# RDA COVID-19 Working Group

## Recommendations and Guidelines

## 3rd release

## 8 May 2020

**RDA Recommendation (2nd Release - 8 May 2020)**
**Produced by: RDA COVID-19 Working Group, 2020**

# Document Metadata

| | |
|---|---|
| *Identifier* | DOI: 10.15497/RDA00046 |
| *Title* | RDA COVID-19; recommendations and guidelines, 3rd release 8 May 2020 |
| *Description* | This is the third draft of the “overarching” RDA COVID-19 Guidelines document, and is intended to provide an update on the progress of the WG, as well as a focus on high-level recommendations that run across all 5 sub-groups in this initial effort, as well as the cross-cutting themes of Research Software andLegal and Ethical . |
| *Date Issued* | 2020-05-08 |
| *Version* | Draft guidelines and recommendations; third release, 8 May 2020, version for public review |
| *Contributors* | RDA COVID-19 Working Group<br>This work was developed as part of the Research Data Alliance (RDA) ‘WG’ entitled ‘RDA-COVID19,’ ‘RDA-COVID19-Clinical,’ ‘RDA-COVID19-Community-participation,' ‘RDA-COVID19-Epidemiology,’ ‘RDA-COVID19-Legal-Ethical,’ ‘RDA-COVID19-Omics,’ ‘RDA-COVID19-Social-Sciences,’ ‘RDA-COVID19-Software,’ and we acknowledge the support provided by the RDA community and structure. |
| *Licence* | This work is licensed under CC0 1.0 Universal (CC0 1.0) Public Domain Dedication |

***Group Co-chairs:*** *Juan Bicarregui, Anne Cambon-Thomsen, Ingrid Dillo, Natalie Harrower, Sarah Jones, Mark Leggott, Priyanka Pillai*

***Subgroup Moderators:***
*Clinical: Sergio Bonini, Dawei Lin, Andrea Jackson-Dipina, Christian Ohmann*
*Community Participation: Timea Biro, Kheeran Dharmawardena, Eva Méndez, Daniel Mietchen, Susanna Sansone, Joanne Stocks*
*Epidemiology: Claire Austin, Gabriel Turinici*
*Legal and Ethical: Alexander Bernier, John Brian Pickering*
*Omics: Natalie Meyers, Rob Hooft*
*Social Sciences: Iryna Kuchma, Amy Pienta*
*Software: Michelle Barker, Hugh Shanahan, Fotis Psomopoulos*

***Editorial team:*** *Christoph Bahim, Alexandre Beaufays, Ingrid Dillo, Natalie Harrower, Mark Leggott, Nicolas Loozen, Priyanka Pillai, Mary Uhlmansiek, Meghan Underwood, Bridget Walker*

History of the discussions in the Working Groups that led to this document can be viewed in the comments made in the associated subgroup documents.

## Table of Contents

# 0. Log Changes

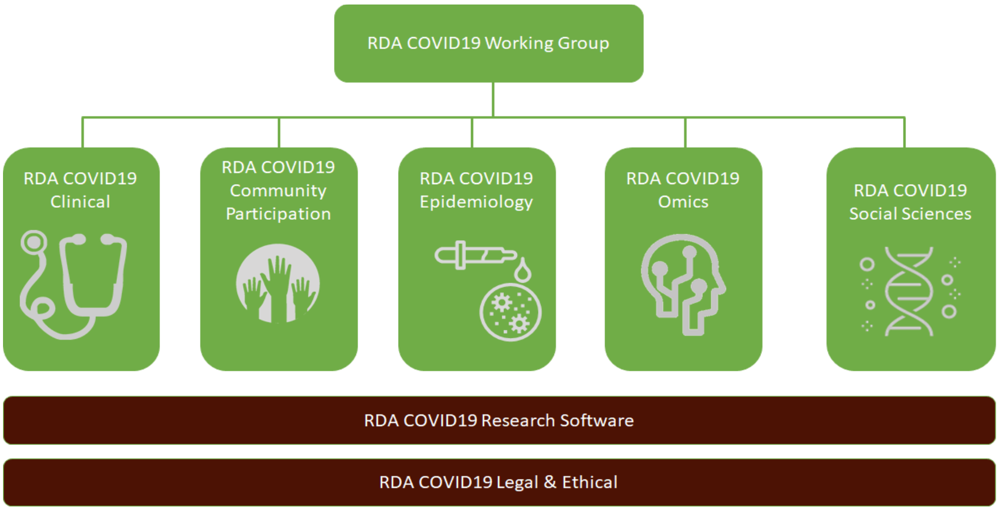
| Document | Changes | Date |
|---|---|---|
| RDA COVID-19; recommendations and guidelines, 1st release 24 April 2020 | First draft of document released for comments and feedback | 24 April 2020 |
| RDA COVID-19; recommendations and guidelines, 2nd release 1 May 2020 | Section 3 - Foundational Principles/Recommendations - modifications<br>Section 4 - Clinical - updated<br>Section 6 - Epidemiology - revised<br>Section 8 - Omics - revised<br>Section 9 - Overarching Research Software Guidelines - new sub-section 9.3 initial guidelines for policy makers included<br>Section 10 - Overarching Legal and Ethical Guidelines added<br>Incorporation of feedback received via Requests for Comments (RfC) process and directly to Co-chairs, moderators and editorial team | 1 May 2020 |
| RDA COVID-19; recommendations and guidelines, 3rd release 8 May 2020 | Section 3 - Foundational Principles/Recommendations - updates and additional content<br>Section 4 - Clinical - Revisions to sections 4.2.3, 4.2.4, and 4.3. Newly added section 4.2.5<br>Section 5 - Community Participation: Use case section 5.2.2 added<br>Section 6 - Epidemiology - major revisions/additional content<br>Section 7 - Omics - updated<br>Section 8 - Social Sciences - revised<br>Section 9 - Software - updated and sections added/reorganized (9.4, 9.5)<br>Section 10 - Legal/Ethical - major revisions<br>Incorporation of feedback received via Requests for Comments (RfC) process and directly to Co-chairs, moderators and editorial team | 8 May 2020 |

# 1. Function of the RDA COVID-19 WG

During a pandemic, data combined with the right context and meaning can be transformed into knowledge for informing public health response. Timely and accurate collection, reporting and sharing of data with the research community, public health practitioners, clinicians and policy makers will inform assessment of the likely impact of a pandemic to implement efficient and effective response strategies.

Public health emergencies clearly demonstrate the challenges associated with rapid collection, sharing and dissemination of data and research findings to inform response. There is global capacity to implement systems to share data during a pandemic, yet the timeliness of accessing data and harmonisation across information systems are currently major roadblocks. The World Health Organisation's (WHO) statement on data sharing during public health emergencies clearly summarises the need for timely sharing of preliminary results and research data. There is also a strong support for recognising open research data as a key component of pandemic preparedness and response, evidenced by the 117 cross-sectoral signatories to the Wellcome Trust statement on 31st January 2020, and the further agreement by 30 leading publishers on immediate open access to COVID-19 publications and underlying data.

The objectives of the RDA COVID-19 Working Group (CWG) are:

1. to clearly define detailed guidelines on data sharing under the present COVID-19 circumstances to help stakeholders follow best practices to maximize the efficiency of their work, and to act as a blueprint for future emergencies;
2. to develop guidelines for policymakers to maximise timely, quality data sharing and appropriate responses in such health emergencies;
3. to address the interests of researchers, policy makers, funders, publishers, and providers of data sharing infrastructures.
4. 

RDA COVID19 Working Group
RDA COVID19 Clinical
RDA COVID19 Community Participation
RDA COVID19 Epidemiology
RDA COVID19 Omics
RDA COVID19 Social Sciences
RDA COVID19 Research Software
RDA COVID19 Legal & Ethical

5. Figure 1. Research Data Alliance COVID-19 working group including thematic and cross themes.

The CWG is addressing the development of such detailed guidelines on the deposit of different data sources in any common data hub or platform. The guidelines aim 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, and principles that can be embedded in research practice. The guidelines to be developed will address general aspects related to the principles that data should adhere to, for example FAIR and others while providing a tool which could serve as the standard of *good enough to decide*. The work has been divided into 5 thematic areas with two cross cutting themes, as a way to focus the conversations, and provide an initial set of guidelines in a tight timeframe.

# 2. Status of the RDA COVID-19 WG Effort

The RDA COVID-19 WG was initiated after a conversation between the RDA Secretary General and European Commission contacts. The first meeting to determine the work was held on March 20th, and included a number of RDA stakeholders. Subsequent to this, the Secretary General reached out to colleagues in the RDA community to act as Co-Chairs, and the first meeting of this group was held on March 30th. The next step was to invite a group of Moderators to facilitate the discussion of the 5 sub-groups, and the first group meetings started taking place soon after.

As of May 8, there are over 440 members of the CWG, relatively evenly spread across the 7 groups. The first two drafts were released respectively on April 24th and May 1st indicating huge progress in the space of 4 weeks, and were opened for comment. The various sub-group drafts collected here reflect different approaches and initial work efforts: future drafts (4th and 5th releases) will synchronize these efforts into a more integrated series of guidelines. The current timeline will produce a new release next Friday May 15, with the subsequent release anticipated for May 28, followed by a round of community feedback, and a final endorsed release of Version 1 of the Guidelines for June 30. The thematic guidelines will be enhanced with a series of "cross-cutting" guidelines, that articulate principles and recommendations common across all areas. Two cross cutting guidelines are included here: Research Software and Legal and Ethical. This effort also reflects the work of a host of other RDA Working Groups, as well as external stakeholder organizations, that has developed over a number of years - we want to recognize and highlight those efforts.

In the spirit of the RDA community and its open process, we are seeking feedback from the COVID-19 WG members, as well as the broader community, early and often during this process. This feedback will inform our work and will be incorporated into the sub-group discussions, and the next set of writing sprints.

*This Working Group and the subgroups operate according to the RDA guiding principles of Openness, Consensus, Balance, Harmonization, Community-driven, Non-profit and technology-neutral and are OPEN TO ALL.*

# 3. Foundational Principles/Recommendations

The thematic sub-groups have each articulated challenges facing researchers working on COVID-19, as well as recommendations/guidelines for improving data sharing; these subgroup guidelines should be considered directly depending on the relevant area of COVID-19 research. However, certain foundational aspects appear across these subgroups, so we present these here as foundational elements that apply across all themes.

## 3.1 Challenges

### 3.1.1 Rapid Pace of Research Under the Pandemic: Speed vs. Accuracy

The unprecedented spread of the virus has prompted a rapid and massive research response, but to make the most of global research efforts, findings and data need to be shared equally rapidly, in a way that is useful and comprehensible. Raw data, algorithms, workflows, models, software and so on are required inputs to research studies, and are essential to the scientific discovery process itself. New findings and understandings need to be disseminated and built upon at a pace that is faster than usual, because decisions are being taken by healthcare practitioners and governments on a daily basis, and it is urgent that they are well-informed.

The inaccuracy and/ low quality of data shared within such short timelines, could have considerable implications, for example:

1) shortcuts with the interpretation of data can create issues, such as the debate on whether COVID-19 is 'just another flu' or not;
2) obligation to share data could orient at least some institutions to reduce testing (suspected cases do not, count, only confirmed ones do, and hence lowering testing allows lowering confirmed case numbers and creates the illusion that the epidemic is under control);
3) in some cases, lack of transparency and publication of false numbers is perhaps worse than no publication at all.

### 3.1.2 Critical Need for Data Sharing

The COVID-19 pandemic has revealed how interconnected we are globally, and how interdependent we are in terms of research, public health, and economy. Data in relation to this pandemic is being collected and created at a high velocity, and it is critical that we can share this data across cultural, sectorial, jurisdictional, and disciplinary boundaries.

The challenge here is the trade off between timeliness and precision. The speed of data collection and sharing needs to be balanced with accuracy, which takes time. The pressure to interpret results, turn around studies quickly and update statistics in almost real-time must not compromise quality and reliability. There is no overarching formula for finding that balance, but documented transparency in the research process and decisions taken can help to mitigate the dangers associated with working at hyperspeed.

### 3.1.3 Lack of Coordinated Standards and Context

Emerging infections are largely unpredictable in nature and there is limited data to support disease investigation. The evidence base generated from early outbreak data is critical to inform rapid response during an emerging pandemic. Lack of pre-approved data sharing agreements and archaic information systems hinder rapid detection of emerging threats and development of an evidence-based response.

While the research and data are abundant, multi-faceted, and globally produced, there is no universally adopted system, or standard, for collecting, documenting, and disseminating COVID-19 research outputs, and many outputs are not reusable by, or useful to different communities, if they have not been sufficiently documented and contextualised, or appropriately licensed. There is an urgent need for data to be shared with minimal contextual information and harmonised metadata so that it can be reused and built upon (see the OECD Open Science Policy Brief).

## 3.2 Recommendations

### 3.2.1 FAIR and Timely

The consensus in this series of guidelines is that research outputs should align with the FAIR principles, meaning that data, software, models and other outputs should be Findable, Accessible, Interoperable and Reusable. However, there is also consensus that outputs need to be shared as quickly as possible in order to have a direct impact on the progress of the pandemic. A balance between achieving 'perfectly' FAIR outputs and timely sharing is necessary with the key goal of immediate and open sharing as a driver. Researchers should be paired with data stewards to facilitate FAIR sharing, and data management should be considered at the start of a study or trial. Immediate open access with open licenses is desirable, but some effort should be put into the quality and documentation of the dataset.

### 3.2.2 Metadata

The key to finding and using digital assets is metadata. COVID-19 research requires access to different assets for different communities. Within a given community, the commonly used metadata standards are well-known, but a researcher working across communities has more difficulty in locating relevant assets. In this case a 'metadata element set' that is generally applicable is required to be associated with each asset so that they can be used under the FAIR principles. A proposed metadata element set is available on the RDA Metadata Interest Group page. At present there are four generic metadata standards that are used widely, Dublin Core (DC), DCAT, DataCite and Schema.org. The latter has a specialisation Bioschemas likely to be particularly useful for COVID-19 research especially with the updated profiles. Providing FAIR access to assets would be much enhanced now if assets had metadata encoded in one of these standards – as well as in the metadata standard(s) used by the particular community. It is to be hoped that in future richer generic metadata standards will be used. For a longer registry of metadata standards, see the Metadata Standards Directory.

However, the use of these standards for machine-to-machine communication depends on how they are implemented. Many DC implementations are in text, HTML or XML form and used more easily by human readability than machine understandability. More recent implementations use Resource Description Framework (RDF) which does provide machine-to-machine capability. Earlier DCAT implementations used XML, more recent implementations use RDF. DataCite uses XML but also schema.org metadata format and JSON-LD while Schema.org uses RDF and JSON-LD. Thus, these metadata standards encourage machine-to-machine interoperation.

Metadata has two aspects: syntax and semantics. The syntax defines the structure of the metadata information and should conform to a formal grammar. The semantics defines the meaning of strings of characters – usually through an associated ontology – and should be declared.  Again, there are generic ontologies (or vocabularies which have less detail on relationships between the terms) and community-specific ontologies (or vocabularies).

### 3.2.3 Documentation

Research outputs need to be documented, which includes documentation of methodologies used to define and construct data, data cleaning, data imputation, data provenance and so on. Software should provide documentation that describes at least the libraries, algorithms, assumptions and parameters used. Equally, research context, methods used to collect data, and quality-assurance steps taken are important. When sharing datasets, other relevant outputs (or documents) should also be made available, such as codebooks, lab journals, or informed consent form templates, so that data can be understood and potentially linked with other data sources. The recent joint statement on the Duty to Document underlines how crucial it is, especially during this time of rapid and unprecedented decision making, to document decisions, and secure and preserve records and data for the future.

### 3.2.4 Use of Trustworthy Repositories

To facilitate data quality control, timely sharing and sustained access, data should be deposited in data repositories. Whenever possible, these should be trustworthy data repositories (TDRs) that have been certified, subject to rigorous governance, and committed to longer-term preservation of their data holdings. As the first choice, widely used disciplinary repositories are recommended for maximum accessibility and assessability of the data, followed by general or institutional repositories. Using existing open repositories is better than starting new resources. By providing persistent identifiers, demanding preferred formats, rich metadata, etc., certified trustworthy repositories already guarantee a baseline FAIRness of and sustained access to the data, as well as citation. In general you can consult re3data.org for a searchable database of research data repositories. Repositories certified by CoreTrustSeal, a result OF the RDA Repository Audit and Certification DSA–WDS Partnership WG and the European Commission's expert group on FAIR data are listed here.

