Horizon 2020 projects working on the 2019 coronavirus disease (COVID-19), the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), and related topics:

Guidelines for open access to publications, data and other research outputs

Version 1.0
April 8, 2020

Research and Innovation

Introduction

On 30 January 2020 the World Health Organization (WHO) declared that the SARS-CoV-2 outbreak constitutes a Public Health Emergency of International Concern (PHEIC). The COVID-19 crisis is putting high pressure on the research community to speed up science discovery, inform the public health response and help save lives, as demonstrated by the activation by the WHO of the R&D Blueprint to accelerate diagnostics, vaccines and therapeutics for this novel virus. A necessary complementary action to accelerate and amplify impact is to ensure that research findings and data relevant to this outbreak, are shared as rapidly, openly and effectively as possible.

Therefore, the European Commission urges researchers of Horizon 2020 grants with research outputs that - in any way - may be used to advance the research on COVID-19, to provide immediate open access to their related publications, data and any other output possible, in line with the guidance offered in this document. These can be projects specifically researching on the corona virus, but also other research fields/disciplines with relevance to tackle the corona crisis. Similarly, the European Commission urges research infrastructures projects, developing and/or providing access services to relevant research tools and resources, to provide priority and customised access to their services for research on COVID-19.

More particularly, the European Commission strongly encourages beneficiaries to follow the guidelines below, thereby exceeding the current Open Access requirements of Horizon 2020 and going beyond the legal obligations enshrined in the Horizon 2020 Grant Agreement (GA), in order to address the current public health emergency. The guidelines build on both the commitments made by the European Commission as a signatory of the Statement on Data Sharing in Public Health Emergency, and on the principles established in the GA. As signatory of the Statement on Data Sharing in Public Health Emergency, the European Commission promotes that:

- *"All peer-reviewed research publications relevant to the outbreak are made immediately open access, or freely available at least for the duration of the outbreak;*
- *Research findings relevant to the outbreak are shared immediately with the WHO upon journal submission, by the journal and with author knowledge;*
- *Research findings are made available via preprint servers before journal publication, or via platforms that make papers openly accessible before peer review, with clear statements regarding the availability of underlying data;*
- *Researchers share interim and final research data relating to the outbreak, together with protocols and standards used to collect the data, as rapidly and widely as possible - including with public health and research communities and the WHO;*
- *Authors are clear that data or preprints shared ahead of submission will not pre-empt its publication in these journals."*

The present document provides detailed guidance for projects working on COVID-19, SARS-CoV-2, and related topics, on:

- The FAIR principles;
- Open access to publications;
- Open access to data;
- Data Management Plans;
- Other research outputs;
- Ongoing data-related efforts under the umbrella of the European Open Science Cloud;
    - Relevant initiatives supported by the European Commission;
    - European Research Infrastructures
- Other useful tools and resources;
    - Publication and data resources;
    - Data repositories;
    - Data standards;
    - Data catalogues;
    - Data Management Plans;
    - Support services and tools.

FAIR principles

- **Manage** all research outputs (publications, data, and other outputs) in line with the FAIR principles to facilitate their re-use in the future and to support actions aiming to link COVID-19 or SARS-CoV-2 related research via dedicated platforms. Every effort should be made to curate all digital assets to satisfy the principles of findability, accessibility, interoperability, and reusability (see the principles on the GO-FAIR website). Digital assets are
    - **Findable** when they are described by sufficiently rich metadata and registered or indexed in a searchable resource. Digital assets should be uniquely identified through the use of Persistent Identifiers that are globally resolvable (PIDs).
    - **Accessible** when they can be obtained by humans and machines upon appropriate authorisation and through a well-defined and universally implementable protocol.
    - **Interoperable** when they follow a formal, accessible, shared and broadly applicable format and when a language for knowledge representation is used.
    - **Re-usable**, when rich metadata and documentation is provided that follow relevant community standards and provide information on provenance.

- **Remember** that the FAIR principles do not impose openness on digital assets, but rather refer to the adequate provisions under which the digital assets are curated. For example, data can be fully in line with the FAIR principles even when access is restricted.

Publications

- **Make** all research publications relevant to the outbreak immediately available, through deposition of a copy of the published, or final, peer-reviewed version, in a repository (through which open access to the deposited copy shall be ensured), at the latest at the time of publication, under a Creative Commons Attribution 4.0 International Public License (CC BY 4.0) or a license with equivalent rights. If you publish in an open access venue or hybrid journal (subscription-based journals in which some content is open access against the payment of a fee), Article Processing Charges (APCs) are eligible costs in the projects. If your discipline does not have a preferred repository and there is no repository in your institution, you may deposit your publications and data with the general-purpose repository Zenodo at no cost.

