Beyond COVID - how the BY-COVID project is increasing European pandemic preparedness

- Katharina B. Lauer. ELIXIR, katharina.lauer@elixir-europe.org, ORCID: 0000-0002-4347-7525
- Romain David, European Research Infrastructure on Highly Pathogenic Agents (ERINHA), romain.david@erinha.eu, ORCID: 0000-0003-4073-7456
- Jonathan Ewbank, European Research Infrastructure on Highly Pathogenic Agents (ERINHA), jonathan.ewbank@erinha.eu, ORCID: 0000-0002-1257-6862
- Nina Van Goethem, Sciensano, nina.vangoethem@sciensano.be, ORCID: 0000-0001-7316-6990
- Elaine Harrison, ELIXIR, elaine.harrison@elixir-europe.org, ORCID: 0000-0003-1149-2242
- Niklas Blomberg, ELIXIR, niklas.blomberg@elixir-europe.org, ORCID: 0000-0003-4155-5910
- Enrique Bernal-Delgado, IACS, ebernal.iacs@aragon.es ORCID:0000-0002-0961-3298
- Philip R Quinlan, University of Nottingham, philip.quinlan@nottingham.ac.uk ORCID: 0000-0002-3012-6646
- Patricia M. Palagi, SIB Swiss Institute of Bioinformatics, patricia.palagi@sib.swiss, ORCID: 0000-0001-9062-6303
- Susanna-Assunta Sansone, Oxford e-Research Centre, University of Oxford, UK, susanna-assunta.sansone@oerc.ox.ac.uk ORCID: 0000-0001-5306-5690
- Allyson Lister, Oxford e-Research Centre, University of Oxford, UK, allyson.lister@oerc.ox.ac.uk ORCID: 0000-0002-7702-4495
- Philippe Rocca-Serra, Oxford e-Research Centre, University of Oxford, UK, philippe.rocca-serra@oerc.ox.ac.uk, ORCID: 0000-0001-9853-5668
- Stian Soiland-Reyes, Dept. of Computer Science, The University of Manchester, UK & Informatics Institute, University of Amsterdam, NL, soiland-reyes@manchester.ac.uk, 0000-0001-9842-9718
- Nick Juty. Dept. of Computer Science, The University of Manchester, UK, nick.juty@manchester.ac.uk, ORCID: 0000-0002-2036-8350
- Frank M. Aarestrup, Technical University of Denmark, ORCID: 0000-0002-7116-2723
- Federico Zambelli, Dept. of Biosciences, University of Milan, federico.zambelli@unimi.it, ORCID: 0000-0003-3487-4331
- Petr Holub, BBMRI-ERIC, petr.holub@bbmri-eric.eu , ORCID: 0000-0002-5358-616X
- Eva Garcia-Alvarez, BBMRI-ERIC, eva.garcia-alvarez@bbmri-eric.eu , ORCID: 0000-0002-3522-5088
- Rudolf Wittner, BBMRI-ERIC, rudolf.wittner@bbmri-eric.eu, ORCID:0000-0002-0003-2024
- Maria Panagiotopoulou, European Clinical Research Infrastructure Network (ECRIN), maria.panagiotopoulou@ecrin.org, ORCID: 0000-0002-4221-7254

- Marco Antonio Tangaro, Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council, ma.tangaro@ibiom.cnr.it, ORCID: 0000-0003-3923-2266
- Philip Gribbon Fraunhofer-ITMP, philip.gribbon@itmp.fraumhofer.de. ORCID: 0000-0001-7655-2459
- David Yu Yuan, PhD, EMBL-EBI, davidyuan@ebi.ac.uk, ORCID: 0000-0003-1075-1628
- Graziano Pesole, Department of Biosciences, Biotechnology and Environment, and Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council, graziano.pesole@uniba,it, ORCID: 0000-0003-3663-0859
- Matteo Chiara, Dept. of Biosciences, University of Milan, matteo.chiara@unimi.it, ORCID: 0000-0003-3983-4961
- Dr Isabel Kemmer, Euro-BioImaging ERIC Bio-Hub, European Molecular Biology Laboratory (EMBL) Heidelberg, Meyerhofstrasse 1, 69117 Heidelberg, Germany, isabel.kemmer@eurobioimaging.eu, ORCID: 0000-0002-8799-4671
- Salvador Capella-Gutierrez, Life Sciences Department, Barcelona Supercomputing Center (BSC). salvador.capella@bsc.es, ORCID: 0000-0002-0309-604X
- Julia Lischke, Lygature, julia.lischke@lygature.org, ORCID: 0000-0002-5524-2838
- Robin Navest, Lygature, robin.navest@lygature.org, ORCID: 0000-0002-0152-2092
- Jeroen Belien, Amsterdam UMC, Vrije Universiteit Amsterdam, Department of Pathology, De Boelelaan 1117, Amsterdam, The Netherlands, jam.belien@amsterdamumc.nl, ORCID 0000-0002-7160-5942
- Michaela Th. Mayrhofer, BBMRI-ERIC, michaela.th.mayrhfer@bbmri-eric.eu ORCID: 0000-0001-6932-0473
- Markus Perola, Finnish Institute for Health and Welfare (THL), markus.perola@thl.fi, ORCID: 0000-0003-4842-1667
- Katarina Öjefors Stark, SciLifeLab Data Centre, Uppsala University katarina.ojefors.stark@scilifelab.uu.se ORCID: 0000-0001-7970-7778
- Liane Hughes, SciLifeLab Data Centre, Uppsala University, liane.hughes@scilifelab.uu.se. ORCID: 0000-0002-4784-5436
- Tom C Giles, The Digital Research Service, University of Nottingham, tom.giles@nottingham.ac.uk ORCID: 0000-0003-1356-4289
- Carole Goble, Department of Computer Science, The University of Manchester, carole.goble@manchester.ac.uk ORCID: 0000-0003-1219-2137
- Katja Moilanen, Finnish Social Science Data Archive, katja.moilanen@tuni.fi. ORCID: 0000-0002-7668-5427
- Luca Pireddu, Center for Advanced Studies, Research and Development in Sardinia (CRS4), luca.pireddu@crs4.it. ORCID: 0000-0002-4663-5613
- Simone Leo, Center for Advanced Studies, Research and Development in Sardinia (CRS4), simone.leo@crs4.it. ORCID: 0000-0001-8271-5429
- Corinne S. Martin, ELIXIR Hub, South Building, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. ORCID: 0000-0002-5428-2766
- Rafael Andrade Buono, VIB-UGent Center for Plant Systems Biology, Ghent, Belgium, rafael.buono@psb.vib-ugent.be ORCID: 0000-0002-6675-3836
- Vasso Kalaitzi, Data Archiving and Networked Services (DANS), The Netherlands, vasso.kalaitzi@dans.knaw.nl ORCID: 0000-0001-8337-120X
- Simon Saldner, Data Archiving and Networked Services (DANS), The Netherlands, simon.saldner@dans.knaw.nl. ORCID: 0000-0002-1145-7829

- Carazo, Jose-Maria, Spanish National Center for Biotechnology, Instruct Image Processing Center, CNB-CSIC, Campus Universidad Autonoma, 28049 Madrid. ORCID: 0000-0003-0788-8447
- Sorzano, Carlos Oscar S. Spanish National Center for Biotechnology, Instruct Image Processing Center, CNB-CSIC, Campus Universidad Autonoma, 28049 Madrid. ORCID: 0000-0002-9473-283X
- Aastha Mathur, Euro-BioImaging ERIC Bio-Hub, European Molecular Biology Laboratory (EMBL) Heidelberg, Meyerhofstrasse 1, 69117 Heidelberg, Germany, aastha.mathur@eurobioimaging.eu, ORCID: 0000-0001-9734-9767
- Jordi Rambla, (1) Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain. (2) Universitat Pompeu Fabra (UPF), Barcelona, Spain ORCID: 0000-0001-9091-257X
- Aina Jené, Jordi Rambla, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain
- Babita Singh, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain
- Arcadi Navarro. (1) IBE, Institute of Evolutionary Biology (UPF-CSIC), Department of Medicine and Life Sciences, Universitat Pompeu Fabra. PRBB, C. Doctor Aiguader N88, 08003 Barcelona, Spain. (2) Institució Catalana de Recerca i Estudis Avançats (ICREA) and Universitat Pompeu Fabra. Pg. Lluís Companys 23, 08010, Barcelona, Spain. (3) Center for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Av. Doctor Aiguader, N88, 08003 Barcelona, Spain. (4) BarcelonaBeta Brain Research Center, Pasqual Maragall Foundation, C. Wellington 30, 08005, Barcelona, Spain ORCID: 0000-0003-2162-8246
- Marek Ostaszewski, University of Luxembourg, ORCID: 0000-0003-1473-370X
- Francesco Messina, National Institute for Infectious Diseases “L. Spallanzani” - IRCCS, Rome, Italy, ORCID: 0000-0001-8076-7217
- Marialuisa Lavitrano, University Milano Bicocca, Milano Italy, ORCID: 0000-0003-4852-1318
- Clementina Elvezia Cocuzza, University of Milano-Bicocca, Milano, Italy, ORCID: 0000-0001-6166-1134
- Paolo Romano, IRCCS Ospedale Policlinico San Martino, Genoa, Italy, paolo.romano@hsanmartino.it, ORCID: 0000-0003-4694-3883
- Colman O’Cathail, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. ORCID: 0000-0002-0151-0657
- Philipp Gormannns, INFRAFRONTIER GmbH, Neuherberg, Germany, philipp.gormanns@infrafrontier.eu, ORCID: 0000-0001-9823-1621
- Henning Hermjakob, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK, ORCID:0000-0001-8479-0262