### 3.2.5 Ethics & Privacy

The ethical and privacy considerations around participant and patient data are significant in this crisis, and several guidelines note the need to find a balance that takes into account individual, community and societal interests and benefits whilst addressing public health concerns and objectives. Access to individual participant data and trial documents should be as open as possible and as closed as necessary, to protect participant privacy and reduce the risk of data misuse. While the privacy protection and anonymisation challenges are substantial (ie. as evidenced with current discussions about contact tracing) solutions which allow algorithms to 'visit' data, asking specific research questions which can be answered while not allowing direct access to data, should be considered.

### 3.2.6 Legal

Technical solutions that ensure anonymisation, encryption, privacy protection, and data de-identification will increase trust in data sharing. The implementation of legal frameworks that promote sharing of surveillance data across jurisdictions and sectors would be a key strategy to address legal challenges. Emergency data legislation activated during a pandemic needs to clearly outline data custodianship/ownership, publication rights and arrangements, consent models, and permissions around sharing data and exemptions.

# 4. Clinical Sub-Group Guidelines

## 4.1 Sub-Group Focus and Description

Clinical activities are at the forefront to combat the COVID-19 pandemic. Although many aspects of such actions were considered in the scope of the sub-group, the work of the Clinical Subgroup centers first on dealing with consent on data sharing, how clinical trials are conducted, how clinical information (personal and health data) and results are shared and consumed in a trustworthy and efficient manner. For this 3rd release, recommendations for consent and clinical trials have been slightly modified, recommendations for clinical data sharing have been deeply revised and some elements for immunological data and imaging data have been added and will be further supplemented.

## 4.2 Initial Sub-Group Guidelines

### 4.2.1 Consent on COVID-19

1. Procedures on data sharing specific for COVID-19 in the general consent for clinical trials
2. should be in accordance with ISO/TS 17975:2015 (Health informatics — Principles and data requirements for consent in the Collection, Use or Disclosure of personal health information)[1]
3. The access to sensitive person-related data should be covered by Data Access Agreements (DAAs) between the data holder and the data user, ideally managed by Data Access Committees
4. For genomic and health-related data the data sharing should follow recommendations of the GA4GH Consent Policy
5. More information on the ethical and legal bases will be found in the chapter from the legal and ethical sub-group.

### 4.2.2 Clinical trials on COVID-19

Clinical trials are an important research area to discover and make available safe and effective treatments for COVID-19[2] . International, regional, and national legal and methodological frameworks exist for clinical trials[3], that also take into account ethical and legal principles. Specific recommendations on registering, performing, and sharing ongoing clinical research are the following:

1. Lawful fast track approval procedures of clinical trials in cases of public health emergencies exist that speed up processes while protecting adequately individual rights. Platforms that point to them in the various national and international institutions should be further developed and administrations should apply them diligently and transparently.

2. Clinical trials in COVID-19 should be registered at or before the time of first patient enrollment and protocols possibly published in order to favor harmonization of studies, collaboration among centres as well as to avoid duplication of efforts
3. Multi-centres multi-countries studies including a sample size calculation according to the primary objective should be recommended to generate sound evidence on COVID-19 treatments. Collaborative trials and multi-arms studies comparing different interventions are advisable.
4. Heterogeneity between registries regarding the number of studies listed and the information available for individual studies should be overcome through a dialogue among different platforms
5. Protocols should follow standard criteria for data collection, stratification of the randomized population, type of intervention and comparator, a minimal set of primary outcome measures (e.g. SPIRIT: Standard Protocol Items: Recommendations for Interventional Trials) and adhere to FAIR data principles.
6. When regulatory bodies allow compassionate use of approved repurposed drugs such a use should be reported; if a fast track for approval of proved COVID-19 drugs exists it is also useful to report it. Adaptive study designs and post-authorization efficacy and safety studies after exceptional or conditional approval should be planned with sponsors in order to favor early access of severe patients to promising medicines.

## 4.2.3 Sharing Clinical Data

### *General aspects*

In the COVID-19 situation promotion of data sharing is of utmost importance because many studies are performed under enormous time pressure, with weaknesses in the methodology (e.g. no control) and preliminary results published without any review. Sharing of data, and related documentation (e.g. protocols) will reduce duplication of effort and improve trial design, when many similar studies are being planned or implemented in different countries. Clinical data outside clinical trials (case studies, descriptive cohorts of patients) may also be of high value and should be reported using appropriate reporting guidelines (see EQUATOR Network guidelines). Relevant principles and recommendations are especially based on a re-consideration of Ohmann et al. [9]

#### *Publications and other formats*

1. Availability for *timely publication* of results - even for negative and withdrawn studies - and for data sharing should be declared by investigators and sponsors at the time of study registration and included in the study documents (e.g. protocol, patient information and consent form). In the COVID-19 crisis, publication cannot be the criterion for data sharing. Data sharing should be performed as soon as the study is completed[10].
2. *Preprint publishing* and other forms of knowledge sharing and exchange are also encouraged. *Full reports* should be made available immediately upon communication of results, e.g. through a press release.
3. Where possible, *open source journals* and adherence to OpenAire initiatives and the likes are also encouraged.

There is a danger that this is not adequately being done in the current situation and solutions have to be found to assure data sharing even under these circumstances.

### *Credit and attribution*

1. In a situation where there is a strong need for data sharing, at an early stage and before primary papers might be written, having *credit for the data* becomes more important. Initiatives to support rewards and credits for data sharing should be strengthened (RDA: Sharing rewards and credit (SHARC) IG, https://www.rd-alliance.org/groups/sharing-rewards-and-credit-sharc-ig, FORCE11: Joint declaration of data citation principles, https://www.force11.org/datacitationprinciples).
2. *Persistent Identifiers for data sources* (e.g. DOI) should be included in a secondary analysis to recognize primary data providers.
3. *Financial models to support data sharing* for COVID-19 studies should be implemented and funding specifically targeted on such activities should be provided. This could include additional costs for preparing data sharing as well as making data as inter-operable as possible.

### *Biological samples as data sources*

1. In the context of a pandemic the *access to biological samples that are data sources* might be of high interest and policies should be in place for facilitating their access; they should be developed in full respect of legal and safety regulations, protection of patients and with recognition of the value of the work performed to constitute such collections with relevant metadata.
2. Main principles are delineated in the Access policy of BBMRI-ERIC, the European research infrastructure consortium for biobanking and biomolecular resources. https://www.bbmri-eric.eu/services/access-policies/

### *Consent for data sharing*

1. Data and trial documents should be made available for sharing
2. Individual participant data sharing should be based on broad consent by trial participants (or if applicable by their legal representatives) to the sharing and reuse of their data for scientific purposes, according to applicable law.
3. Where real-world data are collected from patient registries or similar data sources not involving specific consent to participate, patients' privacy must be adequately protected[4].
4. Procedures on data sharing specific for COVID-19 in the informed consent for clinical trials should be in accordance with standards and recommendations (e.g. ISO/TS 17975:2015 (Health informatics — Principles and data requirements for consent in the Collection, Use or Disclosure of personal health information, https://www.iso.org/standard/61186.html; GA4GH Framework for responsible sharing of genomic and health-related data ("Framework"), https://www.ga4gh.org/wp-content/uploads/GA4GH-Final-Revised-Consent-Policy_16Sept2019.pdf)[1]

### *Protection of trial participants*

1. Because of the pressure to publish and then make data available, there may be a greater risk of data not being properly de-identified (anonymised) for data sharing. For this reason, the importance of measures to protect the data is paramount (e.g. specific data use agreements). It should be clarified whether for public health emergency situations specific legislation for data sharing is in force and simplified approaches could be used. This information should be available centrally on a web page.

### *Data standards*

1. Data standards should be applied in COVID-19 studies. Among the *various data standards* available, those from CDISC should be considered as offering an appropriate starting point currently available for defining and coding data and metadata in a consistent way. But there are alternative standards to CDISC for academic teams. Ontologies and thesauri for standardising terms and formats should be used.
2. More *support is likely to be needed for academic researchers to apply these standards* (a 'simplified CDISC' for COVID-19 may be useful).
3. In the current situation standards related to data sharing should be made *accessible without licensing fees*. Openness should become the rule in pandemic situations.

### *Rights, types and management of access*

1. In order to expedite the process of data sharing, standardized agreements for sharing of data between data providers, repositories and data requestors for COVID-19 clinical trials should be developed and implemented (e.g. data transfer agreements, data access and data use agreements)
2. In the COVID-19 situation access to data should be as open as possible. This does not necessarily mean completely open access, as they also need to be as closed as necessary, but measures to control and manage risk (anonymization, data use agreements) can be used to make access as easy as possible, while adequately protective. If a Data Access Board or a similar third-party mechanism is involved in decisions about data sharing, there is a need for a transparent and fast track process.

## 4.2.4 Trustworthy Sources of Clinical Data

During a pandemic like COVID-19, it is important to concentrate efforts on scrutinizing *reliable data sources* that provide data and metadata of high quality and guarantee the authenticity and integrity of the information. The recommendations are:

1. Data and trial documents should be transferred to a suitable and secure data repository to help ensure that the data are properly prepared, are available in the longer term, are stored securely and are subject to rigorous governance. Repositories that explicitly support data sharing for COVID-19 trials should be announced (e.g. Vivli[5]).
2. Trustworthy repositories should be leveraged as a vital resource for providing access to and supporting the depositing of research data. However, as an emerging and evolving area in biomedical domains, trustworthiness assessment should not be limited to certification[7] [8]or accreditation. A wide -range of community-based standardized quality criteria and best practices should also be considered.
3. If analysis environments, allowing in situ analysis of data sets but preventing downloads or allowing different data sets from different repositories to be combined on a temporary basis, are available they should be provided to the end-user researchers, in a pandemic situation, without fees.
4. Adequate tools should be implemented for collection and analysis of reliable real-world data on drugs approved for the treatment of COVID-19.

Discoverability and metadata are important elements to optimise sharing and accelerate data use. To prepare data for sharing clinical trial data should always be associated with adequate and standardised metadata to improve discoverability ("F" in FAIR).

1. Tools should be developed to enable regular harvesting of metadata objects from clinical trials, allowing identification of trials and all related data objects (e.g. protocol, data set, a summary of results, publication, data management plan) through one portal (e.g. ECRIN: Clinical Research Metadata Repository (CRMDR), https://www.ecrin.org/clinical-research-metadata-repository).
2. Critical in the current situation is to have datasets easily findable. Resolvable persistent identifiers like DOIs, e.g. linking to a repository or network of repositories, would play a large part in making the data available.
3. For COVID-19 a variety of study designs is applied, covering interventional trials, observational studies, cohorts and registries. Metadata schemas between these study types should be aligned to improve discoverability of studies and associated data objects.

More information is at clinical trials and clinical aspects documentation.

### 4.2.5 Other Types of Data

In COVID-19 clinical presentation and evolution, diagnostic and prognostic data including immunological data, virology tests results and imaging, especially lung scan in case of respiratory distress, are important elements. All values for metadata and assay results should be defined with the use of domain specific controlled vocabularies. These data standards are recommended for the following data types:

1. Flow Cytometry (FACS) and Mass Cytometry (CyTOF) Experiments for ImmunoPhenotyping

   The primary cytometry data in .fcs format is greatly enhanced by inclusion of interpreted data (e.g. the cell population name, definition and frequency). (https://www.immport.org/docs/standards/Cytometry_Data_Standard.pdf)

   Cell population names should be the standard name from a curated reference source (e.g. Cell Ontology). Use of standardized Cell population names in flow cytometry and CyTOF experiments improves the ability to compare datasets.

   Cell population definitions are based on the biomarker expression pattern or 'gating strategy'. Biomarker names, when the biomarker is a monoclonal antibody, should use the antibody's antigen name from Protein Ontology, UniProt, or ChEBI. Cell population frequency units should be defined. Inclusion of the monoclonal antibody's clone name enhances the confidence that this crucial assay reagent is the same across datasets.
2. Chemokine and Cytokine Measurements (e.g. ELISA, Luminex xMAP, MBAA)
   Chemokine and cytokine assay methods are often based on monoclonal antibodies and findability and interoperability is facilitated by standardized naming of the antibody's antigen using Protein Ontology, UniProt, ChEBI, the antibody detector, the antibody's clone name and the vendor. Data standards and deposition guides are available (https://www.immport.org/resources/dataTemplates).
3. Neutralizing Antibody Titer
   Standardized names for viral targets using reference sources (e.g. NCBI Taxonomy) is recommended. Description of the neutralizing antibody type (e.g. IgM, IgG) and detector enhances interoperability.
4. Virus Presence and Titer
   Standardized names using reference sources (e.g. NCBI Taxonomy) for measurement of virus presence is recommended.

5. Imaging data
   Standards for medical images and interoperability protocols such as those described in [11] should be applied.

## 4.3 References

1. ISO. ISO/TS 17975:2015. ISO. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/11/61186.html. Published 2015. Accessed May 1, 2020.
2. NIH. ClinicalTrials - Listed clinical studies related to the coronavirus disease (COVID-19). U.S. National Institutes of Health - Information on Clinical Trials and Human Research Studies - National Library of Medicine. https://clinicaltrials.gov/ct2/results?cond=COVID-19. Published 2020. Accessed April 17, 2020.
3. CDISC. Interim User Guide for COVID-19. CDISC. https://www.cdisc.org/standards/therapeutic-areas/covid-19. Published 2020. Accessed May 3, 2020.
4. FAIR4HealthConsortium. FAIR4Health at RDA Germany Conference 2020 - Resources. FAIR4Health - Resources. https://www.fair4health.eu/en/resources. Published 2020. Accessed April 15, 2020.
5. VIVLI. Center for Global Clinical Research Data. https://vivli.org/. Published 2020. Accessed April 24, 2020.
6. European Commission. 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. April 2020. https://www.rd-alliance.org/system/files/documents/H2020_Guidelines_COVID19_EC.pdf. Accessed April 17, 2020.
7. Audit and Certification of Trustworthy Digital Repositories. 2011:77.
8. CoreTrustSeal Standards and Certification Board. CoreTrustSeal Trustworthy Data Repositories Requirements: Extended Guidance 2020–2022. November 2019. doi:10.5281/zenodo.3632533
9. Ohmann et al., BMJ Open, 2017; https://bmjopen.bmj.com/content/7/12/e018647
10. Birney et al. Prepublication data sharing. Nature. 2009 Sep 10;461(7261):168-70. doi: 10.1038/461168a.
11. Persons, K.R., Nagels, J., Carr, C. et al. Interoperability and Considerations for Standards-Based Exchange of Medical Images: HIMSS-SIIM Collaborative White Paper. J Digit Imaging 33, 6–16 (2020). https://doi.org/10.1007/s10278-019-00294-0

# 5. Community Participation Sub-Group Guidelines

## 5.1 Sub-Group Focus and Description

**The context in which we work — data and community participation**
Public health emergencies require profound and swift action at scale with limited resources, often on the basis of incomplete information and frequently under rapidly evolving circumstances. The current COVID-19 pandemic is one such emergency, and its scale is unprecedented in living history. Worldwide, many communities are coming together to address the emergency in a plethora of ways, many of which involve data in various fashions. For instance, they produce or mobilize data, add or refine metadata, assess data quality, merge, curate, preserve and combine datasets, analyze, visualize and use the data to develop maps, automated tools and dashboards, implement good practices, share workflows, or simply engage in a range of other activities that can or do leave data traces that can be leveraged by others.

While emergency-triggered sharing goes back millennia, data sharing is a relatively new aspect of emergency response, and the size, scale and complexity of the data relevant to the current pandemic are many orders of magnitude greater than even those of other recent epidemics, e.g. SARS, MERS, Zika or Ebola. This abundance of data, while in our favour in principle, can also be our Achilles heel if we - and our technology - are not able to openly share, understand and combine this data to gain the maximum insights it can provide, and to communicate those insights to the communities for which they are relevant and to the wider public.

**The aims of the RDA-COVID-19 Community Participation subgroup**
Our primary aim is to support the work of communities which are sharing data with the goal of improving research outputs and public knowledge. To achieve this, our objectives include highlighting the achievements and outputs of groups who practice sharing and to broaden access to the existing guidelines for sharing best practices. As described in "Principles of data sharing in public health emergencies" and similar publications, guidelines address issues of giving credit for contributions, legality in sharing data, technical considerations in making data Findable, Accessible, Interoperable and Reusable (FAIR), or other similar guidance for collaborating in research during a crisis.

With this objective in mind, the subgroup seeks to also take on an active role of bridging communities and ensuring inputs are streamlined, perspectives from communities are considered, and the collaborative outputs of all the RDA COVID-19 subgroups are widely communicated. The aim of linking communities and supporting communication is also designed to help coordination and avoid duplication of efforts since many communities are driving similar or complementary efforts to help the response to the current public health emergency.

