD29
- A Galaxy workflow for visualising the **differential expression profiles of transcriptomic data** on the COVID-19 Disease Map. HTTPS://WWW.PREPRINTS.ORG/MANUSCRIPT/202403.1211/V1

Impact

Visualisation pipelines are now available to accelerate the use of omics data in biomedical research. Mapping host and viral data to the Covid-19 Disease Map provides context, supporting research on disease mechanisms of SARS-CoV-2 in a reproducible and scalable way.

Mapping host and viral data to the Covid-19 Disease Map provides context, supporting research on disease mechanisms of SARS-CoV-2

LEAD CONTRIBUTOR

uni.lu UNIVERSITÉ DU LUXEMBOURG

LINKS

GITLAB.LCSB.UNI.LU/MINERVA/PLUGINS/BY-COVID-EMPIAR

WWW.EBI.AC.UK/BIOSTUDIES/BIOIMAGES/STUDIES/S-BIAD29

WWW.PREPRINTS.ORG/MANUSCRIPT/202403.1211/V1

An automated SARS-CoV-2 genome surveillance system built around existing components

Challenge

Automated viral genome surveillance is an important element of pandemic preparedness. Many components of such a system exist, but to be used for automated surveillance they must be **linked into a single modular and scalable solution**.

Approach

A range of COVID-19 public frameworks, registries and data repositories were created or enhanced during the COVID-19 pandemic. In this work, they have been **repurposed, expanded and combined** to enable automated, reproducible, shareable, transparent, scalable and high-quality viral **sequencing data analysis** using open-source code. **Galaxy Europe** provides compute, tools and connection functionalities, **WorkflowHub** is used to ensure trustworthy and FAIR workflows. provenance is maintained with **RO Crate** and the CRG COVID-19 **Viral Beacon** provides visualisations.

Impact

A **public infrastructure** is now available for **automated pathogen analysis workflows**, including sample analysis, genome annotation, variant monitoring and the generation of data for dashboards. Genome surveillance initiatives can reuse the entire system on-premises or individual components as required. Outputs have been used in other EOSC projects: **EuroScienceGateway, EOSC4Cancer** and **TIER2**.

We have repurposed, expanded and combined existing tools to enable automated, reproducible, shareable, transparent, scalable and high-quality viral sequencing data analysis using open-source code

Galaxy

LEAD CONTRIBUTOR

LINKS

WWW.INFECTIOUS-DISEASES-TOOLKIT.ORG/SHOWCASE/COVID19-GALAXY

WWW.WORKFLOWHUB.ORG/COLLECTIONS/16

Training and sharing expertise

The Infectious Diseases Toolkit: an online resource to share expertise in infectious disease data management

Challenge:

Considerable **know-how about working with infectious disease data** was generated in centres across Europe during the pandemic but was often **scattered and not easily accessible** by others.

Approach:

The **Infectious Diseases Toolkit** is a community driven web resource that provides best practices and solutions to data challenges which affect the response to infectious diseases outbreaks.

Impact:

In the advent of an infectious diseases outbreak, there is now a **source of curated information containing guidelines, best practices, examples and national resources pages** covering pathogen characterisation, socioeconomic data, human biomolecular data, human clinical and health data.

In the advent of an infectious diseases outbreak, there is now a source of curated information containing guidelines, best practices, examples and national resources pages

LEAD CONTRIBUTORS

UNIVERSITY OF TARTU

VIB

LINK

WWW.INFECTIOUS-DISEASES-TOOLKIT.ORG

Enhancing expertise in infectious disease data infrastructure creation and management

Challenge:

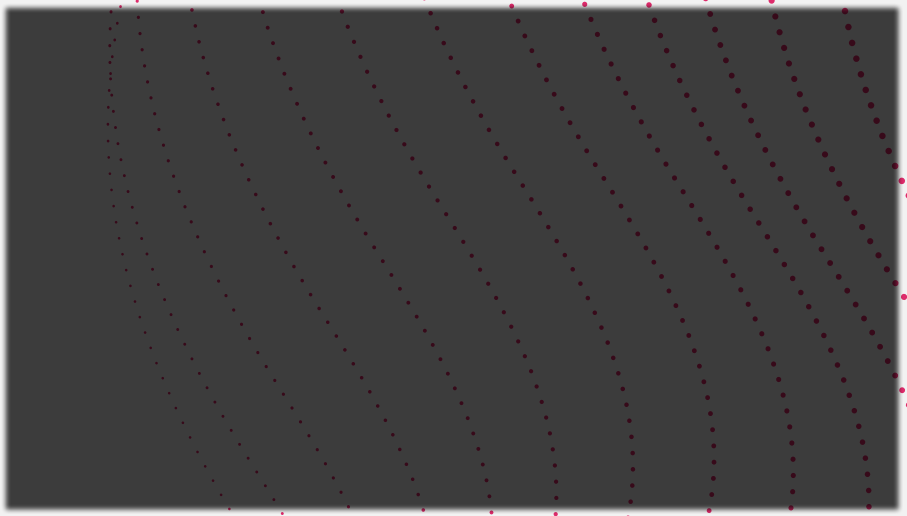
To build **sustainable capacity**, the development of infrastructure tools and guidelines must be **accompanied by awareness raising and technical training** for users and project members.