Introduction

Informed and rapid decision-making during public health emergencies is contingent on the availability of accurate, interoperable and timely data on a global scale. Effective collection,

analysis, reporting and transparent sharing of data and analytical workflows with and between clinicians, researchers and policymakers is therefore vital. Further, the ability to integrate diverse data sets from multiple disciplines is paramount in assisting decision makers in both in surveillance and crisis situations.

It is well established that knowledge and data that are not stored in an organised manner become inaccessible and impossible to reuse (Vines et al., 2014). Effective data sharing is key in the response to future pandemics by shaping priorities for research, preparedness plans and effective policymaking. As nations emerge from the acute phase of the COVID-19 outbreak, it is now time to reflect on the regional, national and international challenges faced in data sharing and consolidate the lessons learnt and innovative solutions into Europe's future toolkit for pandemic preparedness.

BY-COVID is a Horizon Europe-funded project bringing together a consortium covering clinical, public health, social and bio-molecular sciences[1]. It aims to address some of the key challenges in data-driven decision-making, both in the support of the continuing response to COVID-19 and in the preparedness for future infectious disease outbreaks. Extending beyond infectious diseases, many of the project outcomes and resources can be reapplied in other contexts, for example, food security and the biodiversity crisis. This is enabled by using the FAIR principles (Findable, Accessible, Interoperable, Reusable) as a basis for data and knowledge preservation, analysis and reporting methods (Wilkinson et al., 2016).

As we pass the midpoint of the project, we are taking stock and reporting on the progress, value and impact for future outbreak control, disease surveillance and pathogen data sharing (for example, in curation, sequence analysis, genotyping, phenotyping), as well as the impact of collaborations across scientific disciplines.

The challenges of data sharing and analysis in a pandemic

The unpreparedness of the world to respond to large-scale infectious disease outbreaks was exposed both during the pandemic, and by many studies during the last decade (Olliaro et Torreele, 2022; Clark et al, 2022). In particular, concerns have been raised about the paucity of efficient data sharing, limitations and bottlenecks in data sharing practices (Koopmanns et al., 2019; Drury et al., 2019), and the resulting lack in reproducibility (Aarestrup & Koopmans, 2016; Aarestrup et al., 2012; Aarestrup et al., 2020; Aarestrup et al., 2021; Begley et Ioannidis, 2015). When reviewing literature from previous infectious disease outbreaks prior to the COVID-19 pandemic, such as the Ebola virus outbreak (2013-2016),

[1] https://by-covid.org/

or the Zika virus outbreak (2015-2016), experts repeatedly highlight the same limitations and bottlenecks in data sharing practices (Koopmanns et al., 2019; Drury et al., 2019).

Barriers to more extensive data sharing include the fear of getting "scooped" or not being sufficiently credited, losing intellectual property (IP) rights, concerns around misinterpretation or misuse of data, lack of resources or expertise to prepare the data for sharing, and concerns around personal data (particularly, for clinical data). Fear that re-analysis of data might invalidate earlier results, lack of incentives for the data providers to share, and confusion around ethical, legal and societal implications (ELSI) have also been raised (Stuart et al., 2018; Devrient et al., 2021). There is also a paucity of affordable, or free-of-charge, secure infrastructures for sensitive data sharing and reuse.

Tackling these challenges necessitates long-term planning which incorporates learning from the successes and mistakes of previous responses. Time and capital must also be invested in partnerships, data-driven policy support, technical infrastructures, human skills, health literacy and enabling the long-term storage of biobank samples for data sharing. Previous initiatives addressing similar aspects include the Wellcome Trust-organised Fort Lauderdale Meeting in 2003, which discussed in the global context, how "pre-publication data release can promote the best interests of science and help to maximise the public benefit to be gained from research" (Wellcome Trust, 2003).

Technical challenges

Data-related challenges often arise already during their generation processes with fragmented collection systems, even within countries when healthcare is managed on a regional level (Knyazev et al., 2022). Healthcare workers in charge of patient sample collection, for example, are often not trained in data sciences and data management. Further, in a context such as an epidemic, their first priority lies in caring for sick individuals, leaving little time for detailed documentation and lengthy data entry and reporting, which will ultimately lead to a lack of crucial metadata, which is necessary to describe and give context to the data. The lack of information can be amplified further downstream in the data generation pipeline, for example, in a clinical sequencing laboratory with high throughput volumes and little support for IT and data-related queries. In situations where trustworthiness and quality of data are essential, this can lead to widespread problems as demonstrated during the COVID-19 pandemic. InterMine [2022] and Smith et al. [2012] report "obviously corrupt data, such as English sentences embedded in data that purported to be amino acid or nucleotide character sequences."

With no universally adopted system or standard for collecting, generating, documenting, analysing, and disseminating COVID-19 or other infectious disease data, we face fragmentation and non-compatibility of clinical and research data. Data standards are often successfully imposed in one data source, such as a clinical laboratory, but often vary across

different data sources making federated data queries a challenging task. The implementation of interoperable data standards, building on the FAIR data principles, and overseen by an organisation with global reach, as suggested by Yehudi (2022), Thorogood et al. (2021), or Wilkinson et al. (2016), may help improve infectious disease research data management and lead to a more efficient response to outbreaks as well as to increased pandemic preparedness. Unfortunately, the existence of a standard does not mean it will be widely adopted, particularly when it doesn't meet the needs of its users (Yehudi, 2022; Fairchild et al., 2018) or lacks necessary tooling to make it easily adoptable. In fact, the lack of necessary infrastructure for the adoption and implementation of standards has often led to parallel and sometimes conflicting developments.

During the COVID-19 pandemic, the rapid production of omics data contributed to the understanding of the pathophysiology of COVID-19, but also highlighted the need for comparable FAIR data sets, experimental models and metadata across domains (Montaldo et al., 2021). This is because advances in clinical research and evidence-based policy require the integration of public health data (describing population-level health and mobility, for example), with biological and clinical research, biomolecular and socio-economic data. Data need to be taken out of proprietary storage units (silos) and stored and shared in an as open/closed as necessary, connected, and FAIR environment to guarantee research advances into clinical (for example, the use of high-flow oxygen plus dexamethasone in COVID-19 treatment) and evidence-based policy use for the benefit of society (for example, the deployment of the COVID-19 vaccination programme). Some data that could be of particular interest but have been widely underused are electronic medical records, laboratory data and patient registries.

The databases of the INSDC[2] (EMBL's European Bioinformatics Institute[3], the NIG DNA Data Bank of Japan[4] and the National Library of Medicine's National Center for Biotechnology Information at NIH[5]) are examples of good practice. These resources are committed to sustainable capture, organisation, preservation and presentation of data as part of the open scientific record. Consortia such as the INSDC, which are trusted by the wider research community, can help with evidence-based sustainable solutions to prevent short-lived non-sustainable practices, federation fatigue of data generators (due to depositing data in many different databases with different requirements) and limit the effort associated with data sharing.

Publishers should connect with these consortia for data sharing and sharing of results, both positive and negative. Sharing negative results (what did not work) is a valuable contribution to a complete knowledge map, and is already supported by prestigious journals such as

[2] https://www.insdc.org
[3] https://www.ebi.ac.uk
[4] https://www.ddbj.nig.ac.jp
[5] https://www.ncbi.nlm.nih.gov

Science or Nature (Couzin 2013, Nature 2020). Sharing of negative results conveys a more comprehensive picture, fosters new research approaches and reduces resources wasted due to duplication of efforts. Other examples of efforts that have driven open knowledge sharing are curation efforts like the International Molecular Exchange Consortium[6] (IMEx), the SIGnaling Network Open Resource[7] (SIGNOR), the COVID-19 Pathways Portal[8], or more automated solutions such as Integrated Network and Dynamical Reasoning Assembler[9] (INDRA).