- **Make** research findings available via preprint servers before journal publication, or via platforms that make publications openly accessible before peer-review. Include clear statements regarding the availability of underlying data. Preprint servers are increasingly acceptable in most fields and accepted by most journals for later publication, but you must be aware to clear that data or preprints shared ahead of a submission will not pre-empt its publication in these journals. Some reliable and currently very relevant preprint archives are bioRxiv (life sciences), medRxiv (medical), PsyArxiv (behavioural sciences), SocArXiv (social sciences), ArXiv (o.a. physics, mathematics, computer science) and Open Science Framework (OSF) preprints or Zenodo (the latter two are multidisciplinary archives).

- **Provide** information via the repository about any research output or any other tools and instruments needed to re-use and/or validate the conclusions of the scientific publication. This includes for example software, workflows, models, materials etc. If possible, provide access to the tools or instruments themselves.

- **Include** metadata of deposited publications under a Creative Commons Public Domain Dedication (CC 0 1.0) or equivalent, in line with the FAIR principles (in particular machine-actionable) and provide information at least about the following:
  - Publication: author(s), title, date of publication, publication venue;
  - Framework Programme and the action: the terms "European Union" (EU) and "Horizon Europe" or "Euratom", respectively, the name of the action, acronym, grant number;
  - Licensing terms;
  - Persistent identifiers for the publication (e.g. DOI or Handle), the author(s) (e.g. ORCID, ResearcherID), and, if possible, for the institution(s) (e.g. ROR) and the grant (e.g. DOI) covered by this agreement.
  - Where applicable, persistent identifiers for any research output or any other tools and instruments needed to validate the conclusions of the publication.

Research data

- **Make** research data openly accessible immediately, and in accordance with the FAIR principles. Currently, the Grant Agreement may require (if Art 29.3 option 1c for health actions targeting public health emergencies is active) that you make data accessible at the latest within 30 days of generation. Given the current circumstances, we ask that you consider going beyond your legal obligations and provide immediate open access to all your relevant research data. The use of harmonised protocols in collaboration with other actors is recommended for this purpose. Open data should be licensed under Creative Commons Attribution 4.0 International Public License (CC BY 4.0) or a Creative Commons Public Domain Dedication (CC0 1.0) or a licence with rights equivalent to the above.

- **Develop** provisions for access to the data if open access is not possible because of exceptions as described in GA Article 29.3, so long as reasons for exceptions are respected. The principle 'As open as possible, as closed as necessary' applies. If open access is not provided to the data or any other research outputs needed to validate the conclusions of a scientific publication, the beneficiary should provide the access –digital or physical- needed for validation purposes to the extent that its legitimate interests or constraints are safeguarded.

- **Provide** open access to all data that may be useful to researchers. This includes protocols and standards used to collect the data as well as raw and other data that is not necessarily used for publication.

- **Deposit** quality-controlled research data in a data repository as soon as possible and within the deadlines set out in your data management plan (DMP). Data generated in the action, both those underpinning a scientific publication, but also stand-alone data, should be deposited in a repository (further resources under the Other useful tools and resources section below) that minimally, ensures the following:
    - Persistent and unique identifiers (PIDs)
    - Long term sustainability
    - Metadata
    - Curation and quality assurance
    - Access (e.g. free and easy access to re-use)
    - Security
    - Privacy
    - Common format
    - Provenance (e.g. maintains a detailed logfile of changes to datasets and metadata)

- **Provide** information via the repository about any research output or any other tools and instruments needed to re-use or validate the data. This includes for example software, workflows, models, materials etc. If possible, provide access to the tools or instruments themselves.

- **Include** metadata of deposited data under a Creative Commons Public Domain Dedication (CC 0 1.0) or equivalent, in line with the FAIR principles (in particular machine-actionable) and provide information at least about the following:
  - Dataset: description, date of deposit, author(s), venue and (if applicable) embargo;
  - Framework Programme and the action: the terms "European Union" (EU) and "Horizon Europe" or "Euratom", respectively, the name of the action, acronym, grant number);
  - Licensing terms;
  - Persistent identifiers for the dataset: the author(s) (e.g. ORCID, ResearcherID), and, if possible, for the institution(s) (e.g. ROR) and the grant (e.g. DOI);
  - Where applicable, a persistent identifier for the related publication(s) (e.g. DOI, Handle) and other research outputs.

Data Management Plans (DMPs)

- **Provide** a data management plan (DMP) preferably with the proposal or at the latest before grant signature. The data management plan should address the relevant aspects of making the data findable, accessible, interoperable and re-usable (FAIR), including:
  - A description of the data generated/collected (including data types) and an estimate of its size.
  - Whether and how the data will be made accessible for verification and re-use, along with relevant security and privacy considerations.
  - How the research data will be managed (organized, curated, accessed, shared, preserved)
  - Timelines of when generated data will be made open access.
  - An estimation of curation and storage/preservation costs; person/team responsible for data management and quality assurance processes.

