BY-COVID

POLICY BRIEF

Research data infrastructure for sustainable global pandemic preparedness

BY-COVID recommendations from consultation with key stakeholders from academia, industry and policy on data infrastructures to support European pandemic preparedness.

Authors:
Peter Maccallum (ELIXIR) ORCID 0000-0001-5260-5915;
Henning Hermjakob (EMBL-EBI);
Maria Tyler (ELIXIR);
Elaine Harrison (ELIXIR) ORCID 0000-0003-1149-2242;
Yun-Yun Tseng (ELIXIR)

December 2023

Contact:
by-covid-admin@elixir-europe.org

Funded by the European Union

Summary

This second Policy Brief from the BY-COVID project supplements the first, on 'Open data to support pandemic preparedness', by addressing the role of Research Infrastructures in the global pandemic preparedness landscape.

The BY-COVID project has built data partnerships to support pathogen research and pandemic response. The project brought together researchers, infrastructure and database providers and policymakers, initially around COVID-19 resources but gradually developing a vision for a general capability extensible to all pathogens.

Core to the project have been the ten participating European Research Infrastructures. They brought established services to build on, expertise to develop new solutions and have the potential to provide long-term hosts for knowledge and new services.

Through the project and consultation with our partners we have developed a picture of what is required in a future pandemic response architecture. The network of international reference and deposition databases are at the heart to support data sharing. Portals and analysis environments developed to support the COVID-19 response can be developed for all pathogens in partnership with infrastructure providers. Development of data standards, reusable analytic tools and communities of practice can be supported by the Research Infrastructures.

Here, we offer recommendations for policymakers and other stakeholders: that consolidated resources for pathogen data are developed at the international and national levels; that pathogen researchers partner with the e-Infrastructures to ensure technologies are aligned and capacity for outbreak response is available; and that the Research Infrastructures maintain pathogens communities of practice in preparation for a future pandemic.

POLICY BRIEF

BY-COVID – data partnerships in pandemic response

The response to COVID-19

In 2021 the European Union was mobilising its scientific and technological assets to support the response to the COVID-19 epidemic and increase preparedness for future outbreaks[1]. The BeYond-COVID (BY-COVID) project[2,3] was funded under a call for "FAIR and open data sharing in support to European preparedness for COVID-19 and other infectious diseases" (HORIZON-INFRA-2021-EMERGENCY-01)[4] and ran from October 2021 to September 2024.

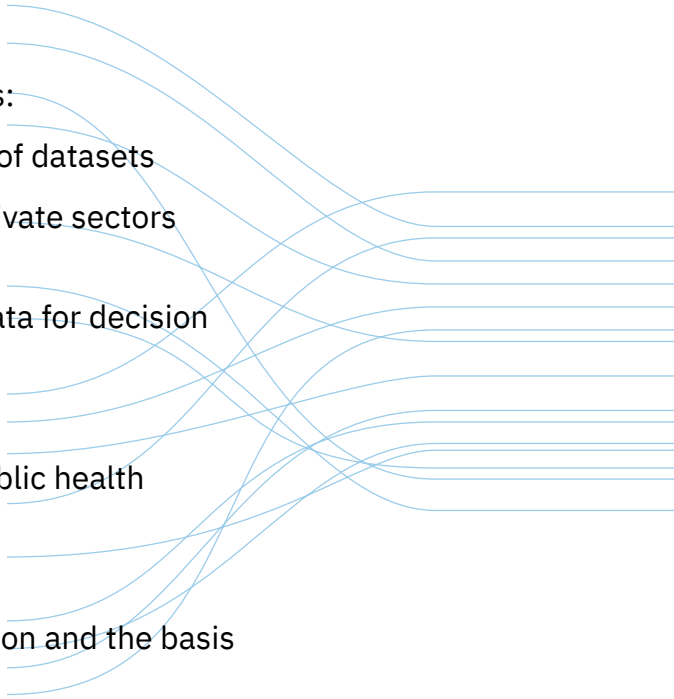
BY-COVID took the approach of developing common approaches to the preparation and sharing of datasets – 'data mobilisation' – based on existing tools and databases integrated with custom portals and resources (such as the COVID-19 Data Portal[5]). Linking research data, clinical data and social science datasets is crucial to understand the full impact of this complex event.