During the pandemic, researchers found it challenging to address important scientific questions that spanned different scientific disciplines. For instance, investigating the effectiveness of updated vaccines requires connecting studies across the social sciences, public health, clinical research and molecular biology, with the attendant challenge of integrating diverse data types. As reported by Yehudi et al (2022), "Sometimes different datasets have the potential to meaningfully answer a question present in the other dataset, but nevertheless cannot be combined. This barrier may be technical – i.e. data sources using incompatible frameworks, conflicting standards, or even non-harmonised vocabularies – or it may be a socio-legal barrier: some licence terms prevent data remixing". When analysing different processes or cell types, using different methodologies may make data incompatible despite being standardised and FAIR.

There are, however, also more technical challenges that arise due to limited standardisation across analysis workflows and software heterogeneity. Part of the solution comes from the use of open-source research software and workflows shared through publicly accessible repositories. This enables tools to be maintained, further developed, and repurposed by the wider research community. An example of this approach is the SARS-CoV-2 analysis platform hosted on Galaxy instances (USA, Europe and Australia) using public computational infrastructure and open-source software (Maier, 2021).

Publication and attribution

The fragmented, sometimes contradictory responses to SARS-CoV-2 were based on incomplete data reflecting a partial view of the complexity of the pandemic which were insufficient to inform policy decisions at local levels[10]. The need to publish COVID-19–related findings has been supported by many ethics committees, funders and journal editors, for example by 'fast-tracking' the publication of COVID-19 manuscripts. This has resulted in concerns about the quality of the reported studies even in highly ranked medical journals,

[6] http://www.imexconsortium.org
[7] https://signor.uniroma2.it
[8] https://classic.wikipathways.org/index.php/Portal:COVID-19
[9] http://www.indra.bio
[10] https://www.who.int/publications/i/item/WHO-WHE-SPP-2022.1

with some studies being retracted at a later stage. To assist in balancing quality and speed, it becomes even more necessary to access study data to perform quality checks.

An example highlighting the need of data sharing is the retraction of two major studies on the use of hydroxychloroquine and cardiovascular mortality associated with COVID-19 from the Lancet and the New England Journal of Medicine (Zdravkovic, 2020). In addition, many other data types other than clinical study data need to be made available for a coherent response to a pandemic and to combat infectious diseases, for example, socioeconomic data. Other challenges include mobilising and linking data from electronic medical records, insurance claims, clinical lab data and patient registries to perform rapid-cycle processing and analyses, and provide nowcasts to inform policy decisions. To avoid premature publications and retractions in the future we need to understand the main limitations - and set out solutions - for effective data sharing, interoperation and exploitation of data across disciplines.

In scientific research, unlike in other professional sectors, the career progression system is very much based on recognition of research efforts and outcomes by peers - a publish-or-perish system which can discourage data sharing. A significant challenge repeatedly cited as a roadblock to fast data sharing is the fear of missing out on attribution. In an ecosystem where a researcher's career and future depend on attribution, increased efforts are required to enforce and incentivise ways to formally reward and recognise these crucial contributions. Persistent identifiers, such as accession numbers and digital object identifiers (DOIs), can be assigned to a dataset or other research object and consist of a unique identifier. These can then be linked to unique personal identification services such as ORCID iD[11] and enable data generators to connect to individual contributions in a similar way to publications. An example is the European Nucleotide Archive (ENA), where studies and projects can now be claimed against an ORCID iD.

Public health decision-making

As also witnessed during the COVID-19 pandemic, effective and timely data sharing can lead to repercussions for scientists when unpopular policies such as lockdowns are based on data coming out of a particular laboratory. Here, policymakers need to provide a guarantee of protection to scientists who provide crucial pieces of information about a pathogen or health emergency.

In a context such as the COVID-19 pandemic, regulatory bottlenecks and a lack of adequate data sharing agreements can pose a threat to people’s lives and health. A balance needs to be struck between protection of sensitive data versus transparency to enable and safeguard collaborative, interdisciplinary science. Data-access bureaucracy needs to be simplified,

[11] https://orcid.org

harmonised, and standardised to allow timely and efficient action when there are healthcare emergencies (Stone Graham et al. 2016; Yehudi et al. 2022). Fortunately, the open databases and repositories supported by supranational legislation and organisations such as the European Commission, possess the required expertise in regulatory and privacy matters, and help data generators move beyond the existing dominant culture of siloed data (Kim and Zhang 2015).

Still, even when technically everything seems ready, barriers related to lack of general policies enforcing non-sensitive data deposition also come into play. An analysis of the percentage of depositions of raw cryo electron microscopy data to the Electron Microscopy Data Bank (EMDB) showed the percentage of final atomic structures supported by deposition of initial data was the same before and after the first two years of the pandemia, not being larger than 10% (and with big differences among world regions (Carazo, J.M., European Microscopy Society Yearbook 2021). Here, transparency in the governance of resources and licences that allow redistribution and enable effective exploitation of the data, along with reproducibility of analysis and results, play an important role.

Information on infectious disease outbreaks and public health threats due to pathogens needs to be rapidly available to allow for efficient interventions by policymakers and public health officials. To facilitate surveillance and research efforts on a global scale we need to build skilled human capacity and up-to-date data infrastructure. Outdated IT systems and computational infrastructure, as well as a lack of expertise in resourcing, configuring, and maintaining these, create tight technical bottlenecks (Maier et al, 2021). High-performance computers are still an underused technology in biological, public health, clinical and social science research and examples pioneering successful usage are needed to make this technology more attractive to a broader community. The mirroring of large datasets across multiple virtual locations also enables cloud-based federated analysis, so data can be shared whilst remaining in a local environment, circumventing many regulatory hurdles in data transfer.

BY-COVID's approach to increasing European readiness for public health emergencies

Introduction to BY-COVID

The COVID-19 pandemic has demonstrated the need for a globally coherent pathogen-agnostic approach to research activities, ongoing surveillance and outbreak responses. International alignment and collaboration, along with capacity building and equitable open data sharing, are key components in building a healthier future for all. Here, using established networks, services and open infrastructures is important to avoid wasting time and resources

in reinventing the wheel, especially in emergency situations where a rapid response is essential.

BeYond-COVID (BY-COVID) is a €12-million project funded by the Horizon Europe programme tackling data challenges to enable effective pathogen surveillance and outbreak control. The core aim of the project is to ensure that data on SARS-CoV-2 and other infectious diseases can be found and used by researchers, healthcare professionals, policymakers, or the general public. BY-COVID's approach is unique in its interdisciplinary integration of data from a wide range of sources.

By linking established and emerging research infrastructures and data resources from social sciences, life sciences, medical research and clinical trials, national public health institutes, health care providers and leading infectious disease research groups across Europe into a single effort. This connected approach is demonstrated by the COVID-19 Data Platform[12], which integrates efforts spanning technical and regulatory harmonisation of data sources, technology for streamlined data processing and sharing, and use cases demonstrating real-life applications. Further, BY-COVID is providing ways forward for fruitful interactions between the research communities involved in the study of a pandemic, the future European Health Space for secondary use (HealthData@EU) and the EOSC ecosystem.

On a technical level, BY-COVID creates a flexible and interlinked core of FAIR data and analysis pipelines. As such, provides answers to the constantly evolving scientific questions arising during a pandemic and beyond the immediate emergency response. The commitment to common, open standards and integration methodologies, such as high-level indexing of COVID-19-related knowledge across a broad range of domains, allows organisations outside the consortium to make use, contribute, and sustain this ecosystem as the data backbone of Europe's pandemic preparedness and response. In summary, the technical core activities of the project are data mobilisation, storage, harmonisation, discovery, analysis, integration, and reuse, thus enabling innovation, research, public health and policy activities.

BY-COVID's technical solutions

BY-COVID is building a portfolio of robust and adaptable technical resources to support data-driven decision making in public health emergencies. This includes web portals, tools and workflows. The flagship offering is the COVID-19 Data Platform, which acts as an entry point for the integration of data from disparate sources. This is complemented by the development of a knowledge and experience exchange platform, the Infectious Disease Toolkit, which offers solutions and expertise to enhance pandemic responses.

[12] https://www.covid19dataportal.org

COVID-19 Data Platform

BY-COVID provides the COVID-19 Data Platform, a single point-of-entry web portal for a diverse range of COVID-19-related data, thus addressing the issue of dispersed and disconnected data resources. With its flexible, tiered indexing system that connects data across disciplines, the COVID-19 Data Platform sits at a pivotal position in the data journey – linking the upstream mobilisation efforts with downstream reuse and analysis (Figure 1).