Approach:

During the BY-COVID project, **20 training related events** were held, reaching close to **2000 participants**. These included courses, workshops, hackathons and events bridging disciplines and technical areas. More than **70 training materials** were produced, made FAIR and shared.

Impact:

Researchers, database managers and developers, bioinformaticians and infrastructure specialists are **more aware of the data and resources** available for pandemic preparedness and either are **trained in their use or know where to find high quality training material**.



We made over 70 training materials openly available and FAIR

We held 20 training events reaching over 2000 people

LEAD CONTRIBUTOR

SIB Swiss Institute of Bioinformatics

LINKS

HTTP://WWW.INFECTIOUS-DISEASES-TOOLKIT.ORG/TRAINING-RESOURCES-LIST

HTTPS://BY-COVID.ORG/RESOURCES

HTTPS://FAIRSHARING.ORG/EDUCATIONAL

Engagement and outreach

Boosting awareness of research data infrastructure amongst citizens

Challenge:

Good data infrastructure is key to fighting the next pandemic, but **citizens are generally unaware of what is involved in data governance and data management**, making it harder to participate in public debate about issues of data sharing.

Approach:

The **BY-COVID Educational Toolkit** provides teachers with resources to create lessons with young people (aged 14-19+) in English, French, Spanish, Czech and Dutch to **discuss the ethics of data use in infectious disease research and pandemic preparedness**.

Impact:

Citizens have increased awareness of the components of data ecosystems and are **better informed and empowered to participate in societal debate**.

LEAD CONTRIBUTOR

sciensano

LINK

HTTP://WWW.BY-COVID.ORG/USECASES

Influencing European and international policy in infectious disease management

Challenge:

To achieve maximum impact, **BY-COVID projects tools and guidelines need to be adopted at the European and international scale**.

Approach:

Engagement with policymakers, including inviting policymakers to events, responding to consultations, meeting with intergovernmental organisations, contributions to policy reports and speaking at events.

Impact:

The BY-COVID project and its outputs have been mentioned in **25 European and international policy documents**, showing their **high visibility and influence**.