These guidelines aim to facilitate the timely sharing of data relevant to the COVID-19 response and build much-needed capacity for similar events in the future. An effective and efficient response to a public health emergency, such as the current pandemic, demands and holds immense value for both public and science communication, informing opinions and understanding, whilst supporting decision-making processes.

Although these principles have been developed with research data in mind, it is also desirable that data created directly by citizens (be that in a role as citizen scientists or not), patients, communities and other actors in a health emergency be produced, curated and shared in line with the spirit of these sharing principles. For example, community projects such as OpenStreetMap and Wikidata generate very valuable FAIR and open data, which can be analysed and used along with data from professional research and other sources.

# 5.2 Initial Sub-Group Guidelines

## 5.2.1 Stakeholders

The intended audience for this subgroup's outputs includes

1. ***Researchers*** undertaking activities along the entire life-cycle of pertinent data, especially those not covered by the other RDA COVID-19 WG subgroups and involving broad-scale community participation but also data stewardship of the community-generated data.
    1.1. **Citizen scientists** undertaking research activities and in need of guidance (e.g. in terms of ethics) as well as means to seamlessly contribute to a common body of knowledge and collaborate with other actors involved.
2. ***Policymakers*** are involved in setting the framework for community participation, funding innovation, working on research policy or focusing on integrating data in decision making.
3. ***Patients***, caregivers and the communities around them that are involved in leveraging data to improve prevention, diagnostics or treatment (this complements the work of the RDA COVID-19 Clinical subgroup).
4. ***Developers*** involved in the creation or maintenance of applications targeted at community data collection that are specific to COVID-19 (e.g. contact tracing apps or exposure risk indicator apps) or more generic in nature (e.g. health or neighbourhood apps).
5. ***Device makers*** involved in developing sensors and data generating products for the community to use.
6. ***Communicators*** involved in informing communities and societies at large about data-related aspects of the COVID-19 pandemic, translating data into meaningful and easy to grasp information, and circulating graphics or key messages in conventional or social media.
7. ***Citizens and the public at large***, i.e. members of any community wanting to contribute to the COVID-19 response in ways that involve data and who want to have a say in how to balance that with legal and ethical issues surrounding such data.
8. ***Other actors (individuals or organisations)*** who are involved in community-based activities around COVID-19 related data.

## 5.2.2 Use Case: Application Development for Community-generated Data

This document is intended to look at data management and sharing issues and only reflect at the technical, social, legal and ethical considerations from that perspective.

**Who are we speaking to?**
**Stakeholders**

This document is intended to provide guidance and recommendations to the following groups of stakeholders.

1. Data subjects: Informed and forms of dynamic consent should be obtained from the data subject before personal data1 is collected from/about them and whenever there are changes to the data collection process, e.g. patients, citizens, general public.
2. Data processors/ data custodians/ data controller: determine the purposes and methods of the processing of personal data, perform the data processing, including analysis, anonymisation, storing and preservation, sharing e.g. researchers, app developer, funders, policy makers, health authority

**What do we mean by app development for community-generated data?** We are referring mostly to:

1. Symptom tracking apps (health monitoring apps where users self-report COVID-19 symptoms)
2. Contact tracing apps (mobile phone tracking used to identify potential geographic spread of COVID-19)
3. Services app (including service volunteers such as healthcare, shopping, entertainment, religious services)

**Disclaimer**: RDA does not endorse any products. Any products mentioned in this document is for illustrative purposes only, and does not constitute an endorsement by RDA. Please view the official RDA statement on this as referenced in the overall COVID-19 Guidelines and Recommendations section.

### Transparency and community participation

**Challenge**

Achieving a balance between timely contact tracing and community safety alongside individual privacy concerns such as surveillance, unauthorized use of personal data and forms of abuse that might result from the identification of subjects.

**Guidelines**:

Establish appropriate and transparent governance mechanisms to have oversight of the data and its management. An open and transparent approach allows for the community to have a say and suggest improvements e.g. Guidelines from the Ada Lovelace Institute https://www.adalovelaceinstitute.org/our-work/covid-19/covid-19- exit-through-the-app-store/

### Data collection

**Challenge:**

App developers are not always aware of all the ethical and legal implications of the data they gather and might not be familiar with protocols for collecting and sharing data.

**Guidelines:**

1. Encourage public and patient involvement (PPI) throughout the data management lifecycle from inception of the research question, implementation of the data collection and final data sharing and usage.
2. Ensure apps are developed with the research and health care question as the central concept and only gather data needed to address these questions.

3. Applications designed to collect data should be developed as open source, with early release on a public code repository and made available under an open source licence (c.f. section on Research Software in this report), to build confidence in the public about security and privacy. It also allows for the rapid identification and removal of vulnerabilities.
4. Protecting personal data is of utmost importance when developing applications. Use protocols and methods that aim to protect personal data e.g. DP-3T.

**Recommendations:**

1. Ensure developers, data stewards, healthcare professionals, epidemiologists, researchers and the public are represented in the teams driving the development of the data collecting apps.
2. Consider the use of the data - clinical, social etc. This will help identify useful standards and disciplinary norms, provide additional directions on the necessary contextual information and harmonised metadata which will allow reuse and sharing across various information systems. Other sections of the RDA COVID-19 Guidelines and Recommendations provide guidance on some of these.

**Data quality and documentation**

**Challenge:**

In the race against time to collect the data required to combat the COVID-19 pandemic, there is risk that data is collected without sufficient attention to quality and reliability of data (e.g. level, or rather lack of any basic provenance of the data, quality of the sources, versioning and level of maintenance).Application developers may not always be aware of the required quality for data to be usable or reusable.

**Guidelines:**

1. Follow standardised ways of collecting and curating community generated data and select a secure data collection platform and trusted digital repositories as a way of standardising COVID-19 data whilst ensuring quality and facilitating sharing. Compilation of recommended repositories can be found here and RDA recommends the use of CoreTrust approved repositories.
2. When collecting and curating the data, ensure detailed metadata is captured with the data. As a minimum the effort should be taken to include the following:
    - 2.1. Provide contextual metadata to help processing, visualization, analysis, storage, publishing, archiving and reuse.
    - 2.2. Include detailed descriptions of the methods, to aid verification of results.
    - 2.3. Include details on the consent and type of consent associated with the collected data.
    - 2.4. Metadata should also include any retention (and deletion) obligations associated with the data.
    - 2.5. Also, where possible, consider including as metadata, specific information on technology characteristics and their limitations (eg: efficiency of underlying technology of app, eg: Bluetooth or GPS).

**Recommendations:**

The research data community has been addressing these challenges, developing standards, vocabularies and ontologies, workflows and various disciplinary norms, as well as a key set of key principles to ensure data quality, findability, accessibility, interoperability and reuse (FAIR).

Implementing the FAIR data principles will ease sharing and increase efficiency, especially important considering the time constraints we are facing.

### Data Storage and long - term preservation

**Challenge:**

Considerations for long term storage and preservation of data generated from apps in relation to COVID-19 is not always apparent. For example, what are the retention periods that apply for COVID-19 related data? Due to the unprecedented nature of this pandemic, much of this is only being considered at present.

**Guidelines:**

1. Ensure that all prevailing national and international legal and ethical requirements for health data and medical studies (e.g. for data retention periods) is adhered to.
2. Ensure that provision is made to enable easy updating of the data collection, storage and preservation to meet any changes to existing requirements.
3. Long-term preservation should be considered in the case of high-value data that could help in modelling future pandemics. Depositing the data in trusted and certified repositories that are widely used by the community aids in achieving this. FAIRsharing.org maintains a comprehensive catalogue of repositories which can be considered for this purpose.
4. Data should be available under an open licence that enables reuse, such CC-BY, unless there are legal and ethical considerations.
5. Consider benefits and challenges of either a centralised or decentralised model for data storage and processing. e.g. View the dedicated section on the processing in a centralised vs decentralised manner from the COVID SafePaths report "COVID-19 Contact-Tracing Mobile Apps: Evaluation And Assessment For Decision Makers" https://drive.google.com/file/d/1A9Ft7-YpB9IOCbaLrRHrR34XP2SiSet5/view

### Legal and ethical aspects

**Challenge**

Ethical considerations have to be made regarding the two-way sharing of information using mobile-tracking apps.

**Recommendations:**

1. Adequate medical, social and emotional support networks need to be established before apps relay to users they may have been in close proximity to a COVID-19 positive individual.The app project owners need to work with relevant local, national and international authorities to ensure appropriate support networks are in place and the app coordinates with these authorities in such matters. .
2. Make sensitive technical consideration such as transmitting anonymised codes as a means to alert individuals to exposure

### Software development

Contact tracing apps should adhere to the same development recommendations as other software, particularly to build public trust. While it has been highlighted that scientists must openly share the code behind modelling software so that the results can be replicated and evaluated (Barton et al. 2020), the transparency provided by open sharing can only address security concerns.

### 5.2.3 Our Approach Going Forward

Whenever possible, we aim to reuse and share applicable recommendations that already exist for specific communities and/or types of data. To this end, we will adopt a standardised approach to identify existing guidance related to specific use cases in communication with relevant communities.

For existing guidance, the subgroup aims to collaborate with relevant communities to review and help refine it and support a broader distribution. If guidance is needed but not available yet, the subgroup will help identify issues and support drafting applicable recommendations. Beyond that, we encourage community members to help translate such recommendations (i) between languages; (ii) from prose into practice, including code and other formalized workflows; (iii) from one community or data type to similar ones.

Topics that we anticipate to be relevant in the context of the above-mentioned use cases include but are not limited to: collaborative data collection, collaborative service or software development initiatives, crowdsourcing of data curation services, data sovereignty when sharing across communities, citizen-led community responses, participatory disaster response strategies, digital platforms or apps to enable public participation and/or offer open data, digital tools to enable public participation.

Furthermore, the group plans to leverage the strengths of the RDA as an international community of data specialists and practitioners as well as reach out beyond to ensure expert input in addressing overarching topics such as ethics and social aspects, indigenous data, global open research commons, metadata standards, persistent identifiers and scientific annotation.

## 5.3 Additional Working Documents & Links

- RDA-COVID-19
- RDA-COVID-19 Community Participation
- Initial scoping doc for Community participation recommendations
- Parent document of RDA COVID-19 WG
- Root folder

# 6. Epidemiology Sub-Group Guidelines

## 6.1 Sub-Group Focus and Description

Responses to the COVID-19 pandemic have been massive and multifaceted worldwide. An immediate understanding of the disease’s epidemiology is crucial to slowing infections, minimizing deaths, and making informed decisions about when, and to what extent, to impose mitigation measures, and when and how to reopen society. As economies “open up,” improved and innovative surveillance and follow-up across the globe is key to minimizing resurgence.
We are still in the midst of the current COVID-19 pandemic. Currently, data and models are incomplete, provisional, and subject to correction under inconsistent and changing conditions. New data and newly applied analytics will provide a better understanding of our current situation and new insights surrounding improved SARS-CoV-2 and COVID-19 responses. Despite the our reliance on, and the importance of evidence based policy and medical decisions, there is no standard or coordinated system for collecting, documenting, and disseminating COVID-19 related data and metadata, making their reuse for timely epidemiological analysis challenging due to issues with documentation, interoperability, completeness, and reliability of the data.

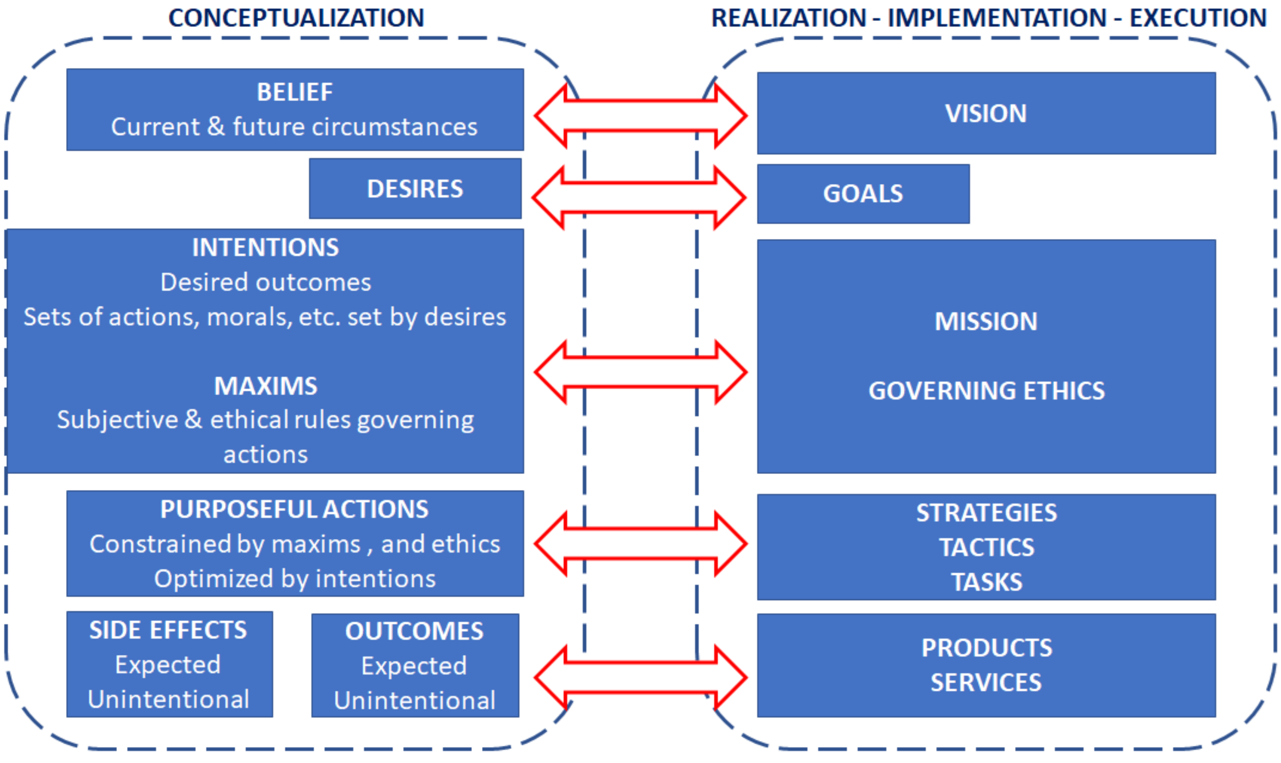
The key elements that block sharing and reuse of epidemiology data are common across many domains. These include non-machine-readable data (e.g., pdf, jpg), heterogeneous measurement standards, divergent metadata formats or lack of metadata, lack of version control, fragmented datasets, delays in releasing data, non-standard definitions and reporting parameters, unavailable or undocumented computer code, copyright and usage conditions, translation requirements, consents, approvals, and legal restrictions. In addition, clinical, eHealth, surveillance, and research systems within and across jurisdictions do not integrate well, or at all.

The present crisis demonstrates more than ever before just how intimately connected and interdependent the world is across countries and organizations. It also lays bare the stark reality and shortcomings of our largely antiquated data systems and data sharing agreements within and between domains, that severely hinder rapid detection of emerging threats and development of a science-based response to them. Barriers are encountered between countries and between jurisdictions within countries, and between national and international organizations.

The epidemiology of COVID-19 is dependent on input data from across a wide variety of domains that include not only clinical and surveillance data, but also administrative, demographic, socioeconomic, and environmental data amongst others. A process of scientific data modernization and related policies in all of these domains is urgently needed to support epidemiologic analyses and modeling that provide critically important insights and understanding of the newly emergent SARS-CoV-2 virus and the COVID-19 disease that it causes.

Implementation of the principles and tools of Open Data and Open Science (e.g. Open Access and FAIR data) that have been under development for several years would solve many of these problems. While science has been gradually moving in this direction, it will require a concerted effort by governments, policy makers, research institutions, clinicians and scientists worldwide to achieve the culture change needed for full adoption. The COVID-19 pandemic highlights the urgent need to remove barriers and accelerate this process now to better respond to the current need for rapid discovery, acquisition and integration of relevant data, and sharing of accurate data to support evidence-informed public health decisions during this rapidly evolving catastrophe.

The RDA-COVID19-Epidemiology Work Group is presently working on how to move from an emotional response to the pandemic to a practical response, which means moving from conceptualization to realization, implementation, and execution (Figure 2).



Contributed to the RDA-COVID19 EPI sub-WG by Gary Mazzaferro    License: CC BY-NC-SA

*Figure 2. Moving from conceptualization to implementation (v0.01). Contributed by Gary Mazzaferro. LICENCE: CC BY-NC-SA 3.0*

## 6.2 Initial Sub-Group Guidelines

### 6.2.1 Summary

**Policy**

1. Urgently update data sharing policies and Memoranda of Understanding (MOUs).
2. Implement a "data first" publication policy in research.
3. Accelerate the implementation of Open Data and Open Science tools and methods.
4. Add "Open Science" to the Open Government Partnership (OGP) list of policy areas.
5. Build and maintain public trust with policies of openness, transparency, privacy protection and honesty.