The COVID-19 Data Platform offers its global user base open access to comprehensive SARS-CoV-2 and COVID-19 data, enables researchers to upload, access and analyse COVID-19-related reference data and datasets, and allows the wider community to develop new services and tools that add value to the original data. It is further used as the entry point for accessing data from disciplines beyond molecular biosciences, such as social sciences and public health, and is linked to national data portals that showcase and highlight COVID-19 research data from each of the participating countries. The Data Platform is built on open science and FAIR data principles, and draws on decades of experience in providing public infrastructure by BY-COVID project partners.

The Platform comprises the following technical components:

- COVID-19 Data Portal - the web and programmatic entry point into a wealth of integrated Covid-19 data and services
- SARS-CoV-2 Data Hubs - a tool for the analysis and sharing of viral sequence data
- Federated European Genome-phenome Archive (FEGA)- a federated data management system for sensitive patient-related data
- Integration mechanisms for other partner resources - for example, linking to other thematic portals or diverse data

The lessons learned from the implementation of the COVID-19 Data Portal, and the technology developed, underpin the newly launched, and more general, Pathogens Portal[13].

[13] https://www.pathogensportal.org

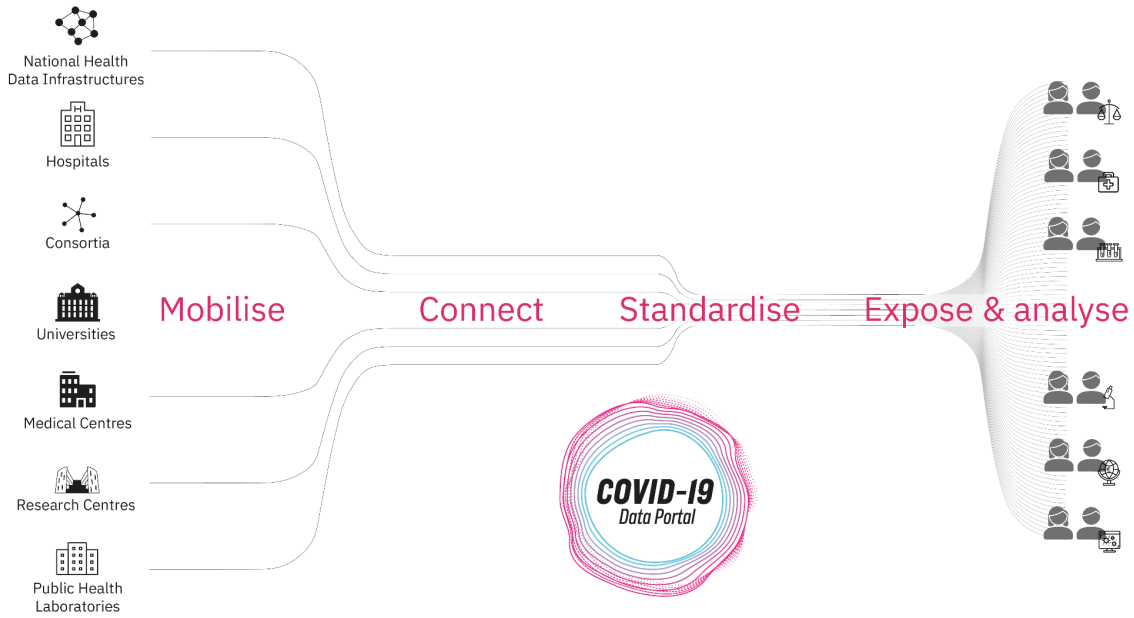

Figure 1: The BY-COVID "Data Journey" concept.

Cross-disciplinary integration

Recognising the importance of cross-discipline connections, BY-COVID set out to broaden the original biomolecular focus of the Platform. One early success has been the incorporation of socioeconomic data via the integration of the Consortium of European Social Science Data Archives (CESSDA[14]) Data Catalogue which includes around 40,000 studies. Similarly, clinical observations stored in various electronic data capturing systems, such as the WHO/ISARIC Case Report Forms,[15] are being harmonised and made available at different access levels (public dashboard, restricted access for discovery or request application workflows). Federated tooling relating to individuals and sensitive in nature from Health Data Research UK (Jefferson et al., 2022) that works across Trusted Research Environments and BBMRI-ERIC have been explored and open-source tooling developed.

BY-COVID is connecting infrastructures, and pioneering improved data sharing, by bridging the social sciences and humanities, with clinical, population health, biomedical, and biomolecular sciences. Access to these diverse datasets will enable broader opportunities for analysis, integration and reuse of data. A particular attention is given to quality and standardisation of data, metadata and provenance, as this often represents a bottleneck for effective data exploitation. Initiatives such as Clinical Quality Language (CQL[16]) for

[14] https://www.cessda.eu

[15] https://isaric.org/research/covid-19-clinical-research-resources/covid-19-crf

[16] https://cql.hl7.org

healthcare data exchange standards and HL7 Fast Healthcare Interoperability Resources (FHIR)[17] are used for assessing quality and defining quality assurance levels of the resources linked to the COVID-19 Data Portal.

An example of the type of data that BY-COVID seeks to mobilise is that generated in the framework of ISIDORe (Integrated Services for Infectious Disease Outbreak Research[18]). ISIDORe provides research services and resources to study epidemic-prone pathogens and is the largest and most diverse research and service provider for infectious diseases in Europe, with expertise from structural biology to clinical trials (Richard, Stepanyan & Ewbank, 2021). Both BY-COVID andISIDORe received funding from the European Union's Horizon Europe research and innovation programme, and the two projects were mandated to work together. Thus, one ambition of the BY-COVID project is to provide the resources to manage the results and data generated by ISIDORe partners.

Researchers using ISIDORe resources are given guidance in FAIRification, by experts working within the BY-COVID consortium, as well as practical support to ensure that datasets are linked to the appropriate open repositories, in line with ISIDORe's data management plan (David et al. 2023b). The provision of an ORCID iD is a mandatory element in the application process for ISIDORe services, allowing datasets and authors to be linked automatically. In cases where the recommended procedure is not followed, if researchers fulfil their contractual obligation and acknowledge ISIDORe in any publication, cross-referencing ORCIDs, funding statements and dataset DOIs and/or accession numbers from publications would provide an alternative, if more laborious, method to connect authors to datasets.

Sensitive data and indexing

The data types in BY-COVID cover a range of levels of sensitivity, as shown in Table 1.

Data Type	**Sensitivity**	**Examples**
Aggregated or ecological data	**Not sensitive** Can be openly shared and centrally located	Contextual information at population-level on health and health determinants (life styles, environment, services, socioeconomic features)
Non-identifying biomolecular data	**Not sensitive** Can be openly shared and centrally located	Viral sequence and variation Animal host response gene expression De-identified human metabolomics Reference protein structures

[17] https://www.hl7.org/fhir

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

		Chemical screening data Animal models Bioimaging of cells and tissues
Non-identifying non-biomolecular patient data without dedicated data resources	**Not sensitive following de-identification** Can be openly shared and centrally located	Seroprevalence studies Deep serology data exposure
Population-level data from cohorts or surveys	**Potentially sensitive**	Epidemiological, clinical, contextual data at individual level
Potentially identifying patient-linked data	**Sensitive** Typically held within national jurisdictions, and needs de-identification to clinical features (Ohmann et al., 2017)	Patient genotype Behavioural data (including vaccine hesitancy)

Table 1: BY-COVID data types and sensitivity level

Analysis across different national borders, legal jurisdictions and data types (including sensitive data and linkage to pathogen variants) is a known bottleneck. The Federated EGA model addresses these challenges directly, and expansion of the resource with further data sources is planned. Access to personally identifiable host data is governed by local, national, and European regulations, and participant consent agreements, and work is required to consolidate the data governance procedures to streamline processes. Interoperability challenges are addressed by implementing community-driven standards, for example Global Alliance for Genomics and Health (GA4GH) Data Use Ontology[19] and Passports[20].

For data under access control, harmonisation is achieved at the metadata level, but for publicly available data sources, harmonisation can include the data itself, allowing a more expansive search across different repositories. In BY-COVID, a flexible, tiered metadata discovery system across different domains, metadata standards, and maturity/robustness levels of data sources has been developed. Tier one provides the deepest indexing available, capturing granular record-level identifiers, metadata and attributes. Tier two provides more coarse-grained metadata and attributes, typically at the level of a dataset or study, rather than at the record level. Tier three indexes only resource-level metadata based on a curated FAIRsharing Collection (Hermjakob et al., 2022).