> *BY-COVID works to: Enable researchers, healthcare professionals and citizens (in terms of consent to share) fighting the spread of infectious diseases to store, document, share, access, analyse, link and process research and clinical data across disciplines and national borders*[6]

BY-COVID stakeholders

The project's activities connected a range of stakeholders:

- Data generators handling the deposition or curation of datasets
- Data consumers, researchers from the public and private sectors building analysis workflows and looking for insights
- Policy makers and funders looking for high quality data for decision making
- Infrastructure providers
- National and international infectious disease and public health agencies and collaborators
- Medical service providers and consumers
- Members of the public looking for accurate information and the basis of policy decisions affecting them



> *The sectors covered by the BY-COVID project include social sciences, public health, molecular life sciences and medical sciences.*[7]

The European COVID-19 Data Portal

The BY-COVID project was focused around the European COVID-19 Data Portal, supporting the portal's operation and enhancing its data and capabilities. The portal became a beacon for the benefits of open science and open data and an exemplar of how a single coordinating point for pathogen related data could help organise outbreak and pandemic response planning. The approach of indexing and connecting datasets from a wide variety of high-quality data sources validated the prediction that open data principles help maximise the societal impact of research data re-use.

> *'... the portal contains c.15 million viral sequences and more than one million open scientific articles related to SARS-CoV-2 and COVID-19.[8]'*

COVID-19 Data Portal

From COVID-19 to all pathogens

Two conclusions can be drawn from reflecting on the pandemic response – first, that the complexity of a pandemic caused by a rapidly-evolving pathogen does indeed need a specialised approach and resources, and second, that over-specialisation of response plans to a small number of likely pathogens is unlikely to be effective.

As BY-COVID has developed in the face of maturing COVID-19 countermeasures and understanding, the project turned its attention to generalising tools and approaches which can be redeployed in the event of the widest possible range of unknown future outbreaks. As anticipated, focus has moved from COVID-19 tools and portals to more generalised pathogen resources.

> *'While COVID-19 variants remain a significant public health risk, the focus of the international research community, as well as policy makers, is shifting to broader pandemic preparedness...[8]'*

POLICY BRIEF

The European Research Infrastructures support preparedness and response

Among the 53 participants in BY-COVID, 10 were members of the European Strategy Forum for Research Infrastructures (ESFRI) and one (EMBL) a member of the European Intergovernmental Research Organisation (EIRO) Forum. The Research Infrastructures brought three benefits:

- Background knowledge and services pre-dating the COVID-19 pandemic (members have been established since the 1970s, and most of the ESFRI members were initiated before 2010)
- Skills and expertise in managing and using data and analyses
- Experience of developing and running long-lived services which can last beyond the responsive mode funding

European ESFRI and EIRO Infrastructures participating in BY-COVID

EMBL
BBMRI-ERIC
cessda
eatris
ecrin
elixir
instruct ERIC
INFRAFRONTIER
EURO BIOIMAGING
erinha
European Research Infrastructure on Highly Pathogenic Agents
MIRRI ERIC

Together these infrastructures provided hundreds of individual services, including databases, software and laboratory facilities. During the BY-COVID project, specialised services for COVID-19 researchers were developed but a significant effort also went into the enhancement of existing services and data standards to better support pathogen researchers alongside existing users.

Services specialised for pathogens	Services and standards with application beyond pathogens
COVID-19 Data Portal	ENA European Nucleotide Archive
COVID-19 Data Hubs	BioImage Archive
PATHOGENS Surveillance | Identification | Investigation	crDSR Clinical Research Data Sharing Repository
Pathogen Data Hubs	WorkflowHub
Infectious Diseases Toolkit	RO-Crate
	Galaxy
	FAIRsharing.org standards, databases, policies
	COVID-19 Disease Map

Lessons learned beyond COVID-19

A network of international reference and deposition databases supports data sharing

From the outset of the pandemic, the early publication of viral sequences had a profound impact on the eventual outcome. The public deposition of the viral sequences from Wuhan[9]were the trigger for extensive programmes of vaccine and countermeasure development and the subsequent programmes of vaccination may have saved tens of millions of lives and restored economies from the impact of travel bans and lockdowns.