**IT and data management infrastructure**

6. Invest in information technology (IT) and data management system infrastructure modernization.

**Analysis and modeling**

7. Internationally harmonized COVID-19 intervention policies. protocols.
8. Account for public health decision making in modelling COVID-19 inputs and outputs.

9. Harmonize approaches to comparably quantify side-effects of pandemic mitigation measures.
10. Provide uncertainty quantification for all data and models.
11. Identify hotspots using data-driven approaches.

**Surveillance data**

12. Develop a consensus standard definitions and criteria for COVID-19 surveillance data across public health, clinical and other domains.
13. Document methodologies used to collect and compile data.
14. Develop standardised tools for aggregating microdata to harmonized formats.
15. Develop machine readable citations and micro-citations for dynamic data.

**Interoperability and data exchange**

16. Develop a technical specification for record linkage across epidemiological input and output systems.
17. Develop systems to share pseudonymized data using encrypted person identifiers.
18. Share metadata and aggregated data where there are restrictions in accessing/using the related person-level data.

**Global preparation, detection and response**

19. Create a WHO-led COVID-19 EPIdemiological Translational Research Action Centre (Epi-TRAC).

## 6.2.2 Discussion

**Policy - recommendations**

1. Urgently update data sharing policies and Memoranda of Understanding (MOUs) across all domains, in government, healthcare systems, and research institutions to support Open Data, Open Science, scientific data modernization, and linked data life cycles that will enable rapid and credible scientific and epidemiologic discovery, and to fast-track decision-making. For example, between the countries and the WHO, between the European Commission and the USA, and between sub-national jurisdictions/institutions and their national government.
2. Implement a "data first" publication policy in research by treating publication of data articles in "open" peer-reviewed data journals, including deposit of the data and associated code in a trusted digital repository, as first-class research outputs equal in value to traditional peer-reviewed articles.
3. Rapid development of government and institutional policies to accelerate the implementation of Open Data and Open Science tools and methods across all science and health domains.
4. Call upon the international Open Government Partnership (OGP) to add "Open Science" as one of its Policy Areas to be included in National Action Plans. Member countries would then be held accountable for developing and implementing Open Science commitments via the Independent Reporting Mechanism (IRM) that tracks the progress of OGP members.
5. Build and maintain public trust: Implement a policy of openness, transparency, and honesty with respect to COVID-19 related data and models, and what we know and do not know. Publish situational data, analytical models, scientific findings, and reports used in decision-making and justification of decisions (OGP 2020).

**IT and data management infrastructure - recommendations**

6. Invest in information technology (IT) and data management system infrastructure (devices or hardware, and algorithms or software used to store, retrieve and process data).

6.1. Rapid development of a modern data management system infrastructure will ensure scientific data integrity via data management plans embedded in linked data life cycles that: (a) are fully machine-enabled, and not constrained by non-digital processes; (b) are available online end-to-end; (c) enable synchronous and asynchronous workflows; (d) guarantee tidy, Findable, Accessible, Interoperable, Reusable, Ethical, and Reproducible (FAIRER) data, metadata, and code/scripts; (e) guarantee data security; (f) provide tiered access to restricted data by appropriately credentialed users and machines; and, (g) analytical tools. See, for example, ELIXIR Galaxy.

6.2. When evaluating apps consider the many underlying issues: legal, confidentiality, data completeness, representativeness, data quality, reliability, verifiability, data ownership, data access, data openness, data control, transparency, peer-review, etc.

**Analysis and modeling - recommendations**

7. Develop and implement internationally harmonized COVID-19 intervention protocols based on peer-reviewed empirical modeling and epidemiological evidence, taking into account local conditions.
8. Account for public health decision making in modelling COVID-19 inputs and outputs.
9. Harmonize approaches to comparably quantify side-effects of pandemic mitigation measures on society, for example, shifts in morbidity, mortality, health care utilisation, quality of life, social isolation.
10. Report underlying assumptions and quantify effects of uncertainties on all reported parameters and conclusions for all model predictions, data etc.
11. Implement a data driven approach to identify hotspots.

**Surveillance Data - recommendations**

12. Rapid development of a consensus standard on COVID-19 surveillance data:

12.1. Definition of and reporting criteria for COVID-19 testing, reporting on testing, and testing turnaround times.

12.2. Policies and definitions: interventions, contact tracing, reporting of cases, deaths, hospitalizations and length of stay, ICU admissions, recoveries, reinfections, time from contact if known, symptoms onset and detection, through clinical course and interventions, to death or recovery, comorbidities, follow up to identify serious long-term effects in recovered cases, sequelae and immunity, location, demographic, socioeconomic information, and outcome of resolved cases.

12.3. Uniform standard daily reporting cut-off time.

13. Document methodologies used to collect and compile data, including data management, data cleaning, data quality checks, updating, data imputation, computer code used, definitions used, etc.
14. Rapid development of standardised tools for aggregating microdata to a harmonized format(s) that can be shared and used while minimising the re-identification risk for individual records.
15. Rapid development of: (a) Resolvable Persistent Identifiers, rather than Uniform Resource Locators (URLs), to provide the ability to successfully access the data over decades; (b) Machine readable citations that allow machines to access and interpret the resource; (c) Micro-citations that refer to the specific data used from large datasets; and, (d) Date and Time Access citations for dynamic data (ESIP 2019).

A major difficulty at this time is the lack of contextual data needed to study the evolution of disease in sub-populations. They include, among others, otherwise healthy sub-populations that are vulnerable to serious long-term effects following recovery that we do not know about yet because we don't have the data and because we are focusing on deaths. They also include age-specific vulnerabilities, disadvantaged sub-populations with limited health care, vulnerabilities evident in severe disease associated with comorbidities, and vulnerabilities due to environmental conditions, and due to social and cultural norms. Vigilance will be necessary to follow sequelae and immunity. These data are not collected systematically in the healthcare system nor via different survey instruments. Moreover, merging clinical databases with other types of databases is difficult or impossible due to interoperability and legal reasons.

Conceptualization of an epidemiological surveillance data model (Figure 3) identifies the primary data domains that need to be integrated to understand COVID-19, and to improve surveillance and follow-up: (a) clinical event history and disease milestones; (b) epidemiological indicators and reporting data; (c) contact tracing; (d) person risk factors.

However, standardization challenges within each of these domains remain to be solved before data can be effectively integrated across domains for epidemiology studies. For example, on the clinical side, the U.S. Clinical Data Interchange Standards Consortium (CDISC) new specification (Interim User Guide for COVID-19), and the WHO Core and Rapid COVID-19 Case Reporting Forms used in low- and middle-income Countries (LMIC) require additional harmonization. Surveillance event history must be integrated across an interface with clinical data. In addition to standard treatments, such providing oxygen passively or aggressively to lungs, dialysis for kidney damage, managing coagulopathy/stroke/heart attack/pericarditis, etc., there is a capacity concern including drug availability.

In addition, compassionate or experimental treatments and trials include, for example, extracorporeal oxygenation, and ad hoc drug treatments with limited evidence of outcomes and with little potential to learn from experience. Contact history and location is another unsettled domain given community surveillance in LMICs and elsewhere, and competing visions in the rapid emergence of various apps from the academic, government, and private sectors which may or may not provide an individual’s geospatial location. There is inconsistent collection of person risk factor information. New York State is developing a COVID-19 risk matrix for establishments that they plan to use in “reopening” the state. Some of the person risk factors may be interrelated, and it will take some future data science to reduce the dimensionality of the risk factor space.

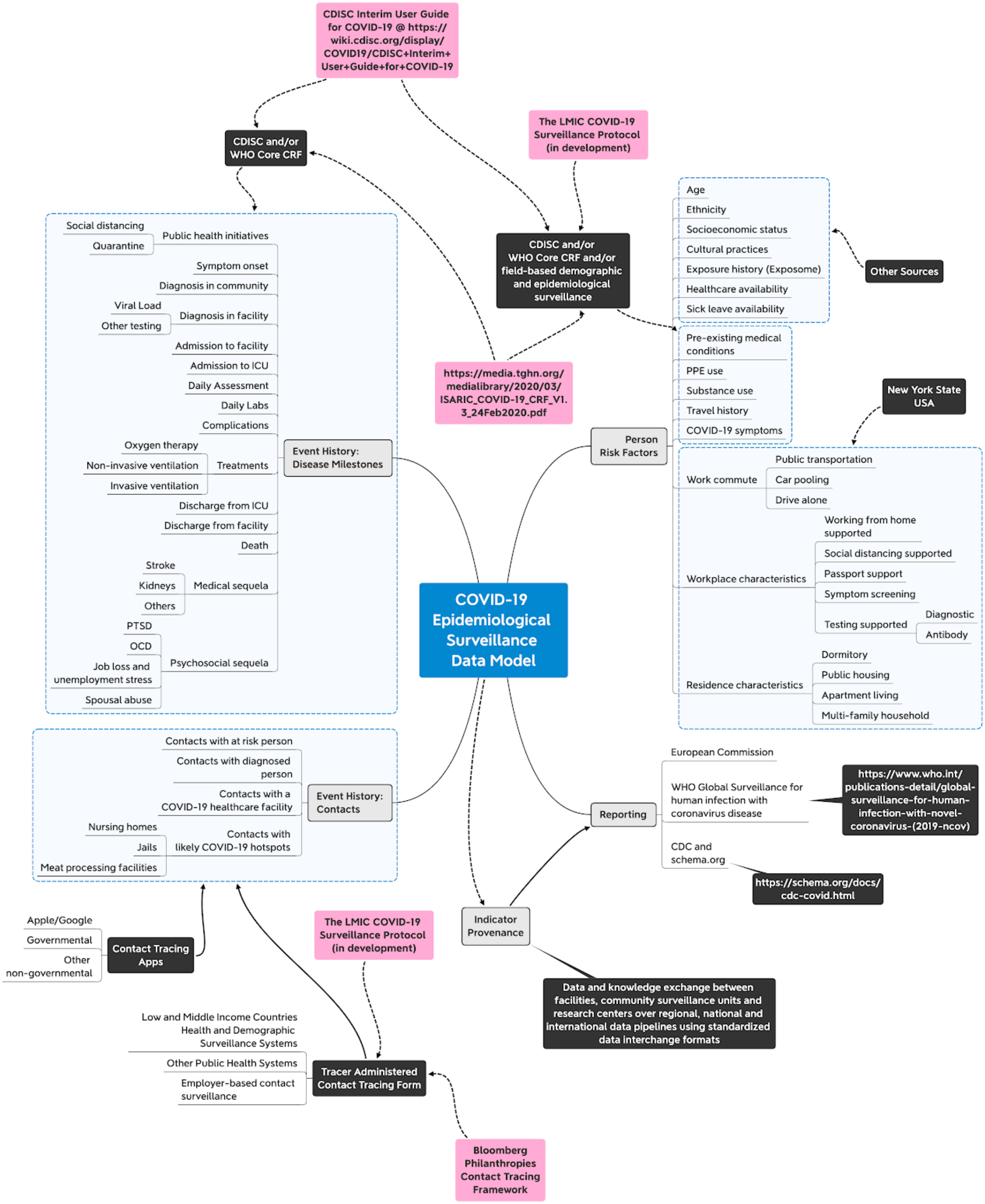

CDISC Interim User Guide for COVID-19 @ https://wiki.cdisc.org/display/COVID19/CDISC+Interim+User+Guide+for+COVID-19
CDISC and/or WHO Core CRF
The LMIC COVID-19 Surveillance Protocol (in development)
CDISC and/or WHO Core CRF and/or field-based demographic and epidemiological surveillance
https://media.tghn.org/medialibrary/2020/03/ISARIC_COVID-19_CRF_V1.3_24Feb2020.pdf
Social distancing
Quarantine
Public health initiatives
Symptom onset
Diagnosis in community
Viral Load
Other testing
Diagnosis in facility
Admission to facility
Admission to ICU
Daily Assessment
Daily Labs
Complications
Oxygen therapy
Non-invasive ventilation
Invasive ventilation
Treatments
Discharge from ICU
Discharge from facility
Death
Stroke
Kidneys
Others
Medical sequela
PTSD
OCD
Job loss and unemployment stress
Spousal abuse
Psychosocial sequela
Event History: Disease Milestones
Age
Ethnicity
Socioeconomic status
Cultural practices
Exposure history (Exposome)
Healthcare availability
Sick leave availability
Other Sources
Pre-existing medical conditions
PPE use
Substance use
Travel history
COVID-19 symptoms
Person Risk Factors
New York State USA
Work commute
Public transportation
Car pooling
Drive alone
Workplace characteristics
Working from home supported
Social distancing supported
Passport support
Symptom screening
Testing supported
Diagnostic
Antibody
Residence characteristics
Dormitory
Public housing
Apartment living
Multi-family household
COVID-19 Epidemiological Surveillance Data Model
Contacts with at risk person
Contacts with diagnosed person
Contacts with a COVID-19 healthcare facility
Nursing homes
Jails
Contacts with likely COVID-19 hotspots
Meat processing facilities
Event History: Contacts
European Commission
WHO Global Surveillance for human infection with coronavirus disease
CDC and schema.org
Reporting
https://www.who.int/publications-detail/global-surveillance-for-human-infection-with-novel-coronavirus-(2019-ncov)
https://schema.org/docs/cdc-covid.html
Apple/Google
Governmental
Other non-governmental
Contact Tracing Apps
The LMIC COVID-19 Surveillance Protocol (in development)
Indicator Provenance
Data and knowledge exchange between facilities, community surveillance units and research centers over regional, national and international data pipelines using standardized data interchange formats
Low and Middle Income Countries Health and Demographic Surveillance Systems
Other Public Health Systems
Employer-based contact surveillance
Tracer Administered Contact Tracing Form
Bloomberg Philanthropies Contact Tracing Framework

*Figure 3. Epidemiology surveillance data model (v0.02). Contributed by Dr. Jay Greenfield, developed from discussions during RDA-COVID19-Epidemiology Work Group meetings. LICENCE: CC BY-NC-SA 3.0. NOTE: We're working on fixing the resolution.*

## Interoperability and data exchange - recommendations

16. Rapid development of an internationally harmonized specification to enable the export/import of epidemiologic data from clinical systems, record linkage to population-based surveillance data, and automatic submission to disease reporting systems and research infrastructures.

17. Develop systems that support workflows to link and share pseudonymized data between different domains, while enabling privacy and security. Use domain specific, time stamped, encrypted person identifiers for this purpose.
18. Share Metadata where there are restrictions in accessing/using the related data.

**Anonymization and pseudonymization**

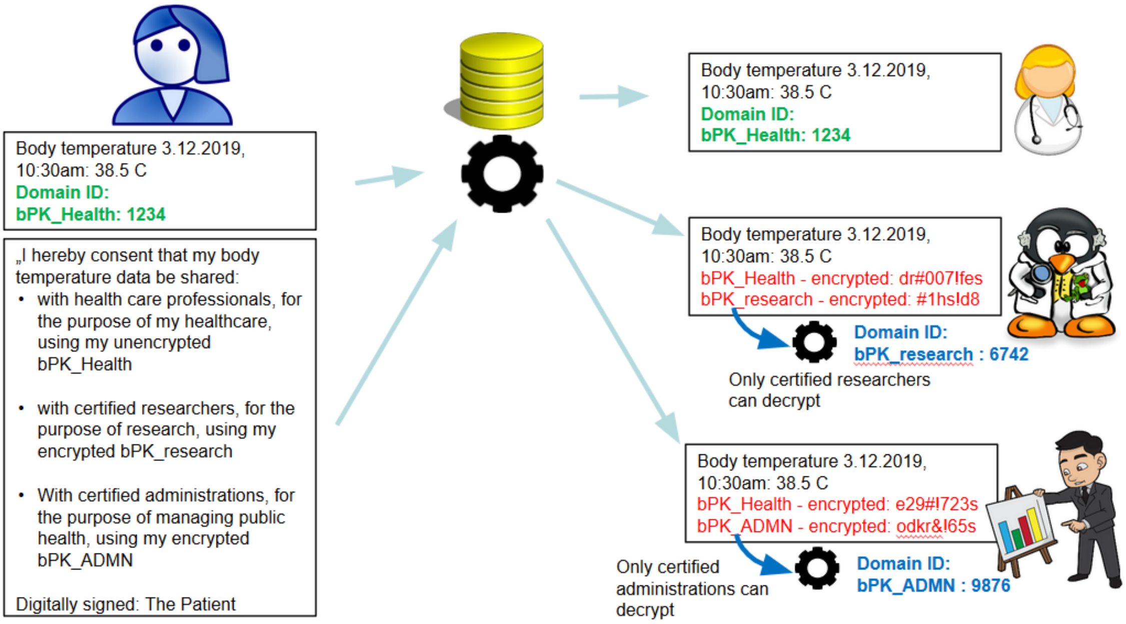
Patient related data are recorded and used in different domains, for example medicine, communities, research, administration, and statistics. Patient specific electronic identifiers (IDs) link specific information to the individual patient records in each domain. Each domain assigns a domain specific ID to each individual patient. In order to satisfy privacy requirements, data that carry a domain ID remain within the home domain, and are not shared.