[19] https://github.com/EBISPOT/DUO

[20] https://github.com/ga4gh-duri/ga4gh-duri.github.io/blob/master/researcher_ids/ga4gh_passport_v1.md

Workflows and tools

The COVID-19 Data Platform ecosystem includes a broad and growing portfolio of existing and newly developed analysis tools and workflows, enabling researchers to explore innovative analyses that can be used to integrate and visualise COVID-19 related data. The Data Platform includes tools for real-time monitoring of surveillance data and interactive analyses, for example, using the Versatile Emerging infectious disease Observatory's Jupyter notebooks[21]. Data discovery is further enabled by the BY-COVID data discovery network, based on the principles of the Beacon Network, an API that allows for data discovery.

Within the BY-COVID project, a lightweight provenance model addresses traceability of ownership and location of data and its precursors, tailored for infectious diseases. This is based on the Common Provenance Model (CPM)[22], a provenance model developed in the EOSC-Life project[23] that is designed for complex, multi-organisational environments (Wittner 2022), and that covers the whole data life cycle from sample acquisition, data generation, and data processing. Additionally, the CPM forms an open conceptual foundation for the ISO 23494 provenance standard series development (Wittner 2023) and aims to be adopted by both academia and industry. The data, provenance, and workflows are packaged using RO-Crate (see below), building on FAIR principles and using established open standards. Specific identifiers allow the provenance of data to be interlinked with the source of biological samples, and provenance traces covering the cross-institutional journey of samples and their derived data can be kept confidential or anonymised.

FAIRsharing

FAIRsharing is a cross-disciplinary resource interlinking standards (for terminologies, formats, reporting guidelines and identifier schemas), databases (knowledge bases and repositories) and policies (Sansone et al., 2019). It is used as the BY-COVID data source catalogue[24] and provides detailed descriptions of the data sources and their relationships, including data types and their mechanisms of access. FAIRsharing also: (i) assists with the selection, documentation and visualisation (as a relation graph) of the standards implemented by the data sources, (ii) maintains a curated record of the data sources indexed in the COVID-19 Data Portal, and (iii) provides additional information on the BY-COVID data sources in the European Open Science Cloud (EOSC) ecosystem.

[21] https://www.veo-europe.eu/
[22] www.commonprovenancemodel.org
[23] https://www.eosc-life.eu/
[24] https://fairsharing.org/3773

RO-Crate

Within BY-COVID, efficient sharing of digital objects (for example, data, analysis workflows) is a managed through RO-Crate[25] [Soiland-Reyes 2022]. It implements the principles of FAIR Digital Objects (De Smedt 2020; Soiland-Reyes 2022b), by combining data and metadata in a single package, as well as through the application of community standards such as schema.org and Bioschemas (Gray 2017). The generic design of RO-Crate allows existing metadata profiles to be adopted within BY-COVID where appropriate, while new profiles can be readily created as required. Indeed, it is precisely through conformance to one or more profiles that RO-Crates become fully typed, machine-actionable digital objects that enable interoperability throughout community ecosystems [Soiland-Reyes 2022b]. Moreover, to ensure the findability and discoverability of data and research outputs, persistent and unique identifiers (for example, digital object identifiers) are used for digital assets referenced within BY-COVID. For example, WorkflowHub.eu is used as a workflow repository (Goble 2021, Goble et al., 2022), with standardised workflow identifiers and metadata descriptions using RO-Crate[26]. Other digital objects are deposited in the Zenodo open research repository as a citable research output (e.g. Mendez-Villalon 2022), which, in the case of RO-Crate, include web rendering using GitHub Pages (e.g. Meurisse 2023a).

BY-COVID is also addressing the provenance of data analysis results through the development of RO-Crate profiles to capture workflow execution history. Established in January 2020, the Workflow Run RO-Crate working group[27] brings together BY-COVID partners, workflow engine developers and the RO-Create community in an effort to build a common, interoperable format for the exchange of machine-actionable information related to the execution of scientific data analysis applications. To accommodate the variety of applications that fall under the broad definition of workflow, several profiles are being defined that describe provenance at increasing granularity levels, from “black box” (where only workflow-level inputs, outputs and parameters are considered) to step-by-step. Implementations of the format are either underway or planned for several workflow engines, including Galaxy, CWL, Snakemake, StreamFlow, Sapporo WES, COMPSs, and WfExS (Leo 2023). An RO-Crate profile that integrates RO-Crate and the CPM has been developed (https://w3id.org/cpm/ro-crate). RO-Crate is also capable of describing interactive notebook computations without formal workflow systems, such as for federated causal inference models to assess vaccine effectiveness (Meurisse 2023b).

[25] https://w3id.org/ro/crate

[26] https://workflowhub.eu/workflows/110?version=7

[27] https://www.researchobject.org/workflow-run-crate

Infectious Disease Toolkit

The Infectious Diseases Toolkit (IDTk) is being developed as part of the BY-COVID project to guide researchers and other stakeholders in the response to infectious diseases outbreaks[28]. The IDTk will serve as a knowledge and experience exchange platform to showcase past and present efforts and solutions towards pandemic responses. From these collected experiences, guidelines and best practices will be distilled and presented in an open and accessible way. Although COVID-19 has been its initial focus, IDTk will offer a platform for exposing, in a contextualised manner, the existence of broader national resources and developments in infectious diseases.

The IDTk will enable researchers to locate tools and guidelines related to the access, analysis and sharing of infectious disease data. The aim of the kit is to inform researchers of best practices for a given type of data, and to facilitate a faster response to disease outbreaks. The IDTk further links and signposts other knowledge resources, including the RDMkit and the FAIR Cookbook[29] (Rocca-Serra et al. 2022). The latter guides users through the key steps of a FAIRification journey via recipes, providing levels and indicators of FAIRness, the maturity model, the technologies, the tools and the standards available, as well as the skills required, and the potential challenges to be faced, to achieve and improve FAIRness.

The knowledge exchange platform is being developed in an open and collaborative way. By design, the IDTk leverages and expands existing technologies. It follows the successful ways of working of the recently launched RDMkit[30], a Research Data Management Toolkit for best practices and guidelines to support FAIR policies in data management. It uses the same technical infrastructure[31] as RDMkit for best practices and guidelines to support FAIR policies in data management. As such, it connects with a variety of established services such as registries for training materials (ELIXIR Training eSupport System, TeSS[32]), for standards and databases (FAIRsharing) and computational tools (ELIXIR bio.tools). The toolkit also integrates with knowledge models where users can get support for generating data management plans via the Data Stewardship Wizard[33].

Applied examples from BY-COVID

Addressing these policy-driven questions often requires data from multiple domains, sources, data owners and countries. The use cases in the BY-COVID project are designed to address this challenge, particularly in a rapid response situation. Specifically, how to

[28] https://www.infectious-diseases-toolkit.org
[29] https://faircookbook.elixir-europe.org
[30] https://rdmkit.elixir-europe.org
[31] https://elixir-belgium.github.io/elixir-toolkit-theme
[32] https://www.tess.elixir-europe.org
[33] https://ds-wizard.org

understand the real-life effect of policy measures, how to improve the timely reporting of surveillance measures, how to repurpose clinical trials results, and how to integrate knowledge on disease pathways at a molecular level to understand mechanisms behind new diseases and their long-term consequences. Such experience could be applied in sharing of viral genomic data of SARS-CoV-2, along with clinical and epidemiological data, collected in a specific surveillance activity, such as on high-risk patient categories (i.e. immunosuppressed) or during specific anti-COVID-19 treatments.

Understanding real-world effects of policy measures

The BY-COVID use cases provide workflows that are generalizable to many situations. For example, to understand the effect of a policy measure (for example, the effectiveness of vaccination in reducing transmission), a workflow is available to link individual sensitive data from multiple heterogeneous data sources (for example, clinical, administrative and socio-economic data) into an observational study which emulates a randomised clinical trial (Meurisse 2023b).

This workflow provides a structured process for causal inference resulting in analytic approaches avoiding apparent paradoxes and common biases (Hernàn et al. 2016). The framework entails several steps, including the framing of the research question into a causal model by using a Directed Acyclic Graph (DAG), the subsequent translation into data requirements, the implementation of a common semantically interoperable data model (Estupiñán-Romero 2023), a technologically interoperable analytical pipeline[34] and the publication of the results following open science principles.

Beyond the actual implementation of the workflow, this use case explores the real-life challenges in the combination of data from multiple countries, the linkage of sensitive data from various data sources, and the distribution of portable computational data analysis solutions based on software containers in the premises of the data holders (González-García et al. 2021). Notably, any workflow which addresses the development of observational studies on human populations risks can be prototyped in this way. Research community reuse is further promoted by publishing and sharing the digital objects in open science initiatives such as Zenodo, and by packaging the workflow as an RO-Crate[35].