- **Update** your DMP regularly: the DMP should be a living document which is updated and enriched with project outputs as the project evolves (e.g. new data sets, new publications, changes in data access or curation policies, etc.).

- **Register** your DMP as a non-restricted, public deliverable that is openly accessible, unless there are reasons (as per GA Article 29.3) to restrict it.

- **Remember** that costs for research data management (for example data storage, processing and preservation) are eligible if they comply with the costs eligibility requirements set out under GA Article 6.2.D.3.

- **Check** under the "Other useful tools and resources" section for useful resources that support the generation of DMPs.

Other research outputs

- **Manage** other research outputs in line with the FAIR principles, and fully document them in your DMP, to facilitate their re-use in the future and to support actions aiming to link corona virus-related research via dedicated platforms. Every effort should be made for other research outputs to be Findable, Accessible, Interoperable and Reusable.

Ongoing data-related efforts under the umbrella of the European Open Science Cloud

In this section, we provide a short overview of some European efforts regarding data sharing, in the context of the COVID-19 crisis, with links through which **we encourage you to explore synergies with, and fully exploit, the initiatives below**. In this section we highlight:

- Relevant initiatives supported by the European Commission;
- European Research Infrastructures.

Relevant initiatives supported by the European Commission

- **EMBL-EBI COVID-19 research data platform**: (https://www.ebi.ac.uk/covid-19)
    - The European Commission and the European Bioinformatics Institute of the European Molecular Biology Laboratory (EMBL-EBI), together with EU Member States and research infrastructure partners such as ELIXIR, are deploying a dedicated European COVID-19 research data platform which will facilitate comprehensive data sharing for the European and global research communities.
    - This joint effort is a priority pilot to realise the objectives of European Open Science Cloud. As such, all data and metadata accessible through this research data platform will be as open and FAIR as possible.
    - **Annex 1** describes the architectural components of the platform.
    - EMBL-EBI has provided **detailed and actionable data deposition guidelines** from up to eleven different data streams. You can find these guidelines in **Annex 2**..
    - EMBL-EBI is hoping to collaborate with governments, organisations and projects doing COVID-19 research and encourages them to direct questions as follows:
        - Questions on **how to contribute/link** to the European COVID-19 platform: ecovid19@ebi.ac.uk
        - **Technical requirements on data sharing**: virus-dataflow@ebi.ac.uk
        - **General enquiries regarding the initiative**: Dr. Rolf Apweiler, Director EMBL-EBI: apweiler@ebi.ac.uk

- **Research Data Alliance Working Group** on COVID-19: (https://rd-alliance.org/groups/rda-covid19)
    - The COVID-19 WG will create a set of deliverables by 24 April:
        - A set of **guideline documents,** aiming at developing a system for data sharing in public health emergencies that supports scientific research and policy making, including an overarching framework, common tools and processes. The guidelines will highlight the primary data sharing resources in five areas, each with different data types, and cross-cutting themes (e.g. ethics, legal, etc.):
            - Omics,
            - Clinical,
            - Epidemiology,
            - Social,
            - Community participation
        - A **set of resources** in each of those areas.
        - A **decision tree tool** to facilitate navigation to specific resources.

- **OpenAIRE**'s Zenodo COVID-19 community and COVID-19 gateway (https://www.openaire.eu/openaire-activities-for-covid-19):
    - OpenAIRE, in collaboration with the European Commission, will provide services to help in the sharing, discovery, navigation and collaboration of the global research community. Together with CERN, a Zenodo Community has been created to collect all research results relevant to COVID-19 and SARS-CoV-2. OpenAIRE is currently developing a service to serve as a single entry point for research results (publications, data, and software) for COVID-19 and SARS-CoV-2. This service will be complementary to, and will link up with, other EU initiatives and global efforts.

- **GO-FAIR's Virus Outbreak Data Network**: (https://www.go-fair.org/implementation-networks/overview/vodan/)
    - GO-FAIR has launched the Virus Outbreak Data Network (VODAN), an initiative to 'FAIRify', and to provide access to, COVID-19 data. Initially, the network will focus on the Clinical Research Form (CRF) model following the WHO standards. Via FAIR Data Points, data will be accessible while the personal data of patients remains in the underlying database of the local institution. With this model, GDPR issues are largely accommodated and data can be 'shared', or rather 'visited', without violating any patient rights.

- **BIP!Finder** for COVID-19: (https://bip.covid19.athenarc.gr/)
    - Ease the exploration of COVID-19 related literature. It is based on a) the CORD-19 dataset released by Semantic Scholar and b) the curated data released by the LitCovid hub.