If data are re-used e.g. for research or public health purposes, the domain specific ID is removed (anonymisation) or replaced by a different ID (pseudonym). Pseudonyms can later be traced back to the original domain ID. This must only occur under well-defined conditions, e.g. if a statistics department needs to clear ambiguities in incoming information together with the organisation that generated the data. In multi-domain and multi-organisation scenarios, consistent management of IDs and pseudonyms is needed to enable cross-linking of data from different sources while ensuring privacy.

The following requirements apply:

1. Patient IDs exist for each domain;
2. The local domain ID must not leave the local domain in clear text, to prevent unintended record linkage between domains;
3. When providing data from a source domain to a target domain, the target domain patient ID (pseudonym) must become available to users in the target domain; and,
4. Domain IDs must enable domains to cooperate e.g. for clearing ambiguities, while preserving privacy and pseudonymisation.

In Austria, for example, eGovernment legislation and IT infrastructures are in place to handle domain specific identifiers e.g. for health care, traffic, taxes, and statistics [1], [2]. This is implemented and in operation for example in the Austrian electronic health care record ELGA [3]. Figure 4 describes how data from a health domain can be linked to records in a research domain in this way. Figure 5 introduces the needed IT infrastructure. Figure 6 shows how IDs are mapped between domains while preserving pseudonymisation.



Body temperature 3.12.2019, 10:30am: 38.5 C
Domain ID:
bPK_Health: 1234

„I hereby consent that my body temperature data be shared:

- with health care professionals, for the purpose of my healthcare, using my unencrypted bPK_Health
- with certified researchers, for the purpose of research, using my encrypted bPK_research
- With certified administrations, for the purpose of managing public health, using my encrypted bPK_ADMN

Digitally signed: The Patient

Body temperature 3.12.2019, 10:30am: 38.5 C
Domain ID:
bPK_Health: 1234

Body temperature 3.12.2019, 10:30am: 38.5 C
bPK_Health - encrypted: dr#007Ifes
bPK_research - encrypted: #1hsId8

Domain ID:
bPK_research : 6742

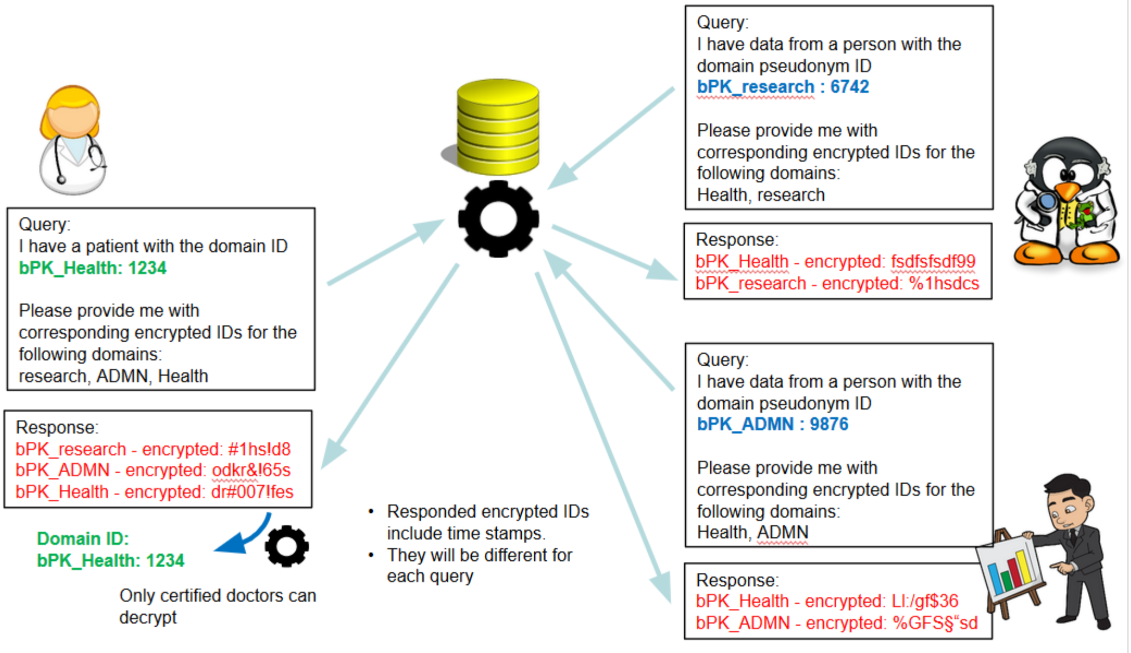
Only certified researchers can decrypt

Body temperature 3.12.2019, 10:30am: 38.5 C
bPK_Health - encrypted: e29#I723s
bPK_ADMN - encrypted: odkr&I65s

Only certified administrations can decrypt

Domain ID:
bPK_ADMN : 9876

*Figure 4. Sharing or linking a body temperature observation from the healthcare domain with a research and administration domain. In the healthcare domain ID (bPK_Health, green), a patient is identified with a specific ID, (1234, colour green, denoting that it is unencrypted). A doctor will receive this data together with the original ID, as the law allows doctors to share IDs unencrypted. The doctor can attach an encrypted ID (#1hsId8, red, denoting it is encrypted) to the data. A researcher who receives the data, decrypts the encrypted ID. This decrypted ID (6742, blue, denoting that it is a pseudonym) is specific to the research domain (bPK_research, blue). The same method applies as data are provided to administrations, e.g. for public health purposes. Contributed by Dr. Carsten Schmidt. LICENCE: CC BY-NC-SA 3.0*



Query:
I have a patient with the domain ID
bPK_Health: 1234

Please provide me with corresponding encrypted IDs for the following domains:
research, ADMN, Health

Response:
bPK_research - encrypted: #1hsId8
bPK_ADMN - encrypted: odkr&I65s
bPK_Health - encrypted: dr#007Ifes

Domain ID:
bPK_Health: 1234

Only certified doctors can decrypt

- Responded encrypted IDs include time stamps.
- They will be different for each query

Query:
I have data from a person with the domain pseudonym ID
bPK_research : 6742

Please provide me with corresponding encrypted IDs for the following domains:
Health, research

Response:
bPK_Health - encrypted: fsdfsfsdf99
bPK_research - encrypted: %1hsdcs

Query:
I have data from a person with the domain pseudonym ID
bPK_ADMN : 9876

Please provide me with corresponding encrypted IDs for the following domains:
Health, ADMN

Response:
bPK_Health - encrypted: LI:/gf$36
bPK_ADMN - encrypted: %GFS§"sd

*Figure 5. IT infrastructure for cross-domain IDs management. A user in the medical domain queries the IT service using the ID of the health domain, asking for encrypted IDs of other domains. The service responds with the encrypted IDs. Users in other domains can use the same mechanism. This enables users in different domains to co-operate: For example the researcher can attach the encrypted bPK_Health ID to a message to the doctor, asking for details to support clearing ambiguities in the data the doctor provided earlier. The doctor can then decrypt the patient ID, access the patient related information in the health IT system, and finalise the clearing with the researcher. Contributed by Dr. Carsten Schmidt. LICENCE: CC BY-NC-SA 3.0*

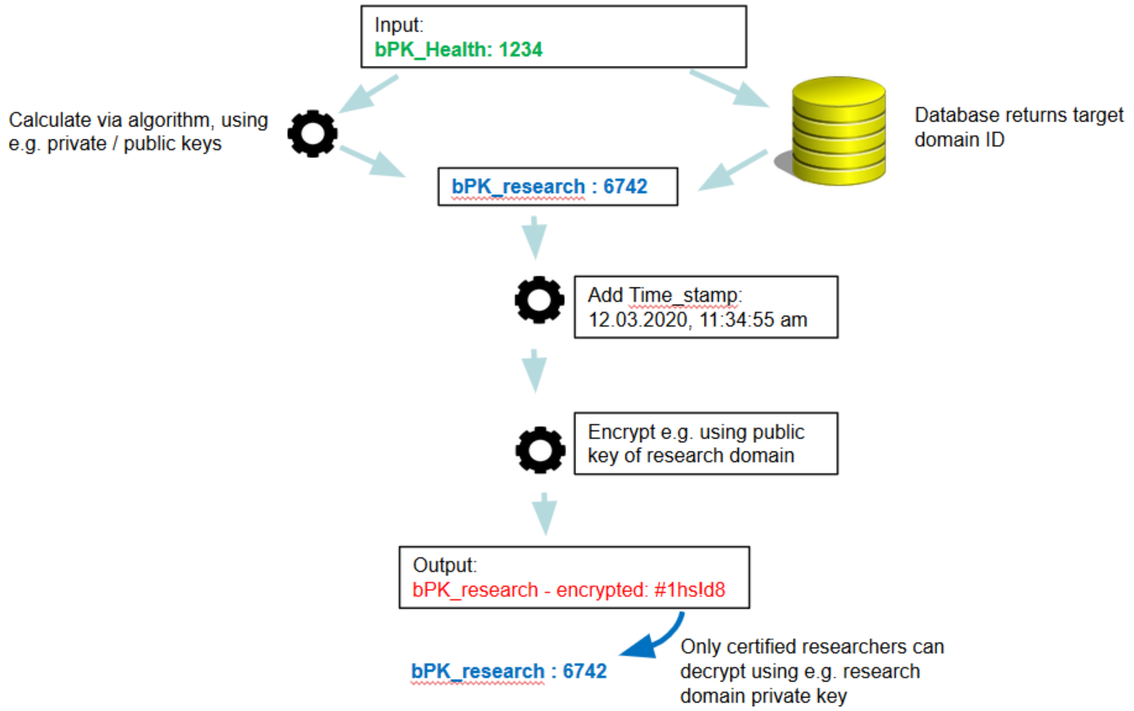

Input:
bPK_Health: 1234
Calculate via algorithm, using e.g. private / public keys
Database returns target domain ID
bPK_research : 6742
Add Time_stamp: 12.03.2020, 11:34:55 am
Encrypt e.g. using public key of research domain
Output:
bPK_research - encrypted: #1hs!d8
bPK_research : 6742
Only certified researchers can decrypt using e.g. research domain private key

*Figure 6 Data flow for deriving an encrypted ID (#1hs!d8) for a target domain (research). The source ID (1234) can be mapped to the target identifier for example in two ways. A mathematical algorithm is used (left branch) to calculate the target domain ID (6742), or a database query returns the ID. A time stamp is then attached to this ID. ID and timestamp together are then encrypted, e.g. using the public key of the target domain. This assures that no two encrypted IDs for the same patient and the same domain are identical, in this way preventing the unintended linking of records. Contributed by Dr. Carsten Schmidt. LICENCE: CC BY-NC-SA 3.0*

One use case for anonymization, pseudonymization, cross-border data discovery and cross-border data transfer are the international efforts underway to create interoperable COVID-19 demographic and epidemiological surveillance questionnaires. These initiatives are domain-specific and support the localization of questions and answers in order to assure cross-country and cross-cultural comparability.

### COVID-19 questionnaire initiatives

There are a number of actively developing COVID-19 questionnaire initiatives that figure into one or more international efforts to create interoperable COVID-19 epidemiological surveillance questionnaires (Tables 1 and 2). These initiatives are very much a work in progress at this time. Indeed, some of us are already participating in these international efforts and are just now learning about each other's work via the RDA-COVID19-Epidemiology Work Group.

| Country | Initiative | Target population | Development stage | Language | Provenance (Influenced by...) | Comments |
|---|---|---|---|---|---|---|
| 1 Australia | NSW Case questionnaire | Patients | | English | | |
| 2 Brazil | Brazil Prevalence of Infection Survey | Rapid tested, Tested positive | In development | English | | |

| Country | Initiative | Target population | Development stage | Language | Provenance (Influenced by...) | Comments |
|---|---|---|---|---|---|---|
| 3 Europe | Questionnaire by WHO Europe | General population | | German, Russian | | Single Instrument |
| 4 Germany | Covid-19 research dataset | Patients | In development | German | | National Network of German University Clinics to study COVID19. |
| 6 Germany | GESIS Panel Special Survey on the Coronavirus SARS-CoV-2 Outbreak in Germany | General population | In use | German | | As the largest European infrastructure institute for the social sciences GESIS provides essential and internationally relevant research-based services |
| 7 Israel | One-minute population wide survey (Israel) | Isreali population | In use | Hebrew, Arabic, Russian, Spanish, French, English | | Participants asked to fill it out on a daily basis and separately for each family member, including members who are unable to fill it out independently (e.g., children and older people). |
| 8 Low and Middle Income Countries | LMIC Covid Questionnaire | | In development | | UK COVID-19 questionnaire, SAPRIN COVID-19 screening form | |
| 9 South Africa | South African Population Research Infrastructure (SAPRIN) COVID-19 Screening Form | | In development | English, Afrikaans | | |
| 10 Uganda | Perinatal COVID-19 Uganda | Women pre-/perinatal | | English | | |
| 11 UK | UK COVID-19 Questionnaire | Adult Respondent, Children, Key worker, Partner. | In development | English | NIHR Global Health Research Unit | - World Bank code book for metadata.<br>- Becoming a model for some African countries |
| 12 UK | National Institute for Health Research (NIHR) Global Health Research Unit | Telephone sample | | English | | |

| Country | Initiative | Target population | Development stage | Language | Provenance (Influenced by...) | Comments |
|---|---|---|---|---|---|---|
| [13] US | Human Infection with 2019 Novel CoronavirusPerson Under Investigation (PUI) and Case Report Form | Patients | In use | English | CDC | |
| [14] US/WHO | Population-based age-stratified seroepidemiological investigation protocol for COVID-19 virus infection | Patients | In use | English | WHO | This protocol has been designed to investigate the extent of infection, as determined by seropositivity in the general population, in any country in which COVID-19 virus infection has been reported. |

**Footnotes (Domains)**

[1] Clinical symptoms, Disease outcome, Exposure sites, Pre-existing conditions, Risk history, Sociodemographics.

[2] Demographics, Home life, Test results, Transportation

[3] Affect, Behaviour, Conspiracies (perceptions), COVID-19 risk perception: probability and severity, Fairness (perceptions), Frequency of Information, Influenza risk perception: probability and severity,, interventions (perceptions), Knowledge and self-assessed adherence to prevention measures, Knowledge incubation, Knowledge symptoms/treatment, Lifting restrictions (pandemic transition phase), Policies, Preparedness and perceived self-efficacy, Prevention – own behaviours, Resilience (perceptions), Risk group, Rumors (open-ended), Self-assessed knowledge, Socio-demography, Trust in institutions (perceptions), Trust in sources of information, Use of sources of information, Worry.

[4] Clinical symptoms, Complications, Imaging, Laboratory markers, Medical treatments, Medication, Sociodemographics.

[5] Adherence to risk minimization measures, Changes in lifestyle factors, COVID infections and testing, COVID tracing, General health status, Physical symptoms, Respiratory infections, Workplace/changed employment situation.

[6] Changed employment situation, Childcare obligations, Risk perception, Evaluation of political measures & their compliance, Media consumption, Risk minimization measures, Trust in politics and institutions.

[7] Age, Geographic location (city and street), Isolation status, Sex, Smoking habits

[9] Actions in response to COVID-19, Bounded structure, Eligibility for testing, Epidemiological risk, Household enumeration, Household impact, Quarantine and hygiene directions,, Symptom screen, Travel and movement (mobility), Travel history, Visit attempts.

[10] Disease outcome, Sociodemographics, Symptoms

[11] Accommodation type, Away from home environment, Behavior changes, Change in benefits, Digital access, Economic activity before and after lockdown, Environmental attitudes, Environmental impact, Family relations, Financial impact, Food security, Impact on employment, Knowledge, Medication, Mental health, Mental health, Physical health, Pre-existing conditions, Social impact, Symptoms, Volunteering.

[12] COVID-19 Interventions, Current living conditions, Displacement and mobility, Economic impacts, Impact of COVID-19 on health-related behaviors, Mental health, Precautions, Pre-existing conditions, Social aid, Social impact, Symptoms, Treatments for pre-existing conditions.

[13] Diagnostic testing procedures, Clinical course, Medical history, Pre-existing conditions, Risk exposure, Sociodemographics, Symptoms, Treatments.

[14] Laboratory results, Sociodemographics, Symptoms.