Sharing and repurposing clinical data

When it comes to clinical research, data sharing is increasingly regarded as a key requirement for accelerating scientific discoveries. When the research community has access to Individual Participant Data (IPD) that underlie research results, new analyses can be done by other researchers with different ideas and expertise, and data can be pooled for

[34] https://github.com/by-covid/BY-COVID_WP5_T5.2_baseline-use-case
[35] https://by-covid.github.io/BY-COVID_WP5_T5.2_baseline-use-case

meta-analysis to increase statistical power. Sharing and reusing IPD makes the research results more transparent and trusted. The value of data sharing lies in answering new research questions that could not be addressed by individual datasets or by the primary researchers alone. Over the last decade, a wide variety of stakeholders, including funders and publishers, have been pushing for the cultural shift needed to make clinical trial data sharing a reality (Ohmann C, Moher D, Siebert M, et al, 2021; Merson et al., 2022). In 2016, the International Committee of Medical Journal Editors (ICMJE), a small group of medical journal editors, published an editorial stating that 'it is an ethical obligation to responsibly share data generated by interventional clinical trials because participants have put themselves at risk' (Taichman DB, Backus J, Baethge C, et al., 2016).

Naturally, most efforts around clinical trial data sharing are focused on the data sets themselves, but data sharing is much broader. Besides the IPD, other clinical trial data sources should be made available for sharing (for example, research protocols, clinical study reports, statistical analysis plans and blank consent forms) to enable a full understanding of any dataset. A recent study examined the intention to share COVID-19 clinical study data as declared by the investigators during registration of their study in ClinicalTrials.gov and found that despite the call for action, this remains at about 15% (Larson et al., 2022).

To address this challenge, the European Clinical Research Infrastructure Network (ECRIN) has partnered with the University of Oslo to design, develop, implement and operate a file-based repository for IPD from COVID-19 clinical research studies that is compliant with European regulations and in particular with the GDPR. For the storage of the IPD, the services for sensitive data (TSD) infrastructure of the University of Oslo will be used, which is a multi-tenant remote access system with a strong set of built-in security measures (Canaham et al., 2020; Ohmann et al., 2021). Within the BY-COVID project, this service supports researchers through the initial steps of their studies (such as the development of appropriate consent forms, and the use of common data standards such as CDISC) and is extending its scope to cover other infectious diseases beyond COVID-19.

A methodology for the identification of data providers is currently being established, using the ECRIN metadata repository[36]. ECRIN has also planned a study to assess the willingness of researchers to share COVID-19 clinical research IPD (Canham et al., 2022) by evaluating the data sharing statements (DSSs) in a wide range of clinical trial registries, which include ClinicalTrials.gov, EUCTR[37], ISRCTN[38] and WHO ICTRP[39]. This will lead to recommendations on improving the completeness of the DSS fields of the different registries. This type of action

[36] https://crmdr.org

[37] www.clinicaltrialsregister.eu

[38] www.isrctn.com

[39] www.who.int/clinical-trials-registry-platform

is integral to the success of future data sharing efforts, and more broadly to the interoperability of disparate data resources accessible via the COVID-19 Data Portal.

Investigating molecular mechanisms of COVID-19 pathology

One of the responses to the pandemic was a global effort to understand molecular mechanisms of SARS-CoV 2 infection, including, for example, studies of the interactions between the human and viral proteomes (Gordon et al 2020 Science), or the transcriptomic profiles of infected lung cells (Lukassen et al 2020 EMBO J). In parallel, the search began for drugs that could be repurposed to treat COVID-19, supported by systematic, high-throughput bioimaging studies (Ellinger et al 2021 Sci Data). These and many other investigations generated an astounding number of scientific reports, with over 700,000 full-text articles reported in the CORD-19 dataset[40]. However, this deluge of information must be filtered and synthesised to understand how particular mechanisms combine to create the complex picture of COVID-19 pathology. In response to this challenge, the scientific community has engaged in a large-scale biocuration effort, building a number of knowledge repositories, from curated datasets of molecular interactions (Perfetto et al 2020 Database) to molecular pathway diagrams (Ostaszewski et al 2021 Mol Syst Biol). These knowledge bases can then be used for data visualisation, interpretation and generation of new hypotheses (Niarakis 2022 bioRxiv doi: 10.1101/2022.12.17.520865). What remains to be addressed is a streamlined workflow to project newly generated data from individual studies onto such repositories, accelerating research on diagnostics and treatment.

With the COVID-19 Disease Map repository, BY-COVID builds capacity for integration of molecular and bioimaging data[41]. The Map is a large-scale community effort to encode and visually display molecular interactions between SARS-CoV-2 and human cellular pathways for visual exploration and computational analysis. In this way, it helps reach BY-COVID's goal to mobilise sensitive patient data for secure processing in a reproducible way with the support of the Galaxy ecosystem. Such processed data can be then visualised to explore gene or protein expression, pinpoint drug targets and relate bioimaging data to the corresponding molecular mechanisms.

Conclusion and looking to the future

The pandemic has shown the clear need for a transparent, globally coherent approach to combat infectious diseases. To achieve this, a transparent framework that relies on multiple coordinated resources has to be fostered. This approach circumvents the risk of one country or organisation monopolising access to critical resources and guarantees a more equitable landscape of human, animal and environmental health data. A global network of coordinated

[40] https://github.com/allenai/cord19
[41] https://fairdomhub.org/projects/190

resources transparently governed by an international collaboration or body will also support development of global capacity and resources.

As described above, BY-COVID will contribute to many aspects of the overall ecosystem of pandemic preparedness. The anticipated long-term impacts of the project include increased federation of pathogen and human infectious disease data and improved reusability of analysis methods from national and international centres (including biobanks), lending to the acceleration of research progress.

Antimicrobial resistance (AMR) in microorganisms poses another potential threat to public health. Here, the learnings and resources from BY-COVID could assist in making genomic and phenotypic data rapidly available to increase understanding of resistance mechanisms and support public health decision-making. The global spread of antimicrobial resistance genes (ARG) is a result of selective pressure exerted by antimicrobials released from anthropic, zoonotic sources and, more generally, their presence in the environment (Ko et al., 2022; Matuku et al 2022). The standardised sharing of human bacterial pathogen AMR data, integrated with clinician, diagnostic and multi-omics data would strengthen the current antimicrobial resistance surveillance systems network (for example, GLASS[42]).

Social sciences research provides important, yet often under-utilised, insights into public health emergencies, notably by improving surveillance (for example, which social groups are most impacted), prediction (for example, of economic consequences), and intervention (for example, vaccine uptake in different social groups) (Lohse & Canali 2021). Integrating social sciences into the COVID-19 Data Portal provides the possibility to analyse the consequences of the pandemic not only to individuals but also to societies. By connecting socioeconomic data to health science data, it is possible to investigate, for example, the effects of the lockdown on the spread of disease and on the socioeconomic well-being of individuals and societies.

More broadly, BY-COVID will contribute to attaining UN Sustainable Development Goal three on *Good Health and Wellbeing*, with national governments better prepared to tackle future infectious disease outbreaks. Drawing on the UNESCO Recommendations on Open Science, BY-COVID contributes to improving trust in science through increased FAIRness, openness, and quality of scientific research. This result is further supported by more meaningful monitoring and better facilitation of reproducibility, validation and reuse of research results, and by improving pathways for the communication of science – for example, to the public. Overall, this will contribute to attaining UN Sustainable Development Goal nine on *Industry, Innovation and Infrastructure*, by stimulating development and uptake of a wide range of new innovative and value-added services from public and commercial providers.

[42] http://www.who.int/glass/en

BY-COVID is paving the way for transforming how researchers, and other relevant stakeholders in the public and private sectors, create, share, and exploit research outputs (for example, data, publications, protocols, methodologies and software) within and across research disciplines, and with the public health sector. These changes will lead to improved timeliness, better quality data, more innovation, higher productivity of research and a better integration between research outputs and public health policy.

Here, the project contributes to implementing Organisation for Economic Co-operation and Development (OECD) Council *Recommendations on Access to Research Data from Public Funding.* BY-COVID shows that investments in standards, tools and infrastructures, as well as building competences for data management, will help society to get the most out of data-driven innovation and will prepare the world to respond to future healthcare crises in a more effective way. This has already been shown when resources from the project were repurposed for the 2022 Mpox Virus outbreak[43], and will further be developed to fit any infectious disease outbreak.