European Research Infrastructures

It is advised to link with and make use of existing European Research Infrastructures whenever appropriate or necessary. This includes in particular:

- The European distributed infrastructure for **life-science information**, ELIXIR. ELIXIR nodes provide a range of services that can be used by researchers and consortia working on SARS-CoV-2 research (https://elixir-europe.org/covid-19-resources).
- The European research infrastructure for **biobanking** BBMRI-ERIC. Some further guidance relating to bio-banking in the context of SARS-CoV-2 and COVID-19 can be found at https://www.bbmri-eric.eu/services/bbmriqm-covid
- The European research infrastructure for multinational **clinical research** ECRIN-ERIC.
- The European research infrastructure for **translational medicine** EATRIS-ERIC.
- The European research infrastructure for **structural biology** INSTRUCT-ERIC. Structural biology services for research directly related to COVID-19 are proposed through a priority access pathway. A list of resources and national initiatives to support the structural biology community in COVID-19 research, including funded, rapid-access to research infrastructures, open access databases and literature, as well as volunteering opportunities, is available at: https://instruct-eric.eu/news/resources-information-and-collaborations-to-support-covid-19-research/.
- The European research infrastructure for **biological and biomedical imaging** Euro-BioImaging. Services and additional information on imaging research to combat COVID-19 are available at https://www.eurobioimaging.eu/content/Covid19.
- The European research infrastructure for **highly infectious emerging and remerging diseases** ERINHA.
- The **European Virus Archive** EVAg, with specific products available for SARS-CoV-2.
- The **vaccine research and development** infrastructure network TRANSVAC.

Consolidated information on dedicated services offered by Research Infrastructures against the COVID-19 pandemic, can be found at:

- The European Strategy Forum on Research Infrastructures (**ESFRI**) website, in a focused webpage that lists and provides quick links to the information gathered (https://www.esfri.eu/covid-19).

- A dedicated website of some of the European Life Science Research Infrastructures (https://lifescience-ri.eu/ls-ri-response-to-covid-19.html), with expert advice, a list of resources, remote access to services/facilities, and, where possible, with minimised or waived cost for access.

- The COVID-19 Fast Response Service, which is a coordinated and accelerated procedure for researchers to access the academic facilities, services and resources of the three research infrastructures BBMRI-ERIC, ECRIN-ERIC and EATRIS-ERIC.

Other useful tools and resources

Due to the relation of your research with the corona virus, we provide to you in this section useful tools and resources related to:

- Publication and data resources;
- Data repositories;
- Data standards;
- Data catalogues;
- Data Management Plan: tools for their creation and upkeep;
- Support services and tools.

Publication and data resources

In line with the recommendation in the present guidelines to provide immediate open access, we encourage you to refer to the Directory of Open Access Journals (DOAJ) as a useful resource to identify high quality, open access, peer-reviewed journals.

There are various freely-accessible collections and corpora that bring together COVID-19 related research information to search manually or mine using text and data mining techniques:

- **LitCovid** (https://www.ncbi.nlm.nih.gov/research/coronavirus/)
    - selection of publications in PubMed since January 2020
    - searching, full-text access (partly), metadata download
    - 1558 publications

- **Dimensions COVID-19** (https://dimensions.figshare.com/articles/Dimensions_COVID-19_publications_datasets_and_clinical_trials/11961063)
    - selection of publications in Dimensions, including datasets and clinical trials
    - searching, full-text access (partly), metadata download
    - 4164 publications, including preprints

- **1Science Coronavirus Research Repository** (https://coronavirus.1science.com/home)
    - selection of publications in 1Science, including SSRN preprints (not limited to Elsevier publications)
    - searching, full-text access (partly)
    - 35.064 publications

- **bioRxiv/medRxiv** COVID 19 (https://www.biorxiv.org/; https://www.medrxiv.org/)
    - preprints from bioRxiv & medRxiv
    - searching, full-text access

- **CORD-19** dataset (https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge)
    - publications from WHO corpus, PubMed Central, bioRiv/medRxiv and publishers
    - full-text download, metadata download
    - 29K publications (full text), 44K publication (metadata)

- **Global Initiative on Sharing All Influenza Data** (GISAID) (https://www.gisaid.org/)

- **Outbreak Science-PREreview** (https://outbreaksci.prereview.org/)
    - provides fast and journal independent peer-review of COVID-19, epidemic and related papers shared as preprint

- **OpenVirus** (https://github.com/petermr/openVirus/blob/master/README.md):
    - text- and data-mining of scholarly publications on virus outbreaks, especially that of COVID-19

Data Repositories

- **Useful listings of repositories** include the Registry of Research Data Repositories and the Core Trust Seal certified repositories. One key entry point for accessing and depositing related data and tools is the general-purpose repository Zenodo.

- **FAIRSharing COVID-19 resources** include a draft collection containing databases (which includes knoweldgebases and repositories) and standards that are responding to, or appropriate for, use in the COVID-19 pandemic. These resources may be focused on patient response, clinical trials, virology study or other related areas. Resources are continuously updated.