> *"...we estimated that vaccinations prevented 14.4 million deaths from COVID-19 in 185 countries and territories between Dec 8, 2020 and Dec 8, 2021.[10]"*

At the heart of the effort were the partner databases of the INSDC[11]. The European Nucleotide Archive (ENA), hosted by the EMBL European Bioinformatics Institute (EMBL-EBI) was one of the partners in the BY-COVID project, and expertise from the ENA team was crucial in the development of the European COVID-19 Data Portal. This is also true for the BioImage Archive, hosted by EMBL-EBI. The network of existing deposition databases run by the Research Infrastructures ensured that repositories were ready to accept novel data and that existing data of all types from related viruses and similar genes and proteins were available for researchers.

General pathogen resources can be re-purposed from specialised COVID-19 resources

The development of the European COVID-19 Data Portal, and the creation of national data portals based on the template, allowed research datasets and policies to be quickly shared and encouraged re-use. However, the integration of each new dataset took time, and maintenance of the specialised portals required dedicated effort. Once the immediate crisis was over, research priorities necessarily moved away and with COVID-19 under control, attention has moved to preparation for a future pandemic.

Given the unsustainability of specialised portals for each pathogen, the BY-COVID strategy was to support and promote the development of a new general purpose 'Pathogens Portal' which could be used to index and collate datasets on a range of pathogens. This general resource could also be quickly adapted to create specialised resources as the need arises. The new portal, and its partner national services, benefits from ready expertise and the option to use the existing codebase and established workflows.

> *"On July 5, 2023, we formally released the Pathogens Portal. The list of pathogens featured in the portal was collated using the UK's Health and Safety Executive's list of approved biological agents and the WHO's global priority pathogens list.[8]"*

e-Infrastructure are well positioned to support pathogen researchers and public health professionals to analyse and share outbreak data

Researchers not only faced a problem with data management during the pandemic, the unprecedented availability of viral and human nucleic acid sequences required the development of rapid and reliable analysis pipelines which could be deployed at scale. The Galaxy platform[12] offered scalable solutions, but secure environments were also required for analysis of and preparation of clinical outcomes data along with sequence datasets before open publication.

The approach was to build 'Data Hubs' based on pre-existing concepts to provide analysis, visualisation, presentation and submission tools for identified groups of collaborators. Most of the tools provided are generalised and can be used to support the analysis and submission of data from any pathogen.

These solutions have been built by the hosting institutions. It is clear from the course of the COVID-19 epidemic that continued development of the capability is wise; and that when needed, they could be rapidly scaled to deal with population-size datasets. This could be supported by the European e-Infrastructures and High Performance Computing facilities, which have technical skills to help the development of the services during 'preparedness' research phases.

> *'By using this concept, we aim to provide a set of tools spanning these areas in order to support researchers in preparedness, focusing on a select group of priority pathogens that are likely to cause the next infectious disease outbreak.[13]'*

Reusable, interoperable data tools support data quality and decision making

Data and infrastructure were not the only components required; the FAIR principles and the requirements of reproducibility also required the development and promotion of standards for data representation and the description of re-usable analyses. The use of open standards for FAIR digital objects, such as RO-Crate, were demonstrated as tools for interchange of workflow descriptions. In conjunction with registries of computational methods such as the Workflow Hub, the BY-COVID partners were able to encourage the re-use and standardisation of bioinformatics practices, which in turn improves the assurance and quality of deposited datasets and subsequent policy decision making.

As with data, and infrastructure, standards and registries require continued investment. The Research Infrastructures play a significant role here, as a permanent home and testbed for established standards.

> *'The use of RO-Crate to collect crucial analysis parameters on the use of a workflow is thus an important step to increase trustworthiness of analysis results.[14]'*

Knowledge hubs and communities of practice should be maintained to ensure rapid outbreak response

The 53 partners in BY-COVID included a range of professionals, from experts in pathogens and public health to generalist bioinformaticians and infrastructure developers. During the pandemic many became skilled in and contributed to the global research response and most developed some degree of specialist expertise. This expertise was captured and shared with the development of a specialised knowledge base, the Infectious Diseases Toolkit.

> *‘The toolkit captures valuable information across a broad range of domains and topics, from molecular biology to clinical data and socioeconomic data … enhancing the ability of the scientific community to respond effectively to future outbreaks.[15]’*

Beyond COVID-19, the European strategy is to have permanent capabilities such as the Health Emergency Preparedness and Response Authority[16] with a research and preparedness mode and an emergency response mode. This is replicated elsewhere in the Research Infrastructures, working groups of specialists and interested researchers in Research Infrastructures providing a focus to sustain communities of practice with the necessary expertise[17].