*Table 1. Questionnaire instruments: Reference studies. Development of such initiatives is a very active and rapidly changing area at the present time during the COVID-9 pandemic. Contributed by Dr. Carsten Schmidt, Dr Jay Greenfield, Dr. Stefen Sauermann, and Dr. Chifundo Kanjala, developed from discussions held during RDA-COVID19-Epidemiology Work Group meetings*

| Provider | Initiative | Language |
|---|---|---|
| US | NIH Public Health Emergency and Disaster Research Response (DR2) | Diverse |
| NIH | COVID-19OBSSR Research Tools | Diverse |
| PhenX is funded by the National Institutes of Health (NIH) Genomic Resource Grant | PhenX COVID-19 Toolkit | Diverse |

*Table 2. Questionnaire instruments - Resources. Development of such initiatives is a very active and rapidly changing area at the present time during the COVID-9 pandemic*

Some of the questionnaire initiatives shown in Table 1 and Table 2Table 1 are feeding into the construction of a COVID-19 demographic and epidemiological surveillance question bank.
Note that at the time of writing of the present document, Table 1 and Table 2Table 1 are targeted for improvements in several respects:

1. Instruments and other resources that have been identified require additional review to ensure that key initiatives for COVID-19 have not been overlooked;
2. Consistent categorization of domains and cohorts will increase the usefulness of Table 1 and Table 2Table 1;
3. More provenance information could prove useful to researchers seeking to understand the effects of the pandemic internationally; and,
4. Additional comments and/or contextual information could provide tips about what these initiatives do well and what improvements they might make.
5. How to ensure adequate data to answer research questions rigorously, unambiguously and transparently.

## COVID-19 question bank

The Wellcome Trust is participating in the development of a COVID-19 question bank that can be used to form locality specific surveys with both common and distinct questions by domains and cohorts.
Specific initiatives such as the UK COVID-19 Questionnaire, and the Low and Middle Income (LMIC) Questionnaire for Sub-Saharan Africa and Asia (under development) are now being funded. Such funding may kick start the development of a domain- and cohort-specific question bank. This question bank, once it becomes operational, can, in turn, be queried and filtered by domain, cohort, question text and so forth. Based on such queries, new questionnaire products can be developed that are more or less interoperable, depending on the questions selected and the capture of "localization" information in the question metadata when questions are reused from one survey to the next.

Reporting formats vary considerably between governmental agencies, non-governmental agencies, and the various communities of practice. A translation ecosystem needs to grow around reporting formats to facilitate frictionless flow of information during a pandemic. This ecosystem is indicated by the "exchange" arrow at the "disseminate" box in Figure 7 A draft model domain- and cohort-specific COVID-19 question bank to facilitate frictionless flow of information during the pandemic (v0.02). See, also, Tables 1 and 2 Contributed by Dr. Jay Greenfield, Dr. Chifundo Kanjala, and Dr. Carsten Schmidt, developed from discussions held during RDA-COVID19-Epidemiology Work Group meetings and based on work in the U.K., Asia, and sub-Saharan Africa supported by the Wellcome Trust. LICENCE: CC BY-NC-SA 3.0.

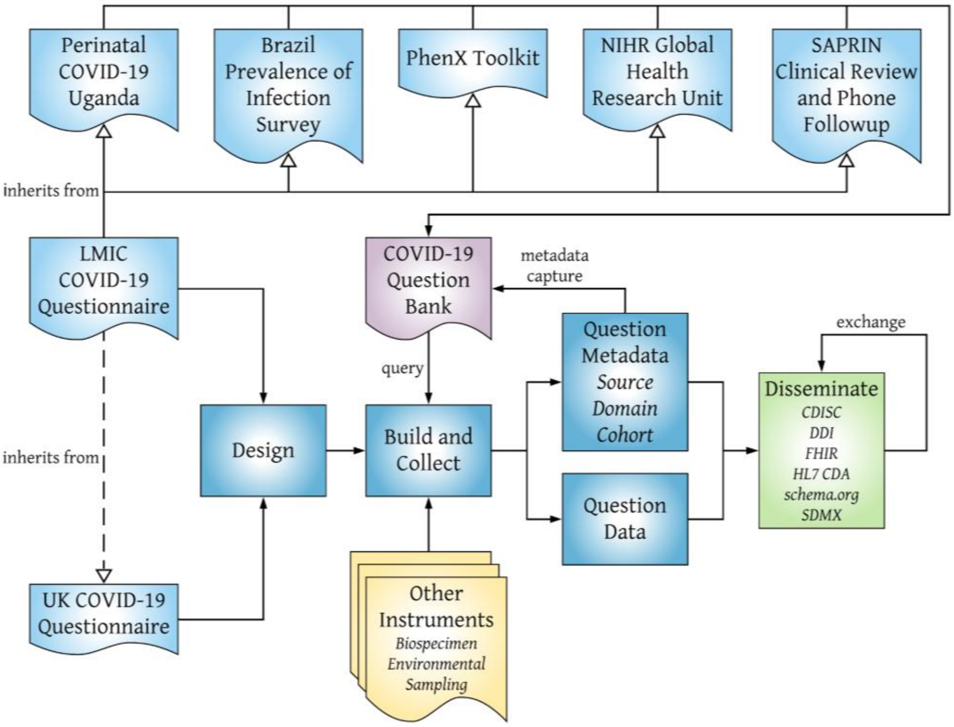


*Figure 7 A draft model domain- and cohort-specific COVID-19 question bank to facilitate frictionless flow of information during the pandemic (v0.02). See, also, Tables 1 and 2 Contributed by Dr. Jay Greenfield, Dr. Chifundo Kanjala, and Dr. Carsten Schmidt, developed from discussions held during RDA-COVID19-Epidemiology Work Group meetings and based on work in the U.K., Asia, and sub-Saharan Africa supported by the Wellcome Trust. LICENCE: CC BY-NC-SA 3.0*

Cross-border, cross domain and semantically interoperable data sources are key to sharing and linking data for pandemic policy making. Patient related data related to the COVID-19 pandemic is handled across clinical, community surveillance (demographic and epidemiological), research, disease management, and social domains. Data in the clinical and community domains support patient care. Regional and national administrations use some of the clinical data for disease management, e.g. in local outbreaks. Researchers generate new knowledge. Contact tracing, telemonitoring, social media are used in the social domain. In order to optimise the outcome, the data flows within and between these domains needs to be further developed to enable secure, safe, timely and reliable automated data processing.

Translational research is already leveraging existing platforms in an impromptu fashion (Westfall et al. 2007). These platforms exchange data and knowledge between facilities, community surveillance units and research centers over regional, national and international data pipelines/networks using standardized data interchange formats. However, these arrangements are developing in an ad hoc fashion and need to be evaluated to determine if they are fit-for-purpose.

In order to accomplish sustainable results, existing programs and initiatives must begin with a well defined set of high priority short term goals, e.g. optimising data use for disease management. In the disease management domain these include the WHO, ECDC, Tessy, Austrian EMS, CDC, NIH and the FDA. Clinical data exchange systems (e.g., in the EU and USA) need to be considered. In the research domain, CDISC must be considered as well as other technical standards from the more clinical space (IHE and HL7). In parallel to short term activities, long-term cooperation needs to be established, under clear coordination and with sufficient resources.

**Global preparation, detection and response - recommendations**

19. Create a WHO-led COVID-19 EPIdemiological Translational Research Action Centre (Epi-TRAC).

WHO's Global Influenza Surveillance Response System (GISRS) is a well-established network of more than 150 national public health laboratories in 125 countries that monitors the epidemiology and virologic evolution of influenza disease and viruses (WHO 2020). On March 26, 2020, WHO published Operational considerations for COVID-19 surveillance using GISRS. This document,

> *"is intended for Ministry of Health and other government officials responsible for COVID-19 and influenza surveillance and summarizes the operational considerations for leveraging influenza surveillance systems to incorporate COVID-19 testing. The enhanced surveillance outputs will support national, regional, and global situation monitoring, knowledge building, risk assessment, and response actions."*

Prior to the COVID-19 outbreak, WHO was already engaged in re-examining GISRS's long-term fitness-for-purpose, reimagining the GISRS based on new themes. In line with these short-term considerations and with GISRS long-term aspirations, we are recommending a real time, adaptable, rapidly-responding system that supports developing countries, and that employs new technology to combat pandemics and other emerging diseases. The RDA-COVID19-Epidemiology WG recommends the creation of a WHO-led Epi-TRAC to add an implementation layer to the existing WHO policies, guidelines, partnerships and information exchange stack (Figure 8).

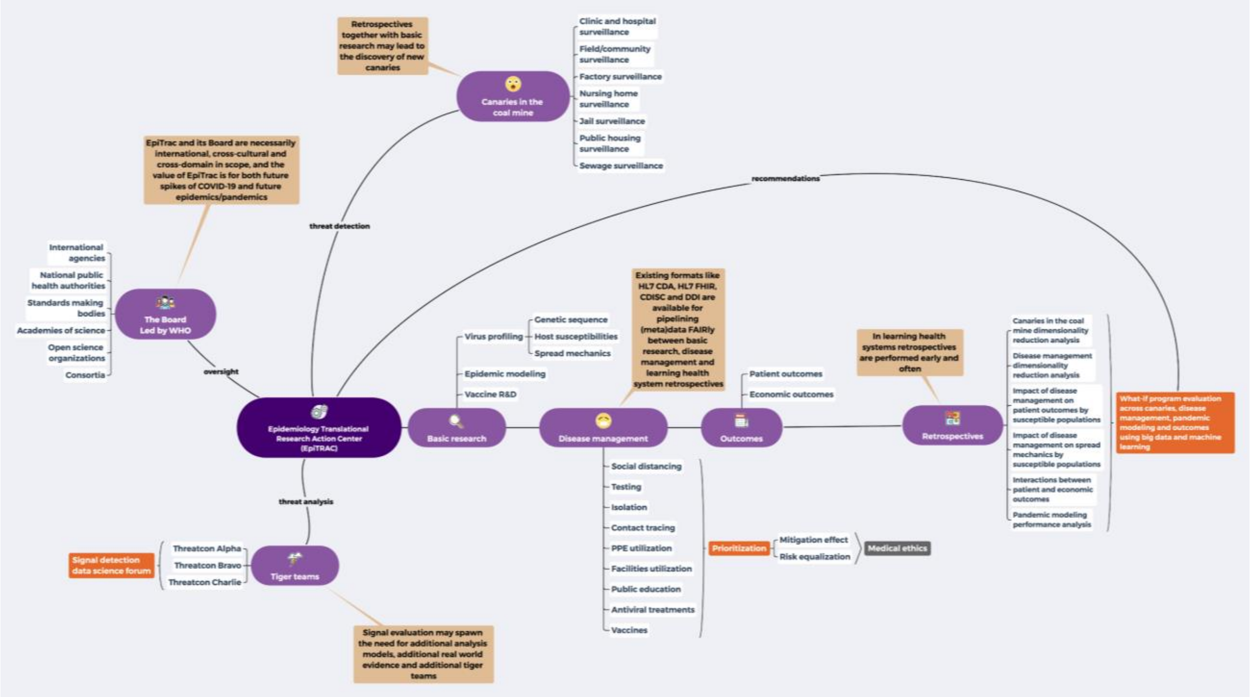

Retrospectives together with basic research may lead to the discovery of new canaries
Canaries in the coal mine
Clinic and hospital surveillance
Field/community surveillance
Factory surveillance
Nursing home surveillance
Jail surveillance
Public housing surveillance
Sewage surveillance
EpiTrac and its Board are necessarily international, cross-cultural and cross-domain in scope, and the value of EpiTrac is for both future spikes of COVID-19 and future epidemics/pandemics
threat detection
recommendations
International agencies
National public health authorities
Standards making bodies
Academies of science
Open science organizations
Consortia
The Board Led by WHO
oversight
Epidemiology Translational Research Action Center (EpiTRAC)
Basic research
Virus profiling
Genetic sequence
Host susceptibilities
Spread mechanics
Epidemic modeling
Vaccine R&D
Existing formats like HL7 CDA, HL7 FHIR, CDISC and DDI are available for pipelining (meta)data FAIRly between basic research, disease management and learning health system retrospectives
Disease management
Social distancing
Testing
Isolation
Contact tracing
PPE utilization
Facilities utilization
Public education
Antiviral treatments
Vaccines
Prioritization
Mitigation effect
Risk equalization
Medical ethics
Outcomes
Patient outcomes
Economic outcomes
In learning health systems retrospectives are performed early and often
Retrospectives
Canaries in the coal mine dimensionality reduction analysis
Disease management dimensionality reduction analysis
Impact of disease management on patient outcomes by susceptible populations
Impact of disease management on spread mechanics by susceptible populations
Interactions between patient and economic outcomes
Pandemic modeling performance analysis
What-if program evaluation across canaries, disease management, pandemic modeling and outcomes using big data and machine learning
threat analysis
Tiger teams
Threatcon Alpha
Threatcon Bravo
Threatcon Charlie
Signal detection data science forum
Signal evaluation may spawn the need for additional analysis models, additional real world evidence and additional tiger teams

*Figure 8. Epi-TRAC (v0.02). A proposed WHO-led COVID-19 EPIdemiological Translational Research Action Centre. Contributed by Dr. Jay Greenfield and Gary Mazzaferro, developed from discussions held during RDA-COVID19-Epidemiology Work Group meetings. Special thanks, also to Dr. Stefan Sauermann, and Gary Mazzaferro. LICENCE: CC BY-NC-SA 3.0*

## 6.3 Additional Working Documents & Links

Additional materials can be found at: RDA-COVID19-Epidemiology

## 6.4 References

Battegay, M., Kuehl, R., Tschudin-Sutter, S., Hirsch, H. H., Widmer, A. F., & Neher, R. A (2020). 2019-novel Coronavirus (2019-nCoV): Estimating the case fatality rate – a word of caution. Swiss Medical Weekly, 150(0506). https://doi.org/10.4414/smw.2020.20203

Brickley D (2020), "Schema.org - COVID Hospital data schema (US CDC)," Feb. 04, 2020. https://schema.org/docs/cdc-covid.html

CDC (2020a). Cases of COVID19 in the U.S. [Dataset]. Centers for Disease Control, USA. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html

CDC (2020b). Weekly provisional death counts from death certificate data: COVID19, pneumonia, flu [Dataset]. Center for Disease Control, USA. https://www.cdc.gov/nchs/nvss/vsrr/covid19/index.htm

CDC (2020c, March 24). Public Health and Promoting Interoperability Programs (formerly, known as Electronic Health Records Meaningful Use). https://www.cdc.gov/ehrmeaningfuluse/introduction.html

CDISC (n.d.). CDISC Standards in the Clinical Research Process. CDISC. https://www.cdisc.org/standards

CDISC. (2020). Interim User Guide for COVID-19. CDISC. ttps://www.cdisc.org/standards/therapeutic-areas/covid-19

Davis, L (2020). Corona Data Scraper [HTML]. COVID Atlas. https://github.com/covidatlas/coronadatascraper

DDI Alliance (2020). Data Documentation Initiative. https://ddialliance.org/

DSI, D. S., CEF eHealth (2020, March 24). EHDSI INTEROPERABILITY SPECIFICATIONS, Requirements and Frameworks (normative artefacts)—EHealth DSI Operations—CEF Digital. https://ec.europa.eu/cefdigital/wiki/pages/viewpage.action?pageId=35210463

ECDC (2019). The European Surveillance System (TESSy). European Centre for Disease Prevention and Control. https://www.ecdc.europa.eu/en/publications-data/european-surveillance-system-tessy

ECDC (2020). Geographic distribution of COVID-19 cases worldwide [Dataset]. European CDC. https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide

ELIXIR (2020). ELIXIR. https://elixir-europe.org/about-us

ESIP (2019). Data Citation Guidelines for Earth Science Data , Version 2. Earth Science Information Partners. https://doi.org/10.6084/m9.figshare.8441816.v1

EU eHealth Network (2019). EHealth Network Guidelines to the EU Member States and the European Commission on an interoperable eco-system for digital health and investment programmes for a new/updated generation of digital infrastructure in Europe, https://ec.europa.eu/health/sites/health/files/ehealth/docs/ev_20190611_co922_en.pdf

European Commission (2016, November 25). European Reference Networks [Text]. Public Health - European Commission. https://ec.europa.eu/health/ern/networks_en

European Commission (2020a). Pseudonymisation tool. EUPID - European Platform on Rare Disease Registration. https://eu-rd-platform.jrc.ec.europa.eu

European Commission (2020b) Commission recommendation (EU) 2020/518 on a common Union toolbox for the use of technology and data to combat and exit from the COVID-19 crisis, in particular concerning mobile applications and the use of anonymized mobility data. https://ec.europa.eu/info/sites/info/files/recommendation_on_apps_for_contact_tracing_4.pdf