- For further details on general and discipline-specific repositories, visit the EUDAT Collaborative Data Infrastructure and the following:
    - **EMBL-EBI's Pathogen Portal:** provides access to genes, protein structures, Electron Microscopy data and scientific publications relating to COVID-19 (https://www.ebi.ac.uk/ena/pathogens/covid-19).
    - **Drug discovery bioactivity data:** https://www.ebi.ac.uk/chembl/.
    - **Protein structures :** https://www.ebi.ac.uk/pdbe/.
    - **Protein sequence and functional information:** https://www.uniprot.org/.
    - ELIXIR's Deposition Databases can be used to store experimental data from life science research and its Recommended Interoperability Resources can be used to make life science data FAIR.
    - **Clinical research data:** https://www.combacte.com/novel-coronavirus-faq/; https://isaric.tghn.org/covid-19-clinical-research-resources/.

Data standards

- EMBL-EBI has prepared a preliminary document addressing the current SARS-CoV-2 sequence data standards for submission and publication of data in the European Nucleotide Archive (ENA) which you will find as **Annex 3** in the present guidelines.

- FAIRsharing, as described above, provides a curated and searchable portal of data standards, databases, and policies in the life sciences and other scientific disciplines.

- For more information on disciplinary metadata standards, visit Digital Curation Centre and Research Data Alliance Metadata Standards Directory.

- Other sources include:
    - **eTRIKS** standards starter pack: https://zenodo.org/record/50825/files/eTRIKS_Standards_Starter_Pack_Release_1.1_April_2016.pdf
    - **Clinical:** Clinical Data Interchange Standards Consortium (CDISC): CDISC's current thinking is a highly expedited, lean process resulting in a COVID-19 Guide for CDISC Standards with a focus on Controlled Terminology (CT) and standardizing tabulation data clinical review. CDISC Chief Standards Officer Peter Van Reusel will lead this effort. If your project is actively starting a COVID-19 trial and would be interested in participating in this short-term initiative, please contact Peter Van Reusel at pvanreusel@cdisc.org.
    - **RWD** (Real World Data): Observational Medical Outcomes Partnership/ Observational Health Data Sciences and Informatics OMOP/OHDSI Common Data Model: https://forums.ohdsi.org/t/ohdsi-virtual-study-a-thon-to-support-covid-19-response-to-take-place-26-29mar2020-collaborators-wanted/9810

Data Catalogues

- Relevant data catalogues include:
  - WHO International Clinical Trials Registry Platform, **WHO ICTRP:** https://www.who.int/ictrp/en/.
  - The **EU Clinical Trials Register** and the European Union Drug Regulating Authorities Clinical Trials database, **EudraCT database**: https://eudract.ema.europa.eu/.
  - The **ClinicalTrials.gov database**: https://clinicaltrials.gov/
  - European Network of Centres for Pharmacoepidemiology and Pharmacovigilance, **ENCePP**: http://www.encepp.eu/.

Data Management Plans: tools

The following resources support the generation of DMPs:

- A template for the DMP can be found in the online H2020 guide for applicants, here.
- The DMPONLINE tool, specifically supports the development of project data management plans: https://dmponline.dcc.ac.uk
- ARGOS is an online tool to automate the creation, management, sharing and linking DMPs with the research artefacts to which they correspond. It is the joint effort of OpenAIRE and EUDAT CDI to deliver an open platform for Data Management Planning that addresses FAIR and Open best practices and assumes no barriers for its use and adoption (https://argos.openaire.eu).
- The Data Stewardship Wizard is a joint ELIXIR CZ and ELIXIR NL tool, bringing a simple but powerful solution for researchers to help them understand what is needed for good, FAIR-oriented data stewardship, and build their own Data Management Plans (https://ds-wizard.org/).

Support services and tools

- **European Medicines Agency (EMA):** EMA is ready to support medicine developers with all available regulatory tools to advance and expedite the development of effective measures to fight and prevent the spread of COVID-19. Developers of potential therapeutics or vaccines against COVID-19 are encouraged to contact the Agency as soon as possible with information about their proposed development, by emailing 2019-ncov@ema.europa.eu.
- **CROWDFIGHTCOVID19:** Service for researchers via platform aiming to redirect scientific resources towards the fight against COVID-19 at https://crowdfightcovid19.org/

- **Crowdhelix**: free match-making platform for COVID-19 researchers at https://network.crowdhelix.com/covid-19
- **Informed Consent Forms (ICF) templates**: DO-IT: e.g. https://cordis.europa.eu/project/id/116055/results; https://europa.eu/!QP48PW (Documents, reports)
- **FAIR self-assessment tool**: https://www.ands-nectar-rds.org.au/fair-tool

Annex 1: Components of the European COVID-19 Research Data Platform

Concept for the intended data platform

The European Research COVID-19 Data Platform will provide an open, trusted, and scalable pan-European environment where researchers can store and share relevant datasets including:

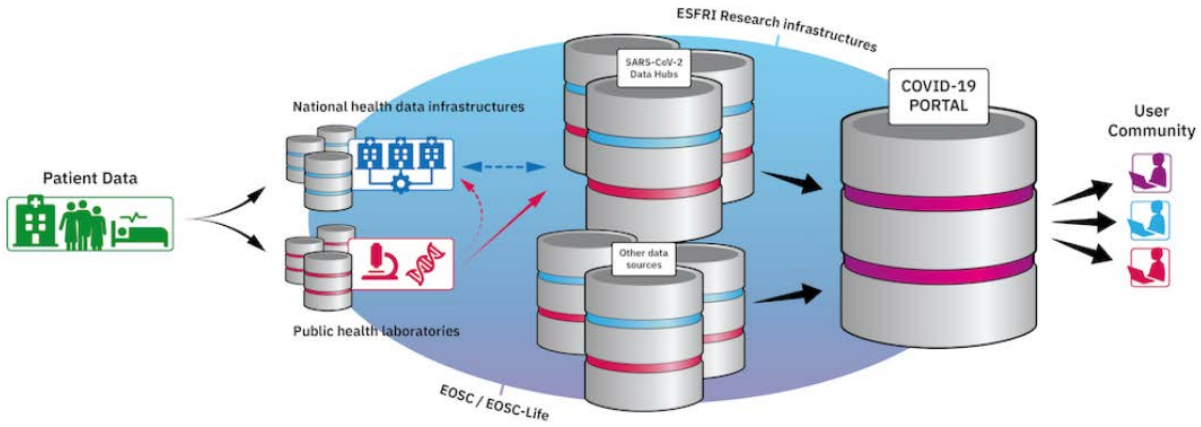
- ✓ **Omics data** for the characterisation and quantification of biological molecules (including sequence data on both virus genomes and human genomes) and other high-dimensional data such as microbiome data;
- ✓ Data from **pre-clinical research** to **test drug candidates**, **vaccine interventions**, or **other treatments**, for **efficacy, toxicity and pharmacokinetic information**;
- ✓ Research data from **clinical trials** and from **observational studies**;
- ✓ **Epidemiological data**, models, codes and algorithms.

It will also provide systems for data exploration and visualisation and a cloud compute facility where scientists and public health workers can collaborate.

The European COVID-19 Research Data Platform consists of two connected components. The COVID-19 Portal will be the main interface for the researchers, bringing together and continuously updating relevant COVID-19 datasets and tools. The SARS-CoV-2 Data Hubs will organise the flow of research data from the outbreak and feed the COVID-19 Portal. Essential metadata will be captured from the various Data Hub which will differ to reflect national and regional efforts and requirements:

- The COVID-19 Portal will be the main interface for the researchers, bringing together and continuously updating relevant COVID-19 datasets and tools. In a first stage, the COVID-19 Portal will feature relevant datasets from EMBL-EBI data resources such as the European Nucleotide Archive (ENA), UniProt, Protein Data Bank in Europe (PDBe), the Electron Microscopy Data Bank (EMDB), Expression Atlas, and Europe PMC. The portal will also include the outbreak sequence data and a Cohort Browser for searching clinical and epidemiological data (including by means of a metadata catalogue). It will also enable scientists to upload, search, and explore specialist datasets. In a second stage, additional datasets and tools from other European projects / existing platforms will be accessible with the long-term objective of including data from other international projects and European research infrstructures. This will be achieved with the help of the Commission, the EOSC governance and ELIXIR, the intergovernmental organisation that brings together life science data and resources from across Europe, and other collaborators.
- The SARS-CoV-2 Data Hubs will organise the flow of research data from the outbreak and provide comprehensive open data sharing for the research communities, starting with genomics data and expanding toother types of data. It will build on the EMBL-EBI infrastructure and will mainly be used by scientists (and public health agencies) responsible for generating viral sequences, microbiome data, data on host genetics and immune response or epidemiological modelling at national or regional levels. The research data in each Data

Hub will differ to reflect national and regional efforts and requirements. Essential metadata will be captured, including sampling time, method, geographical location, sequencing technology, and the health status of the host. The Data Hubs will also provide systematic data processing, visualisation, and phylogenetic analysis tools.



Source: EMBL-EBI[1]

[1] www.embl.org/news/science/embl-ebi-leads-international-collaboration-to-share-covid-19-research-data/

Annex 2: European COVID-19 Research Data Platform data submission routes

This document (prepared by EMBL-EBI) lays out the optimal routing of scientific COVID-19 data of different types from those who generate the data into the *European COVID-19 Research Data Platform*.