BY-COVID’s overall recommendations

Through our experience in the BY-COVID project, consultation in policy events in 2023 and 2024 with European and international policy partners, and looking ahead to the future European and international landscape, we make the following recommendations for the sustainability of pathogen preparedness embedded in the Research Infrastructures. These should be taken alongside recommendations in the first BY-COVID policy brief on the use of open data to support European pandemic preparedness.

Support core databases and pathogen data resources

Instead of building specialised resources for individual pathogens, direct support to a network of resources with existing deposition databases at its heart, and a suite of generalised international or national pathogen portals and tools which can be rapidly re-configured for outbreak response.

Partner with e-Infrastructures for pathogen research and pandemic response

A network of on-demand national or regional data hubs designed to common standards is the best way to support basic research and, in partnership with large scale infrastructure providers, rapidly scale to the analysis population scale datasets.

Maintain pathogens communities of practice

Each of the Research Infrastructures found expertise and tools to support the global pandemic response. This expertise should be maintained, not just in the infrastructures where pathogens and public health are a primary concern but across all of the life science Research Infrastructures.

Recommendations specific to policymakers

1

Encourage the use of existing large scale international repositories of data of all types – sequencing, imaging, social and clinical – to maximise re-use and optimise long term sustainability through scale.

2

Invest in the ability to responsively create views and indexes of prior and current data based around outbreaks, pandemics and pathogens of concern, allowing the dynamic creation of national and international resources in response to immediate public health needs.

3

Develop national data hub approaches for the analysis of surveillance and outbreak data which re-use international toolkits of data standards, software, workflows and brokering to large scale repositories.

4

Encourage the development of partnerships between e-Infrastructure and High Performance Computing providers to ensure emergency capacity is available to host and expand data analysis workflows on demand.

5

Support the creation of specialist working groups on pathogen research within each Research Infrastructure, ready to support outbreak response with high quality data as the need arises.

6

Build global connections to exchange best practices, knowledge and infrastructure strategies.

References

1. *"ERAvsCORONA"* Action Plan [https://research-and-innovation.ec.europa.eu/system/files/2020-04/ec_rtd_era-vs-corona.pdf]
2. Beyond COVID [https://cordis.europa.eu/project/id/101046203]
3. BY-COVID [https://by-covid.org]
4. *FAIR and open data sharing in support to European preparedness for COVID-19 and other infectious diseases* [https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/horizon-infra-2021-emergency-01]
5. COVID-19 Data Portal [https://www.covid19dataportal.org]
6. *Open data to support European pandemic preparedness* [https://doi.org/10.5281/zenodo.7950479]
7. *BY-COVID D7.1 – Communication, Dissemination and Exploitation Strategy* [https://doi.org/10.5281/zenodo.6884870]
8. *BY-COVID D3.3.2 – COVID-19 Data Portal* [https://doi.org/10.5281/zenodo.13332975]
9. *A new coronavirus associated with human respiratory disease in China* [https://doi.org/10.1038/s41586-020-2008-3]
10. *Global impact of the first year of COVID-19 vaccination: a mathematical modelling study* [https://doi.org/10.1016/S1473-3099(22)00320-6]
11. International Nucleotide Sequence Database Collaboration [https://www.insdc.org]
12. *Selection Analysis Identifies Clusters of Unusual Mutational Changes in Omicron Lineage BA.1 That Likely Impact Spike Function* [https://doi.org/10.1093/molbev/msac061]
13. *BY-COVID D1.2 – Preparedness Data Hub* [https://doi.org/10.5281/zenodo.10069940]
14. *BY-COVID D4.2 – Common analysis environment* [https://doi.org/10.5281/zenodo.12706837]
15. *BY-COVID D4.1 – Infectious Diseases Toolkit* [https://doi.org/10.5281/zenodo.13909280]
16. *Health Emergency Preparedness and Response Authority* [https://health.ec.europa.eu/health-emergency-preparedness-and-response-hera_en]
17. ELIXIR Pathogen Data Focus Group [https://elixir-europe.org/focus-groups/pathogen-data]