We are mentioned in 25 European and international policy documents

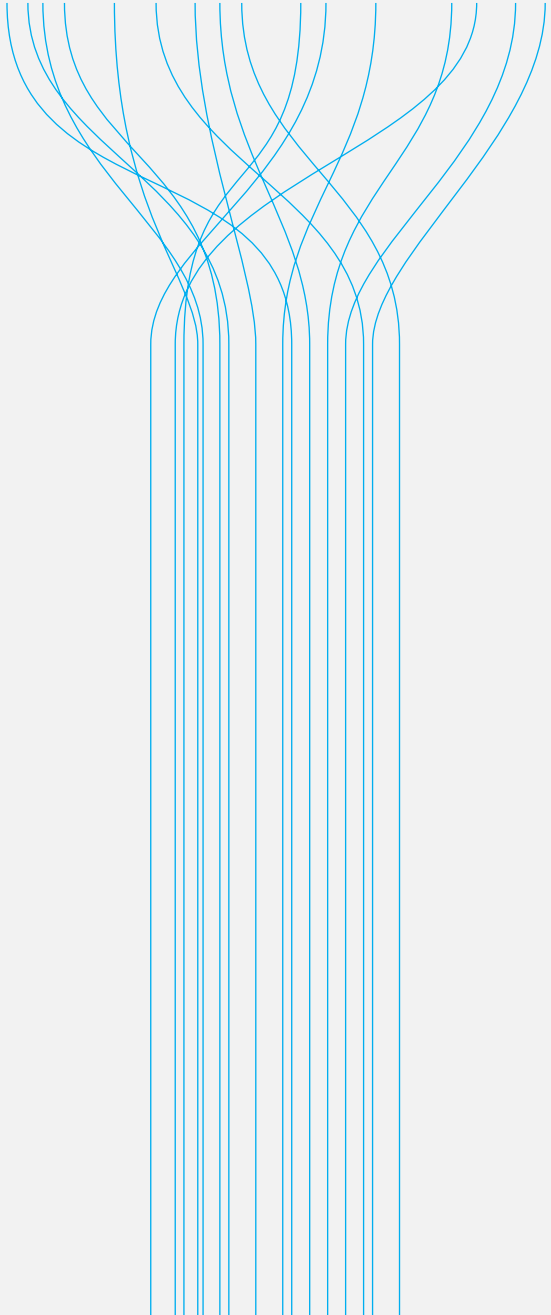
LEAD CONTRIBUTOR

elixir

LINK

HTTPS://ZENODO.ORG/RECORDS/12820912

Selected BY-COVID publications



General

1. Soiland-Reyes, Stian, et al. “Packaging research artefacts with RO-Crate.” Data Science 5.2 (2022): 97-138. HTTPS://DOI.ORG/10.3233/DS-210053
2. Soiland-Reyes, Stian, et al. “Creating lightweight FAIR digital objects with RO-Crate.” Research Ideas and Outcomes 8 (2022): e93937. P (2022) Creating. HTTPS://DOI.ORG/10.3897/RIO.8.E93937
3. David, Romain, et al. “Umbrella Data Management Plans to integrate FAIR data: lessons from the ISIDORe and BY-COVID consortia for pandemic preparedness.” CODATA Data Science Journal 22 (2023): 35. HTTPS://DOI.ORG/10.5334/DSJ2023-03
4. Ohmann, Christian, et al. “ECRIN–CESSDA strategies for cross metadata mappings in selected areas between life sciences and social sciences and humanities.” Open Research Europe 3 (2023). 10.12688/OPENRESEUROPE.16284.2
5. Soiland-Reyes, Stian, Carole Goble, and Paul Groth. “Evaluating FAIR Digital Object and Linked Data as distributed object systems.” PeerJ Computer Science 10 (2024): e1781. HTTPS://DOI.ORG/10.7717/PEERJ-CS.1781

Clinical data

1. Ohmann, Christian, et al. “Linking the ECRIN Metadata Repository with the BBMRI-ERIC Directory to connect clinical studies with related biobanks and collections.” Open Research Europe 4.50 (2024): 50. HTTPS://DOI.ORG/10.12688/OPENRESEUROPE.17131.1
2. Ohmann, Christian, et al. “An assessment of the informative value of data sharing statements in clinical trial registries.” BMC Medical Research Methodology 24.1 (2024): 61. HTTPS://DOI.ORG/10.1186/S12874-024-02168-8

Large-scale analysis

1. Meurisse, Marjan, et al. “Federated causal inference based on real-world observational data sources: application to a SARS-CoV-2 vaccine effectiveness assessment.” BMC Medical Research Methodology 23.1 (2023): 248. HTTPS://DOI.ORG/10.1186/S12874-023-02068-3
2. Rahman, Nadim, et al. “Mobilisation and analyses of publicly available SARS-CoV-2 data for pandemic responses.” Microbial genomics 10.2 (2024): 001188. HTTPS://DOI.ORG/10.1099/MGEN.0.001188
3. Pipek, Orsolya Anna, et al. “Systematic detection of co-infection and intra-host recombination in more than 2 million global SARS-CoV-2 samples.” Nature Communications 15.1 (2024): 517. HTTPS://DOI.ORG/10.1038/S41467-023-43391-Z
4. Martiny, Hannah-Marie, et al. “ARGprofiler—a pipeline for large-scale analysis of antimicrobial resistance genes and their flanking regions in metagenomic datasets.” Bioinformatics 40.3 (2024): btae086. HTTPS://DOI.ORG/10.1093/BIOINFORMATICS/BTAE086
5. Kuzikov, Maria, et al. “Drug repurposing screen to identify inhibitors of the RNA polymerase (nsp12) and helicase (nsp13) from SARS-CoV-2 replication and transcription complex.” Virus Research 343 (2024): 199356. HTTPS://DOI.ORG/10.1016/J.VIRUSRES.2024.199356

Funded by
the European Union

BY-COVID