European Commission. (2020c). 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. European Union. https://www.rd-alliance.org/system/files/documents/H2020_Guidelines_COVID19_EC.pdf

European Commission (2020d). eHDSI Interoperability Specifications, Requirements and Frameworks (normative artefacts) - eHealth DSI Operations - CEF Digital, Mar. 24, 2020. https://ec.europa.eu/cefdigital/wiki/pages/viewpage.action?pageId=35210463

European Parliament and Council, Regulation (EC) No 851/2004 of the European Parliament and of the Council of 21 april 2004 establishing a European Centre for disease prevention and control, EUR-Lex - 32004R0851 - EN - EUR-Lex. 2004. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32004R0851

FDA Sentinel Initiative (2019). Sentinel Common Data Model | Sentinel Initiative. https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model

Finnie, T., South, A., & Bento, A (2016). EpiJSON: A unified data-format for epidemiology. Epidemics, 15(June, 2016), 20–26. https://doi.org/10.1016/j.epidem.2015.12.002

FitzHenry, F., Resnic, F. S., Robbins, S. L., Denton, J., Nookala, L., Meeker, D., Ohno-Machado, L., & Matheny, M. E (2015). Creating a Common Data Model for Comparative Effectiveness with the Observational Medical Outcomes Partnership. Applied Clinical Informatics, 6(3), 536–547. https://doi.org/10.4338/ACI-2014-12-CR-0121

FMSA, "Implementierungsleitfaden Meldung an das Epidemiologische Meldesystem (EMS)–Labor-und ArztmeldungZur Anwendung im österreichischen Gesundheitswesen [1.2.40.0.34.7.9.2.0]," Federal Ministry of Social Affairs, Health, Care and Consumer Protection, 2020. https://hl7.at/wp-content/uploads/2020/04/Implementierungsleitfaden-Meldepflichtige-Krankheiten_v02.20.pdf

Ghani, A. C., Donnelly, C. A., Cox, D. R., Griffin, J. T., Fraser, C., Lam, T. H., Ho, L. M., Chan, W. S., Anderson, R. M., Hedley, A. J., & Leung, G. M. (2005). Methods for Estimating the Case Fatality Ratio for a Novel, Emerging Infectious Disease. American Journal of Epidemiology, 162(5), 479–486. https://doi.org/10.1093/aje/kwi230

GIDA (2019). Global Indigenous Data Alliance, https://www.gida-global.org

Google. (2020). Google Dataset Search. https://datasetsearch.research.google.com/

Hamilton et al. (2020). PhenX Toolkit—COVID-19 Protocols. https://www.phenxtoolkit.org/covid19

Hausman, J., Stall, S., Gallagher, J., & Wu, M. (2019). Software and Services Citation Guidelines and Examples. ESIP. https://esip.figshare.com/articles/Software_and_Services_Citation_Guidelines_and_Examples/7640426

Healthcare Systems Research Network (2019). VDW Data Model. http://www.hcsrn.org/en/Tools%20&%20Materials/VDW/

Healy, Kieran (2020). Rpackage—COVID19 Case and Mortality Time Series (Version version 0.1.0) [R package]. https://kjhealy.github.io/covdata

HL7 (2020). HL7 Standards Product Brief—CDA® Release 2 | HL7 International. Retrieved April 21, 2020, from http://www.hl7.org/implement/standards/product_brief.cfm?product_id=7

HL7 (). "HL7 FHIR® Implementation Guide: Electronic Case Reporting (eCR) - US Realm." http://hl7.org/fhir/us/ecr/index.html

HL7 (). "HL7 Standards Product Brief - CDA® Release 2 | HL7 International." http://www.hl7.org/implement/standards/product_brief.cfm?product_id=7

HL7 (2019). "Fast Healthcare Interoperability Resources Standard," HL7®, Nov. 01, 2019. https://www.hl7.org/fhir/

IHE International (2019). "IHE Patient Care Device (PCD) Technical Framework, Volume 1, IHE_PCD_TF_Vol1.pdf," Dec. 12, 2019. https://www.ihe.net/uploadedFiles/Documents/PCD/IHE_PCD_TF_Vol1.pdf

IHE International (2019). "IT Infrastructure (ITI) Technical Framework," Integrating Healthc. Enterp., vol. 1, p. 334, 2019. https://www.ihe.net/uploadedFiles/Documents/ITI/IHE_ITI_TF_Vol1.pdf

IHME (2020). COVID-19 Projections. Institute for Health Metrics and Evaluation. University of Washington, Institute for Health Metrics and Evaluation https://covid19.healthdata.org/projections

IHME (2020). Global Health Data Exchange | GHDx. University of Washington, Institute for Health Metrics and Evaluation (IHME), http://ghdx.healthdata.org/

Johns Hopkins University (2020). COVID19 dataset [Dataset]. Johns Hopkins University, CSSEGISandData. https://github.com/CSSEGISandData/COVID-19

Knight, G., Dharan, N., & Fox, G (2016). Bridging the gap between evidence and policy for infectious diseases: How models can aid public health decision-making. Int J Infect Dis., 42, 17–23. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4996966/

National Institutes of Health, U.S. Department of Health and Human Services. (n.d.). Open-Access Data and Computational Resources to Address COVID-19. Open-Access Data and Computational Resources to Address COVID-19 | Data Science at NIH. Retrieved April 17, 2020, from https://datascience.nih.gov/covid-19-open-access-resources

New York Times (2020). Coronavirus (Covid-19) Data in the United States. https://github.com/nytimes/covid-19-data

NIH (2020). NIH Public Health Emergency and Disaster Research Response (DR2) COVID-19 Research Tools—Training Material. NIH Public Health Emergency and Disaster Research Response (DR2). https://dr2.nlm.nih.gov/

Observational Health Data Sciences and Informatics (2019). OMOP Common Data Model – OHDSI. https://www.ohdsi.org/data-standardization/the-common-data-model/

Pathak, E. B., Salemi, J. L., Sobers, N., Menard, J., & Hambleton, I. R (2020). COVID-19 in Children in the United States: Intensive Care Admissions, Estimated Total Infected, and Projected Numbers of Severe Pediatric Cases in 2020. Journal of Public Health Management and Practice, Publish Ahead of Print. https://doi.org/10.1097/PHH.0000000000001190

Patient-Centered Outcomes Research Institute (2020). PCORnet. The National Patient-Centered Clinical Research Network. https://pcornet.org/

PEPP-PT (2020). "Pan-European Privacy-Preserving Proximity Tracing." https://www.pepp-pt.org

"European Rare Disease Registry Infrastructure (ERDRI) | EU RD Platform," Apr. 13, 2020. https://eu-rd-platform.jrc.ec.europa.eu/erdri-description_en

Semantic Scholar (2020). CORD-19. https://pages.semanticscholar.org/coronavirus-research

The Atlantic (2020). The COVID Tracking Project [Dataset]. https://covidtracking.com/

Translational Cancer Research Network, "Translational Research – Defining the 'T's'", http://www.tcrn.unsw.edu.au/translational-research-definitions

U.N (2018). The Humanitarian Exchange Language (HXL). United Nations, Office for the Coordination of Humanitarian Affairs (OCHA), Centre for Humanitarian Data. https://hxlstandard.org/standard/1-1final/

U.N (2020). The Humanitarian Data Exchange (HDX). United Nations, Office for the Coordination of Humanitarian Affairs (OCHA), Centre for Humanitarian Data. https://data.humdata.org/

University of Maryland (2020, April 24). University of Maryland COVID-19 Impact Analysis Platform. COVID-19 Impact Analysis Platform. https://data.covid.umd.edu/

University of Oxford (2020a). COVID19 dataset [Dataset]. https://github.com/owid/covid-19-data

University of Oxford (2020b). COVID19 government response tracker [Dataset]. University of Oxford. https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker

University of Washington (2020). COVID19 data: Beoutbreakprepared. https://github.com/beoutbreakprepared

Westfall, J. M., Mold, J., & Fagnan, L. (2007). Practice-Based Research—"Blue Highways" on the NIH Roadmap. JAMA, 297(4), 403–406. https://doi.org/10.1001/jama.297.4.403

WHO (2020a). Novel Coronavirus (2019-nCoV) situation reports. World Health Organization. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

WHO (2020b, February). COVID-19 CRF • ISARIC. https://isaric.tghn.org/COVID-19-CRF/

WHO (2020c). Global surveillance for COVID-19 caused by human infection with COVID-19 virus: Interim guidance (WHO/2019-nCoV/SurveillanceGuidance/2020.6; p. 4). World

Health Organization. https://apps.who.int/iris/bitstream/handle/10665/331506/WHO-2019-nCoV-SurveillanceGuidance-2020.6-eng.pdf

WHO (2020d, March 23). WHO Global COVID-19 Clinical Platform Case Record Form (CRF). https://www.who.int/publications-detail/global-covid-19-clinical-platform-novel-coronavius-(-covid-19)-rapid-version

WHO (2020e, March 29). Modes of transmission of virus causing COVID-19: Implications for IPC precaution recommendations. https://www.who.int/news-room/commentaries/detail/modes-of-transmission-of-virus-causing-covid-19-implications-for-ipc-precaution-recommendations

WHO (2020f). Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). https://www.who.int/publications-detail/report-of-the-who-china-joint-mission-on-coronavirus-disease-2019-(covid-19)

WHO (2020g). Surveillance, rapid response teams, and case investigation. Coronavirus Disease (COVID-19) Technical Guidance. World Health Organization. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/surveillance-and-case-definitions

WHO (2020h). Survey tool and guidance: Behavioural insights on COVID-19. WHO Regional Office for Europe. http://www.euro.who.int/__data/assets/pdf_file/0007/436705/COVID-19-survey-tool-and-guidance.pdf?ua=1

WHO (2020b). WHO COVID-19 forms and checklists. http://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/novel-coronavirus-2019-ncov-technical-guidance/forms-and-checklists-english-and-russian

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Santos, L. B. da S., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1), 1–9. https://doi.org/10.1038/sdata.2016.18

World Bank (2020). Understanding the Coronavirus (COVID-19) pandemic through data [Dataset]. http://datatopics.worldbank.org/universal-health-coverage/covid19/

Worldometers (2020). COVID19 data [Dataset]. https://www.worldometers.info/coronavirus/

Zhang, Y. (2020). The Epidemiological Characteristics of an Outbreak of 2019 Novel Coronavirus Diseases (COVID-19)—China, 2020. Cina CDC Weekly. http://weekly.chinacdc.cn/en/article/id/e53946e2-c6c4-41e9-9a9b-fea8db1a8f51

# 7. Omics Sub-Group Guidelines

## 7.1 Sub-Group Focus and Description

For the purpose of this group, OMICS are defined as data from cell and molecular biology. For most of the data modalities, data can be deposited in existing deposition database resources. Many of these resources are now supporting specific COVID-19 subsets.
Within this scope, the group has prioritized recommendations on data that is already frequently associated with research on COVID-19.

## 7.2 Initial Sub-Group Guidelines

## 7.2.1 Generic Guidelines

This chapter addresses guidelines that are appropriate for all OMICS data types, and potentially also for data addressed in other chapters.

### 7.2.1.1 Guidelines for Researchers Producing Data

1. The FAIR data principles (Wilkinson et al. 2016) address a primary concern that has led to the formation of the group writing these guidelines: availability and re-usability of research data on COVID-19, in order to prevent unnecessary duplication of work. Many of the specific guidelines in this chapter (and others) address what can be done to make the data as FAIR as possible with a reasonable time investment. Some considerations during the Covid-19 pandemic are:
    1.1. Several of the FAIR principles call for rich metadata. Especially where data about human subjects is concerned it is not always possible to share such metadata in an open catalogue. Specifics can be found in guidelines for the individual data types as well as in the chapter on legal issues and ethics.
    1.2. In this time of urgency, making data FAIR should not unnecessarily slow down researchers collecting data. It is better to bring researchers collecting data with data stewards who can help with the FAIRification than to force everybody to learn how to do this themselves.
    1.3. We also need to encourage people to share what they have as-is without fear it is insufficient, and signal that help is needed
    1.4. A generic guideline for increasing FAIRness is to make sure data is made available in existing (certified (CoreTrustSeal Standards and Certification Board 2019)) repositories, rather than starting new local resources. Also, if a choice must be made, submission to domain-specific repositories is preferred over generic repositories and catalogs.
    1.5. Reusability of data requires documented provenance: When sharing any secondary data the generation of which involved comparison against other resources (examples for OMICS data are: reference sequences for mapping, GO annotations for expression analysis, pre-trained models for gene annotation), both the public availability of these used resources and unambiguous referencing of the used resources, including version numbers, should be ensured.
    1.6. Increase the reusability of data with consistent preprocessing: To increase the availability of data ready for analysis and integration, it may be prudent to agree

on a consistent approach to preprocessing OMICS data. This would be a second-phase step that should not unnecessarily slow down researchers collecting data.

2. The FAIR principles do not contain a push for open availability of the data, but they are often accompanied by the credo "as open as possible, as closed as necessary". In these times of the pandemic, this quest for openness gains even more importance. It is therefore critical to pursue "legal interoperability", which in this context practically means to use a CC0 waiver (Creative Commons n.d.) where possible, a CC-BY license (Creative Commons n.d.) if necessary, with a strong preference for not adding any other restrictions (RDA-CODATA Legal Interoperability Interest Group 2016). For more details on this, we refer to the guidelines of the Legal/Ethics WG.
3. Data reproducibility and increased trust in the shared data are important. This is covered in detail in the RDA good software practices section. Summarizing:
   3.1. Software, including scripts and applications that were used to process and analyse the data should be provided along with the data in the publication.
   3.2. The dependencies of the underlying software environment should also be provided with the data in the publication.

   A good example of these principles in action is the Galaxy platform (Galaxy Project 2020) (and associated training materials).

### 7.2.1.2 Guidelines for Funders

1. It is recommended that increased weighting is given in the grant review process to researchers who demonstrate best practice in open data and data reproducibility with their research outputs.
2. Require association of a project with a dedicated data steward, who can be specifically responsible for making data available in a FAIR and timely manner without interfering with the required pace of the research process.
3. Make sure that calls for projects clearly say that for COVID-19 data "timely" publication means "as soon as possible after it has been collected" and not "as soon as the publication has been accepted by the journal".
4. Be clear in the call for proposals that budget for professional data stewards in the project to help make the data more FAIR is eligible for funding.

### 7.2.1.3 Guidelines for Publishers

1. Require publishing of data underlying a study, in an even more timely manner than usual.
2. Make sure that the author recommendations prefer publishing of data in domain-specific repositories where findability is better than in generic or institutional repositories.

### 7.2.1.4 Guidelines for Policymakers

1. Put guidelines into place that give researchers ease of mind when licensing/sharing their data.
2. Promote use of domain-specific repositories instead of or as well as institutional repositories. Benefits of domain-specific repositories are that metadata standards are more likely enforced, assay result formats that promote re-use better supported, standardization of terms and ontologies eases re-use challenges.

### 7.2.1.5 Guidelines for Researchers

1. If you have any existing SARS-CoV, MERS-CoV or EBOV data that have not yet been made public, consider publishing that data now as it can be a useful reference.

2. Think early about systematic naming of filenames. Not thinking about it early enough is often the cause of a lot of extra work when the data is not stored in a database and researchers have to rename a large number of files manually at a later stage.
3. Document the computing time and resources required for data processing. This could help other researchers to assess the time and resources required for the pipeline, therefore to decide whether it is feasible to proceed with the local resources available.
4. When selecting a repository for submission of the data, priority should be given to domain-specific repositories over generic (e.g. institutional) repositories. Domain-specific repositories are easier to find, and often have better visualization and selection facilities for re-users of the data.
5. The repositories listed for deposition are also prime locations for locating existing data. Many now have dedicated sections for new as well as pre-existing data relevant to Covid-19 research.

### 7.2.1.6 Guidelines for Providers of Data Sharing Infrastructures

1. Perform validation that data confirms to recommended metadata/annotation standards in order to help researchers making their data as FAIR as possible.

## 7.2.2 Guidelines Addressing Specific Data Types or Resources

This section contains recommendations that are specific to different named types of research data in the omics field. They are mostly targeted towards researchers who are producing data, and to a minor extent to researchers looking to re-use existing data. Where guidelines are addressing a different audience, they are in recognizable subsections.

### 7.2.2.1 Recommendations for Virus Genomics Data

**Repositories**

There are several genomics resources that can be used to make virus genomics sequences available for further research. A curated list can be found in FAIRsharing (FAIRsharing 2020b) some specific examples are listed below.