The European COVID-19 Data Platform provides two components. The web component and main entry point into the Platform, the *COVID-19 Portal*, is not itself a deposition database; rather, it provides an organised and integrated view of data in existing deposition databases. The sequence data management component of the Platform, the *SARS-CoV-2 Data Hubs*, again builds on an existing data deposition database, the European Nucleotide Archive, but provides users with specific supported interfaces to manage, process, share and navigate their data. As such, outside SARS-CoV-2 sequences, data are routed to the Platform through data deposition processes offered and supported by the respective underlying data repositories while SARS-CoV-2 data are provided with dedicated processed as part of the SARS-CoV-2 Data Hubs.

All data routing towards the European COVID-19 Data Platform follows a "push" model, in which those generating data use web interfaces or online programmatic services to submit their data sets. Data sets typically include not only data files, such as raw sequence data files or 3-dimensional protein structures, but contextual and other metadata that serve to structure and describe the data, such as details of the biological materials that have been assayed, the nature of the experiment that has been performed and links to the scientific literature. The deposition process, then, provides not only for data publication but also to prepare data sets for integration, curation and connection to the overall scientific narrative, with end goal to support discoverability, comparability and thus reusability of data, in line with FAIR principles.

Which deposition database is appropriate for a given data set depends on a number of factors, including the material scope (species or compound) of the assay and the sensitivity of the data (availability for open versus controlled access). In due course, EMBL-EBI will provide a decision tree tool to help those with data find the appropriate deposition route. (This tool will be similar in design to the existing EMBL-EBI submission triage at https://www.ebi.ac.uk/submission/.) In the interim, data is classed as follows:

1. Viral, non-human and cell line sequence data
2. Human molecular biology data
3. Linked viral and human molecular biology data
4. Viral and non-human proteomics data
5. Structural biology data
6. Viral and non-human molecular interaction data
7. Viral and non-human metabolomics data
8. Viral and other non-human molecular biology data
9. Compound and target data
10. Clinical and epidemiological data
11. Non-biological data

In the paragraphs that follow, we outline these classes and point to appropriate deposition databases.

1. Viral, non-human and cell line sequence data

This class includes sequence data from studies targeting virus alone or with co-occurring species. It also includes sequencing from non-human host species (such as from species acting as models for infection) and human cell lines (where data are consented for full open publication). All sequencing library types, all platforms, all library methods and all levels of processing (from raw data to assembled sequences) are included in this class.

Deposition actions:

- Users should submit data to the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home)
- Specific deposition instructions are available for viral data submission (https://ena-browser-docs.readthedocs.io/en/latest/help_and_guides/sars-cov-2-submissions.html)
- Users are encouraged to contact ENA at virus-dataflow@ebi.ac.uk

General depositions and those from users who are managing their data in SARS-CoV-2 Data Hubs are also included in this class.

2. Human molecular biology data

This class covers all sequence and all other molecular biology data types of human source where the data are potentially identifying of the research subject and must therefore be managed under controlled access.

Deposition actions:

- Users should deposit data into the European Genome-phenome Archive (EGA; https://ega-archive.org/)
- Users should contact their national nodes of EGA or other national infrastructure for the management of these data

3. Linked viral and human molecular biology data

This case covers data from studies that consider in combination SARS-CoV-2 and the human research subject that is infected with it. Studies include combined sequencing of host transcriptome and viral genotype; immunome and viral gene expression; pulmonary proteomics and targeted sequence-based typing of viral genes.

Deposition actions:

- Users should register data first into the BioSamples Database (https://www.ebi.ac.uk/biosamples/)
- Users are encouraged to contact EMBL-EBI at virus-dataflow@ebi.ac.uk

4. Viral and non-human proteomics data

Data in this class are derived from spectroscopy-based analyses of occurrence and abundance profiles for proteins in non-human samples, including in vitro viral culture systems.

Deposition actions:

- Users should deposit data into the PRoteomics IDentification Database (PRIDE; https://www.ebi.ac.uk/pride/archive/

5. Structural biology data

This class covers data from methods that elucidate the 3-dimensional structure of proteins and other biological macromolecules in isolation, or in their biological context (such as with bound ligands).

Deposition actions:

- Users should deposit x-ray crystallography and nuclear magnetic resonance data to the Protein Data Bank in Europe (PDBe; https://www.ebi.ac.uk/pdbe/)
- Users should deposit electron microscopy and tomography data at the Electron Microscopy Public Image Archive (EMPIAR; https://www.ebi.ac.uk/pdbe/emdb/empiar/)

6. Viral and non-human molecular interaction data

In this class are data from studies that seek to elucidate the interactions between macromolecules at binary, complex and network scales.

Deposition actions:

- Users should deposit molecular interactions data to IntAct (https://www.ebi.ac.uk/intact/)

7. Viral and non-human metabolomics data

This class comprises spectroscopy-based profiling of metabolites in non-human tissues, such as virus-infected model species and cell lines.