References

- Aarestrup FM, Brown EW, Detter C, Gerner-Smidt P, Gilmour MW, Harmsen D, Hendriksen RS, Hewson R, Heymann DL, Johansson K, Ijaz K, Keim PS, Koopmans M, Kroneman A, Lo Fo Wong D, Lund O, Palm D, Sawanpanyalert P, Sobel J, Schlundt J. Integrating genome-based informatics to modernize global disease monitoring, information sharing, and response. Emerg Infect Dis. 2012 Nov;18(11):e1. doi: 10.3201/eid/1811.120453.
- Aarestrup FM, Bonten M, Koopmans M. Pandemics- One Health preparedness for the next. Lancet Reg Health Eur. 2021 Oct;9:100210. doi: 10.1016/j.lanepe.2021.100210.
- Aarestrup FM, Koopmans MG. Sharing Data for Global Infectious Disease Surveillance and Outbreak Detection. Trends Microbiol. 2016 Apr;24(4):241-245. doi: 10.1016/j.tim.2016.01.009.
- Aarestrup FM, Albeyatti A, Armitage WJ, Auffray C, Augello L, Balling R, Benhabiles N, Bertolini G, Bjaalie JG, Black M, Blomberg N, Bogaert P, Bubak M, Claerhout B, Clarke L, De Meulder B, D'Errico G, Di Meglio A, Forgo N, Gans-Combe C, Gray AE, Gut I, Gyllenberg A, Hemmrich-Stanisak G, Hjorth L, Ioannidis Y, Jarmalaite S, Kel A, Kherif F, Korbel JO, Larue C, Laszlo M, Maas A, Magalhaes L, Manneh-Vangramberen I, Morley-Fletcher E, Ohmann C, Oksvold P, Oxtoby NP, Perseil I, Pezoulas V, Riess O, Riper H, Roca J, Rosenstiel P, Sabatier P, Sanz F, Tayeb M, Thomassen G, Van Bussel J, Van den Bulcke M, Van Oyen H. Towards a European health research and innovation cloud (HRIC). Genome Med. 2020 Feb 19;12(1):18. doi: 10.1186/s13073-020-0713-z.

[43] https://www.ebi.ac.uk/ena/pathogens/v2/monkeypox

- Begley CG, Ioannidis JP. Reproducibility in science: improving the standard for basic and preclinical research. Circ Res. 2015 Jan;116(1) 116-126. doi:10.1161/circresaha.114.303819.
- Canham S, Ohmann C, Thomassen G, Matei M, Demotes J, Panagiotopoulou M. (2020). EOSC-Life Strategic plan for the development of a COVID-19 repository including specification of technical requirements, policies and procedures - v2. Zenodo. https://doi.org/10.5281/zenodo.4341385
- Canham S, Felder G, Ohmann C, Panagiotopoulou M. (2022). Identification of COVID-19 clinical studies intending to share individual participant data for secondary use: Protocol for a pilot study. Zenodo. https://doi.org/10.5281/zenodo.7064624
- Clark H, Cárdenas M, Dybul M, Kazatchkine M, Liu J, Miliband D, Nordström A, Sudan P, Zedillo E, Obaid T, McCarney R, Radin E, Eliasz MK, McNab C, Legido-Quigley H, Sirleaf EJ. Transforming or tinkering: the world remains unprepared for the next pandemic threat. Lancet. 2022 May;399(10340) 1995-1999. doi:10.1016/s0140-6736(22)00929-1.
- Couzin-Frankel J, The power of negative thinking, 2013, Science , Vol. 342: p. 68-69
- David R, Ohmann C, Boiten JW, Abadía MC, Bietrix F, Canham S, Chiusano ML, Dastrù W, Laroquette A, Longo D, Mayrhofer MT, Panagiotopoulou M, Richard AS, Goryanin S, Verde PE. An iterative and interdisciplinary categorisation process towards FAIRer digital resources for sensitive life-sciences data. Sci Rep. 2022 Dec;12(1) 20989. doi:10.1038/s41598-022-25278-z.
- David, R., Richard, A.S., Connellan, C., Lauer, K.B., Chiusano, M.L., Goble, C., Houde, M., Kemmer, I., Keppler, A., Lieutaud, P., Ohmann, C., Panagiotopoulou, M., Khan, S.R., Rybina, A., Soiland-Reyes, S., Wit, C., Wittner, R., Buono, R.A., Marsh, S.A., Audergon, P., Bonfils, D., Carazo, J.-M., Charrel, R., Coppens, F., Fecke, W., Filippone, C., Alvarez, E.G., Gul, S., Hermjakob, H., Herzog, K., Holub, P., Kozera, L., Lister, A.L., López-Coronado, J., Madon, B., Majcen, K., Martin, W., Müller, W., Papadopoulou, E., Prat, C.M.A., Romano, P., Sansone, S.-A., Saunders, G., Blomberg, N. and Ewbank, J., 2023. Umbrella Data Management Plans to Integrate FAIR Data: Lessons From the ISIDORe and BY-COVID Consortia for Pandemic Preparedness. Data Science Journal, 22(1), p.52.DOI: https://doi.org/10.5334/dsj-2023-035
- David R, Rybin A, Burel J-M, Heriche J-K, Audergon P, Boiten J-M, Coppens F, Crockett S, Exter K, Fahrener S, Fratelli M, Goble C, Gormanns P, Grantner T, Gruning B, Gurwitz K, Hancock J, Harmse H, Holub P, Gribbon P. (2023). Preprint: "Be Sustainable", Recommendations for FAIR Resources in Life Sciences research: EOSC-Life's Lessons. In EMBO Journal (6.3). Zenodo. https://doi.org/10.5281/zenodo.8338931
- De Geest P, Coppens F, Soiland-Reyes S, Eguinoa I, Leo S (2022) Enhancing RDM in Galaxy by integrating RO-Crate. Research Ideas and Outcomes 8: e95164. https://doi.org/10.3897/rio.8.e95164

- De Smedt K, Koureas D, Wittenburg P (2020): FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. Publications 8(2):21. https://doi.org/10.3390/publications8020021

- Devriendt T, Borry P, Shabani M. Factors that influence data sharing through data sharing platforms: A qualitative study on the views and experiences of cohort holders and platform developers. PLoS One. 2021 Jul 2;16(7):e0254202. doi: 10.1371/journal.pone.0254202.

- Drury G, Jolliffe S, Mukhopadhyay TK. Process mapping of vaccines: Understanding the limitations in current response to emerging epidemic threats. Vaccine. 2019 Apr;37(17) 2415-2421. doi:10.1016/j.vaccine.2019.01.050.

- Estupiñán-Romero, Francisco, Van Goethem, Nina, Meurisse, Marjan, González-Galindo, Javier, & Bernal-Delgado, Enrique. (2023). BY-COVID - WP5 - Baseline Use Case: SARS-CoV-2 vaccine effectiveness assessment - Common Data Model Specification (1.1.0). Zenodo. https://doi.org/10.5281/zenodo.7572373

- García Álvarez, E., Mayrhofer, M. & Holub, P. BY-COVID - D8.2 - Data Management Plan. (2021) doi:10.5281/zenodo.6884816.

- Goble C, Soiland-Reyes S, Bacall F, Owen S, Williams A, Eguinoa I, Droesbeke B, Leo S, Pireddu L, Rodriguez-Navas L, Fernández J, Capella-Gutierrez S, Ménage H, Grüning B, Serrano-Solano B, Ewels P, Coppens F (2021): Implementing FAIR Digital Objects in the EOSC-Life Workflow Collaboratory. Zenodo https://doi.org/10.5281/zenodo.4605654

- Goble, C, Bacall, F, Soiland-Reyes, S, et al., 2022, 'WorkflowHub – a FAIR registry for workflows'. F1000Research, 11. DOI: https://doi.org/10.7490/f1000research.1118984.1

- Gray A, Goble C, Jimenez R (2017): Bioschemas: From Potato Salad to Protein Annotation. Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks. CEUR Workshop Proceedings 1963:579 https://ceur-ws.org/Vol-1963/paper579.pdf

- Hermjakob H, Kleemola M, Moilanen K, Tuominen M, Sansone S, Lister A, David R, Panagiotopoulou M, Ohmann C, Belien J, Lischke J, Juty N, Soiland-Reyes S. (2022). BY-COVID D3.2: Implementation of cloud-based, high performance, scalable indexing system. Zenodo. https://doi.org/10.5281/zenodo.7129553

- Jefferson E, Cole C, Mumtaz S, Cox S, Giles TC, Adejumo S, Urwin E, Lea D, Macdonald C, Best J, Masood E, Milligan G, Johnston J, Horban S, Birced I, Hall C, Jackson AS, Collins C, Rising S, Dodsley C, Hampton J, Hadfield A, Santos R, Tarr S, Panagi V, Lavagna J, Jackson T, Chuter A, Beggs J, Martinez-Queipo M, Ward H, von Ziegenweidt J, Burns F, Martin J, Sebire N, Morris C, Bradley D, Baxter R, Ahonen-Bishopp A, Smith P, Shoemark A, Valdes AM, Ollivere B, Manisty C, Eyre D, Gallant S, Joy G, McAuley A, Connell D, Northstone K, Jeffery K, Di Angelantonio E, McMahon A, Walker M, Semple MG, Sims JM, Lawrence E, Davies B, Baillie JK, Tang M, Leeming G, Power L, Breeze T, Murray D, Orton C, Pierce I, Hall I, Ladhani S, Gillson N, Whitaker M, Shallcross L, Seymour D, Varma S, Reilly G, Morris A, Hopkins S, Sheikh A, Quinlan P. (2022), A Hybrid Architecture (CO-

CONNECT) to Facilitate Rapid Discovery and Access to Data Across the United Kingdom in Response to the COVID-19 Pandemic: Development Study J Med Internet Res 2022; 24(12):e40035, doi: 10.2196/40035

- Koopmans M, de Lamballerie X, Jaenisch T, ZIKAlliance Consortium. Familiar barriers still unresolved-a perspective on the Zika virus outbreak research response. Lancet Infect Dis. 2019 Feb;19(2) e59-e62. doi:10.1016/s1473-3099(18)30497-3.