Deposition actions:

- Users should deposit molecular interactions data to IntAct (https://www.ebi.ac.uk/intact/)

8. Viral and other non-human molecular biology data

Non-human molecular biological data of relevance to COVID-19 research for which a structured data deposition database does not exist lie in this class, such as qPCR or viral virulence assay data.

Deposition actions:

- Users should deposit these data to BioStudies (https://www.ebi.ac.uk/biostudies/)

9. Compound and target data

Studies with focus on compounds and the investigation of their biological activities with respect to viral and other drug targets lie in this class.

Deposition actions:

- Users should deposit these data to ChEMBL (https://www.ebi.ac.uk/chembl/)

10. Clinical and epidemiological data

Data in this class cover research subject data captured during clinical practice or study relating to cohorts, study groups and case studies and include such data types as serology profiles, treatment histories and disease classifications. These data are typically retained within national health data systems and not shared centrally, but are ideally linked to other centrally managed data.

Deposition actions:

- Users should provide top-level cohort and study group data to virus-dataflow@ebi.ac.uk
- Data should be provided to national health data systems and, where possible, linked to/from data accessible from the COVID-19 Portal
- Where possible, data can be managed within EGA (https://ega-archive.org/)

11. Non-biological data

Non-biological data of relevance to COVID-19, such as travel, trade, meteorology and social distancing behaviour are not managed within the European COVID-19 Data Platform, but where possible are linked to data within the system.

Deposition actions:

- Users should encourage data standards compliance in these non-biological data sets, such as relates to the World Geospatial Consortium and FAIR

Annex 3: SARS-CoV-2 sequence data standards in the European Nucleotide Archive (ENA)

This annex (prepared by EMBL-EBI) provides an outline of data standards applied for submission and publication of data in the European Nucleotide Archive (ENA). The standards have been implemented in ENA tools and services since the beginning of the COVID-19 outbreak and reflect work that has continued for many years between the team at EMBL-EBI and many collaborators from the infectious disease and sequencing data communities. As an implementer of these community-driven standards, ENA continues to work with such organisations as the Genomics Standards Consortium (microbial genomics), the Global Microbial Identifier and the COMPARE Consortium (pathogen surveillance), the International Nucleotide Sequence Database Collaboration (sequence data formats) and the Global Alliance for Genomics and Health (raw data formats and compression).

The data standards comprise three parts: **sample data**, **raw sequence data** and **consensus/assembled sequence data**.

- Sample data
    - Sample records serve to enumerate and describe physical entities created during sampling processes. Sampling processes of relevance to SARS-CoV-2 include clinical procedures (e.g. throat swab, nasopharyngeal aspirate) and environmental testing (e.g. hospital surface testing). Sample records capture sample attributes in four classes: mandatory, recommended, optional and user-defined. The first class of (mandatory) attributes are required for acceptance in ENA and comprise the "minimal standard" while submission tools suggest and enable reporting of attributes in the remaining classes.

      To date, for the most part, SARS-CoV-2 sample record submission have been captured under the "ENA prokaryotic pathogen minimal sample checklist" which defines 7 mandatory fields (scientific name, isolation source, collection date, geographic location, host health state and host scientific name), 3 recommended fields, 11 optional fields and unlimited user-defined fields. Full details can be found at https://www.ebi.ac.uk/ena/browser/view/ERC000028.

      Given our intensified effort to mobilise data and the additional user support that we are able to provide, we are now promoting for SARS-CoV-2 submissions the more appropriate "ENA virus pathogen reporting standard checklist" which defines 10 mandatory fields (scientific name, geographic location, host common name, host subject ID, host health state, host sex, host scientific name, collector name, collecting institution and isolate name), 15 recommended fields and a further 12 optional fields, with unlimited user-defined fields. Full details can be found at https://www.ebi.ac.uk/ena/browser/view/ERC000033.

- Raw sequence data
    - Raw sequence reads are captured in a variety of established data formats, including native (such as Oxford Nanopore Technology's FAST5), and community formats, such as CRAM, BAM and FASTQ. Specifications for these formats are provided at https://ena-docs.readthedocs.io/en/latest/submit/fileprep/reads.html. Raw reads must be accompanied by mandatory library-related metadata fields: instrument platform, library source, library selection and library strategy and further fields are encouraged, such as links to external resources providing library protocols. Full metadata requirements, including dictionaries for metadata fields are provided at https://ena-docs.readthedocs.io/en/latest/submit/reads/webin-cli.html.

- Consensus and assembled sequence data
    - Consensus and assembled sequence data requirements are described at https://ena-docs.readthedocs.io/en/latest/submit/sequence.html and https://ena-docs.readthedocs.io/en/latest/submit/assembly.html, respectively. Existing community data formats must be provided (such as FASTA and AGP; see documentation referenced previously) and annotation, where provided, is captured according to the INSDC Feature Table Definitions (http://www.insdc.org/files/feature_table.html).