- Knyazev, S., Chhugani, K., Sarwal, V. et al. Unlocking capacities of genomics for the COVID-19 response and future pandemics. Nat Methods 19, 374–380 (2022). https://doi.org/10.1038/s41592-022-01444-z

- Ko, K.K.K., Chng, K.R. & Nagarajan, N. Metagenomics-enabled microbial surveillance. Nat Microbiol 7, 486–496 (2022). https://doi.org/10.1038/s41564-022-01089-w

- Larson K, Sim I, von Isenburg M, Levenstein M, Rockhold F, Neumann S, D'Arcy C, Graham E, Zuckerman D, Li R. COVID-19 interventional trials: Analysis of data sharing intentions during a time of pandemic. Contemp Clin Trials. 2022 Apr;115:106709. doi: 10.1016/j.cct.2022.106709. Epub 2022 Feb 16.

- Leo S, Crusoe M, Rodríguez-Navas R, Sirvent R, Kanitz A, De Geest P, Wittner R, P Luca, Daniel Garijo, Fernández J, Colonnelli I, Gallo M, Ohta T, Suetake H, Capella-Gutierrez S, de Wit R, Kinoshita B, Soiland-Reyes S (2023): Recording provenance of workflow runs with RO-Crate. (In preparation)

- Lohse, S., & Canali, S. (2021). Follow *the* science? On the marginal role of the social sciences in the COVID-19 pandemic. European Journal for Philosophy of Science, 11(4), 1-28. DOI: https://doi.org/10.1007/s13194-021-00416-y

- Maier, W., Bray, S., van den Beek, M. et al. Ready-to-use public infrastructure for global SARS-CoV-2 monitoring. Nat Biotechnol 39, 1178–1179 (2021). https://doi.org/10.1038/s41587-021-01069-1

- Mendez-Villalon A, Giles T, Urwin E, & Cox S. (2022). BY-COVID synthetic RO-Crates [Data set]. Zenodo. https://doi.org/10.5281/zenodo.6940413

- Merson L, Ndwandwe D, Malinga T, Paparella G, Oneil K, Karam G, Terry RF. Promotion of data sharing needs more than an emergency: An analysis of trends across clinical trials registered on the International Clinical Trials Registry Platform. Wellcome Open Res. 2022 Mar 21;7:101. doi: 10.12688/wellcomeopenres.17700.1.

- Meurisse M, Estupiñán-Romero F, Van Goethem N, González-Galindo J, Royo-Sierra S, Bernal-Delgado E (2023a): BY-COVID - WP5 - Baseline Use Case: SARS-CoV-2 vaccine effectiveness assessment. BY-COVID Project, RO-Crate. https://doi.org/10.5281/zenodo.6913045 https://w3id.org/ro/doi/10.5281/zenodo.6913045

- Meurisse M, Estupiñán-Romero F, González-Galindo J, Martínez-Lizaga N, Royo-Sierra S, Saldner S, Dolanski-Aghamanoukjan L, Degelsegger-Marquez A, Soiland-Reyes S, Van Goethem N, Bernal-Delgado E, on behalf of BeYond-COVID project contributors (2023b):

Federated causal inference based on real-world observational data sources: application to a SARS-CoV-2 vaccine effectiveness assessment. BMC Medical Research Methodology. (in consideration)

- Montaldo, C., Messina, F., Abbate, I. et al. Multi-omics approach to COVID-19: a domain-based literature review. J Transl Med 19, 501 (2021). https://doi.org/10.1186/s12967-021-03168-8
- Mutuku C, Gazdag Z, Melegh S. Occurrence of antibiotics and bacterial resistance genes in wastewater: resistance mechanisms and antimicrobial resistance control approaches. World J Microbiol Biotechnol. 2022 Jul 4;38(9):152
- Nature editorial, In praise of replication studies and null results, 2020, Nature , Vol. 578: p. 489-490
- Ohmann C, Moher D, Siebert M, et al. Status, use and impact of sharing individual participant data from clinical trials: a scoping reviewBMJ Open 2021;11:e049228. doi: 10.1136/bmjopen-2021-049228
- Ohmann C, Matei M, Canham S, Panagiotopoulou M, Demotes J, Thomassen G, Blomberg, N. (2021). EOSC-Life WP14: COVID-19 Repository Data Sharing Policy (Version 1). Zenodo. https://doi.org/10.5281/zenodo.5519122
- Olliaro P, Torreele E. Global challenges in preparedness and response to epidemic infectious diseases. Mol Ther. 2022 May 4;30(5):1801-1809. doi: 10.1016/j.ymthe.2022.02.022. Epub 2022 Feb 23.
- Rocca-Serra, P., Gu, W,, Ioannidis, V,, et al., 2022, 'The FAIR Cookbook - the essential resource for and by FAIR doers'. Zenodo, DOI: https://doi.org/10.5281/zenodo.7156792
- Sansone, S-A, McQuilton, P, Rocca-Serra, P, et al., 2019. 'FAIRsharing as a community approach to standards, repositories and policies'. Nat Biotechnol, 37: 358–367. DOI: https://doi.org/10.1038/s41587-019-0080-8
- Soiland-Reyes S, Sefton P, Crosas, M, Jael Castro L, Coppens F, Fernández J, Garijo D, Grüning B, La Rosa M, Leo S, Ó Carragáin E, Portier M, Trisovic A, RO-Crate Community, Groth P, Goble C (2022):Packaging research artefacts with RO-Crate. Data Science 5(2) https://doi.org/10.3233/DS-210053
- Stuart, D; Baynes, G; Hrynaszkiewicz, I; Allin, K; Penny, D; Lucraft, M et al. (2018): Whitepaper: Practical challenges for researchers in data sharing. figshare. Journal contribution. https://doi.org/10.6084/m9.figshare.5975011.v1
- Taichman DB, Backus J, Baethge C, et al. Sharing clinical trial data: a proposal from the International Committee of medical Journal editors. PLoS Med 2016;13:e1001950.
- Vines, TH, Albert, AYK, Andrew, RL, et al., 2014, 'The Availability of Research Data Declines Rapidly with Article Age', Current Biology 24, 94–97. DOI: https://doi.org/10.1016/j.cub.2013.11.014

- Wellcome Trust (2003) Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility, Fort Lauderdale. https://www.sanger.ac.uk/wp-content/uploads/fortlauderdalereport.pdf
- Wilkinson, MD, Dumontier, M, Aalbersberg, IjJ, et al., 2016. 'The FAIR Guiding Principles for scientific data management and stewardship'. Scientific Data, 3: 160018. DOI: https://doi.org/10.1038/sdata.2016.18
- Wittner R, Mascia C, Gallo M, Frexia F, Müller H, Plass M, Geiger J, Holub F (2022): Lightweight Distributed Provenance Model for Complex Real–World Environments. Scientific Data 9:503 https://doi.org/10.1038/s41597-022-01537-6
- Wittner R, Holub P, Mascia C, et al (2023). Toward a common standard for data and specimen provenance in life sciences. Learn Health Sys. 2023;e10365. https://doi.org/10.1002/lrh2.10365
- Yehudi Y, Hughes-Noehrer L, Goble C, Jay C (2022): COVID-19: An exploration of consecutive systemic barriers to pathogen-related data sharing during a pandemic. arXiv 2205.12098 [cs.CY] (preprint) https://doi.org/10.48550/arXiv.2205.12098
- Zdravkovic M, Berger-Estilita J, Zdravkovic B, Berger D (2020) Scientific quality of COVID-19 and SARS CoV-2 publications in the highest impact medical journals during the early phase of the pandemic: A case control study. PLoS ONE 15(11): e0241826. https://doi.org/10.1371/journal.pone.0241826