1. We suggest that raw virus sequence data is stored in one of the INSDC archives (INSDC n.d.), as each of these is well known and openly accessible for immediate reuse without undue delays:
   1.1. DDBJ (FAIRsharing 2020a) Sequence Read Archive
   1.2. ENA (FAIRsharing 2015e), for submission documentation see (ENA-Docs 2020)
   1.3. NCBI SRA (FAIRsharing 2015k), for submission documentation see (NIH-NCBI 2020b)
2. For assembled and annotated genomes we suggest deposition in one or more of these archives:
   2.1. NCBI GenBank (FAIRsharing 2015j)
   2.2. DDBJ Annotated/Assembled Sequences
   2.3. ENA
   2.4. NCBI Virus (FAIRsharing 2015l), for submission documentation see (NIH-NCBI 2020a)
3. There are other archives suitable for genome data that are more restrictive in their data access; submission to such resources is not discouraged, but such archives should not be the only place where a sequence is made available.

4. Before submission of raw sequence data (e.g., shotgun sequencing) to INSDC archives it is necessary to remove contaminating human reads.

**Data and metadata standards**

A list of relevant data and metadata standards can be found in FAIRsharing (FAIRsharing 2020c), some specific examples are below.

1. We suggest that data is preferentially stored in the following formats, in order to maximize the interoperability with each other and with standard analysis pipelines:
    1.1. Raw sequences: .fastq (FAIRsharing 2015g); optionally add compression with gzip
    1.2. Genome contigs: .fastq if uncertainties of the assembler can be captured, .fasta (FAIRsharing 2015f) otherwise; optionally add compression with gzip
        1.2.1. De novo aligned sequences: .afa
    1.3. Gene Structure: .gtf (FAIRsharing 2015h)
    1.4. Gene Features: .gff (FAIRsharing 2015i)
    1.5. Sequences mapped to a genome: .sam (FAIRsharing 2015m) or the compressed formats .bam (FAIRsharing 2015a) or .cram. Please ensure that the used reference sequence is also publically available and that the @SQ header is present and unambiguously describes the used reference sequence.
    1.6. Variant calling: .vcf (FAIRsharing 2015o). Please ensure that the used reference sequence is also publically available and that it is unambiguously referenced in the header of the .vcf file, e.g., using the URL field of the ##contig field.
    1.7. Browser: .bed (FAIRsharing 2015b)
2. Consider annotating virus genomes using the ENA virus pathogen reporting standard checklist (European Nucleotide Archive 2020), which is a minimal information standard under development right now and the more general Viral Genome Annotation System (VGAS) (Zhang et al. 2019).
3. For submitting data and metadata relating to phylogenetic relationships (including topology, branch lengths, and support values) consider using widely accepted formats such as Newick, NEXUS and PhyloXML (Stoltzfus et al. 2012). The Minimum Information About a Phylogenetic Analysis (Lapp 2017) checklist provides a reference list of useful tree annotations.

## 7.2.3 Recommendations for Host Genomics Data

Host genomics data is often coupled to human subjects. This comes with many ethical and legal obligations that are documented in a separate chapter and not repeated here.

### 7.2.3.1 Generic Recommendations for Researchers

1. Data sharing of not only summary statistics (or significant data) but also raw data (individual-level data) will foster a build-up of larger datasets. This will eventually allow identifying the determinants of phenotype more accurately.

2. Especially for raw sequencing data make sure to include QC results and details of the sequencing platform used.
3. Common terminologies for reporting statistical tests (e.g with StatO (FAIRsharing 2015n)) enable reuse and reproducibility.
4. Researchers interested in HLA genomics are referred to the HLA COVID-19 consortium (HLA Covid-19 Consortium 2020).

## Repositories

Several different types of host genomics data are being collected for COVID-19 research. Some suitable repositories for these are:

1. Gene expression: A curated list can be found in FAIRsharing (FAIRsharing 2020d) some specific examples are listed below. To achieve load balancing, it is in general recommended to choose the respective regional repository for data deposition. Note that INSDC resources synchronize most for their data sets daily (ex: EGA/JGA/dbGaP, ArrayExpress/GEA/GEO).
    1.1. Transcriptomics of human subjects (i.e., requiring authorized access):
    - European Genome-Phenome Archive (EGA) (FAIRsharing 2015d) (if the data must be stored locally, EGA is working on a software package that can be installed locally and connects to the central metadata archive for findability)
    - DDBJ Genotype-phenotype Archive (JGA) (FAIRsharing 2018a)
    - NCBI Genotypes and Phenotypes Database (dbGaP) (FAIRsharing 2015c)

    1.2. Transcriptomics (cell lines/animals):
    - DDBJ (FAIRsharing 2020a) Sequence Read Archive
    - ENA (FAIRsharing 2015e), for submission documentation see (ENA-Docs 2020)
    - NCBI SRA (FAIRsharing 2015k), for submission documentation see (NIH-NCBI 2020b)

    1.3. Gene expression arrays:
    - NCBI GEO (FAIRsharing Team 2015b)
    - DDBJ GEA (FAIRsharing Team 2018)
    - EBI ArrayExpress (FAIRsharing Team 2015a)
2. Genome-wide association studies (GWAS): GWAS Catalog; EGA; GWAS Central
3. Adaptive Immune Receptor Repertoire sequencing (AIRR-seq)[1] data and annotations can be submitted to dedicated repositories: iReceptor Public Archive (FAIRsharing 2020e) or VDJServer (FAIRsharing 2018b). It is also possible to submit these data to general purpose repositories (SRA, Genbank), for this process there are detailed instructions (AIRR Community 2020).

## Data and metadata standards

1. Gene expression
    1.1. Transcriptomics.
        1.1.1. Preferred minimal metadata standard MINSEQE
        1.1.2. Preferred file formats (sequencing-based):
            1.1.2.1. Raw sequences: fastq (compression can be added with gzip)
            1.1.2.2. Mapped sequences: .sam (compression with .bam or .cram)
            1.1.2.3. Transcript count: TPM .csv
        1.1.3. Also see FAIRsharing using the query 'transcriptomics'
    1.2. Microarrays:
        1.2.1. Preferred minimal metadata standard: MIAME.
        1.2.2. Preferred file formats tab-delimited text, raw data file formats from commercial microarray platforms (Affymetrix, Illumina etc)

[1] Adaptive Immune Receptor Repertoire sequencing (AIRR-seq) samples the diversity of the immunoglobulins/antibodies and T cell receptors present in a host. The respective gene loci undergo random and irreversible rearrangement during lymphocyte development, therefore this data is fundamentally distinct from conventional genome sequencing.

2. Genome-wide association studies (GWAS):
    - 2.1. Preferred minimal metadata standard: MIxS
    - 2.2. Preferred file formats: for binary files: .bim .fam and .bed; for text-format files .ped and .map.
3. Adaptive Immune Receptor Repertoire sequencing (AIRR-seq).
    - 3.1. Preferred minimal metadata standards: MiAIRR
    - 3.2. Preferred file formats:
        - 3.2.1. AIRR repertoire metadata (formatted as .JSON or .YAML), AIRR rearrangements (formatted as .TSV)
    - 3.3. Also see FAIRsharing using the query 'AIRR'.

#### 7.2.3.2 Recommendations for Policymakers

1. Although there is a growing number of consortia for genetic determinants (https://www.covid19hg.org/), due to the high costs involved with advanced high-throughput genomics, data is disproportionately not available from Low and Middle Income Countries (LMICs) and from minority ethic populations in high income countries, thus leading to improper extrapolation of results to unrepresented population groups. A strong policy framework is required to facilitate research and encourage more inclusive participation.

#### 7.2.3.3 Recommendations for Funders

1. Due to the high costs involved with high-throughput genomics, little data is available from Low and Middle Income Countries (LMICs) and from minority ethic populations in high income countries, thus leading to improper extrapolation of results to unrepresented population groups. Research that improves the coverage could be worth preferential treatment for funding.

### 7.2.4 Recommendations for Structural data

**Repositories**

Several different types of structural data are being collected for Covid-19 research. Some suitable repositories for these are:

1. Structural data on proteins acquired using any experimental technique should be deposited in the wwPDB: Worldwide Protein Data Bank; a collaborating cluster of three regional centers at (1) Europe EBI PDBe (PDBe-KB) and The Electron Microscopy Data Bank EMDB; (2) USA RCSB PDB; and (3) Japan PDBj. Data submitted to either of these resources will be available through each of them.
2. A public information sharing portal and data repository for the drug discovery community, initiated by the Global Health Drug Discovery Institute of China (GHDDI) is the GHDDI Info Sharing Portal and includes the following: a) compound libraries including the ReFRAME compound library (the world's largest collection of its kind, containing over 12,000 known drugs), a diversity-based synthetic compound library, a natural product library, a traditional Chinese medicine extract library, b) Drug Discovery Cloud Computing System on Alibaba Cloud, c) Data mining and integration of historical drug discovery efforts against coronavirus (e.g. SARS/MERS) using AI and big data, d) Molecular chemical modeling and simulation data using computational tools.

**Locating existing data**

1. A community data repository and curation service for Structure, Models, Therapeutics, Simulations related computations for research into the SARS-CoV-2 / COVID-19 / ”Coronavirus” pandemic is maintained by The Molecular Sciences Software Institute (MolSSI) and BioExcel and can be found here.

**Data and metadata standards**

1. X-ray diffraction
   - 1.1. There are no widely accepted standards for X-ray raw data files. Generally these are stored and archived in the Vendor’s native formats. Metadata is stored in CBF/imgCIF format (See: catalogue of metadata resources for crystallographic applications)
   - 1.2. Processed structural information is submitted to structural databases in the .pdf or .mmCIF format.
2. Electron microscopy
   - 2.1. Data archiving and validation standards for cryo-EM maps and models are coordinated internationally by EMDataResource (EMDR).
   - 2.2. Every cryo-EM structure (map, experimental metadata, and optionally coordinate model) is deposited and processed through the wwPDB OneDep system (deposit.wwpdb.org), following the same annotation and validation workflow also used for X-ray crystallography and NMR structures. EMDB holds all workflow metadata while PDB holds a subset of the metadata.
   - 2.3. Most electron microscopy data is stored in either raw data formats (binary, bitmap images, tiff, etc.) or proprietary formats developed by vendors (dm3, emispec, etc.).
   - 2.4. Processed structural information is submitted to structural resources as PDBx/mmCIF.
   - 2.5. Experimental metadata include information about the sample, specimen preparation, imaging, image processing, symmetry, reconstruction method, resolution and resolution method, as well as a description of the modeling/fitting procedures used and are described in EMDR, see also Lawson et al 2020.
3. NMR
   - 3.1. There are no widely accepted standards for NMR raw data files. Generally these are stored and archived in single FID/SER files.
   - 3.2. One effort for the standardization of NMR parameters extracted from 1D and 2D spectra of organic compounds to the proposed chemical structure is the NMReDATA format.
   - 3.3. There is no universally accepted format, especially for crucial FID-associated metadata. The NMR-STAR is the archival format used by the Biological Nuclear Magnetic Resonance data Bank (BMRB), the international repository of biomolecular NMR data and an archive of the Worldwide Protein Data Bank (wwPDB 2018).
   - 3.4. The nmrML format specification (XML Schema Definition (XSD) and an accompanying controlled vocabulary called nmrCV) is an open mark up language an ontology for NMR data.
4. Neutron scattering
   - 4.1. The nuclear data evaluations have been separately released from different countries. ENDF/B-VI of Cross-Section Evaluation Working Group (CSEWG) and JEFF of OECD/NEA have been widely utilized in the nuclear community. The latest versions of the two nuclear reaction data libraries arer JEFF-3.3 in 2017 (Cabellos

et al., 2017) and ENDF/B-VIII.0 in 2018 (Brown et al., 2018) with a significant upgrade in data for a number of nuclides (Carlson et al., 2018).

4.2. Neutron scattering data are stored in the internationally-adopted ENDF-6 format maintained by CSEWG.

5. Molecular Dynamics (MD) simulations of SARS-CoV-2 proteins
    5.1. Raw trajectory files containing all the coordinates, velocities, forces and energies of the simulation are stored as binary files: .trr, .dcd, .xtc and .netCDF; See also a description of metadata standards to be considered.
    5.2. Refined structural models from experimental structural data.
6. Computer-aided drug design data
    6.1. Virtual screening results are stored in chemical data formats such as .mol/.sdf, .mol2, .pdb, SMILES, InChi.

## 7.2.5 Recommendations for Proteomics and Metabolomics

Proteomics and metabolomics studies are used to find biomarkers for disease and susceptibility. Lipidomics is a special form of metabolomics, but is also described in more detail in a separate section below because of its special relevance to COVID-19 research.

**Repositories**

1. For a curated list of relevant repositories see FAIRsharing using the query 'proteomics' and 'metabolomics'.
    1.1. Metabolomics/Lipidomics data can be submitted to MetaboLights (in Europe) or Metabolomics Workbench (USA)
    1.2. The ProteomeXchange Consortium provides: PRIDE, PeptideAtlas, MassIVE, jPOST | Japan Proteome Standard Repository/Database, iProX - integrated Proteome resources, Panorama
    1.3. Non-mass spectrometry based proteomics (ELISA, Luminex, ELISPOT, neutralizing antibody titer), Flow Cytometry, Mass Cytometry/CyToF and HLA/KIR typing data can be submitted to ImmPort.

**Data and metadata standards**

1. For a curated list of relevant standards see FAIRsharing using the query 'proteomics' and 'metabolomics'. Specific examples:
    1.1. Metabolomics/Lipidomics:
        1.1.1. CIMR standard, SMILES, InChl, ISA-Tab (MetaboLights)/mwTab (Metabolomics Workbench )
        1.1.2. Formats: for LC-MS data use: ANDI-MS, mzML; for NMR data: nmrCV, nmrML
    1.2. Proteomics
        1.2.1. use the minimal information model specified in MIAPE, and these are filled using the controlled vocabularies specified by the Proteomics Standards Initiative: PSI CVs
        1.2.2. formats: (gelML), TraML, mzML, mzTab, mzQuantML, mzIdentML
    1.3. Flow Cytometry
        1.3.1. Formats: .FCS / .ACS / .GatingML (ISAC standards)

## 7.2.6 Recommendations for Lipidomics

**Background**

Lipidomics reveal an altered lipid composition in infected cells (serum lipid levels in patients with preexisting conditions). Lipid rafts (lipid microdomains) play a critical role in viral infections facilitating virus entry, replication, assembly and budding. Lipid rafts are enriched in glycosphingolipids, sphingomyelin and cholesterol. It is likely that coronavirus (SARS-CoV-2) enters the cell via angiotensin-converting enzyme-2 (ACE2) that depends on the integrity of lipid rafts in the infected cell membrane.

### 7.2.6.1 Generic Recommendations for Researchers

1. Lipidomics analysis should follow the guidelines of the Lipidomic Standards Initiative

**Repositories**

1. The largest repository for lipidomics data is Metabolights

**Data and metadata standards**

1. Metadata standards
   - 1.1. Metadata should follow recommendations from the CIMR standard by the Metabolomics Standards Initiative. It should be made available as tab or comma separated files (.tsv or .csv).
2. Data standards
   - 2.1. Data can be stored in LC-MS file, tabular (.tsv) or comma (.csv) formats :
3. Data analysis
   - 3.1. Most of the analysis is usually performed using the software delivered by the suppliers of the instrumentation. In line with generic software recommendations it should be made sure that the process and parameters are well described, and that the output is converted to a standard format.
   - 3.2. Workflow for Metabolomics (W4M)
   - 3.3. R software packages from Bioconductor (xcms, camera, mixOmics)
4. Compound identification
   - 4.1. Manual identification using Lipid Maps tools
   - 4.2. Library templates for compounds identification
   - 4.3. LipidBlast
   - 4.4. MSPepSearch
   - 4.5. MS-DIAL
   - 4.6. Lipids classification and nomenclature should follow the LIPID MAPS guidelines

## 7.3 Additional Working Documents & Links

Additional materials from the Working Group can be found at: RDA-COVID19-Omics

## 7.4 References

Addshore, Mietchen D, Willighagen E, Yayamamo. SARS-CoV-2-Queries. 2020. https://egonw.github.io/SARS-CoV-2-Queries/. Accessed 4 May 2020.

AIRR Community. Adaptive Immune Receptor Repertoire - Data Commons API V1 - AIRR Standards 1.3.0 documentation. Adaptive Immune Receptor Repertoire (AIRR) - Common Repository Working Group (CRWG) - AIRR Data Commons API V1 — AIRR Standards 1.3.0 documentation. 2020. https://docs.airr-community.org/en/latest/api/adc_api.html. Accessed 24 Apr 2020.

AIRR Community. MiAIRR-to-NCBI Implementation. AIRR Standards 1.3.0 documentation. 2020. https://docs.airr-community.org/en/latest/miairr/miairr_ncbi_overview.html. Accessed 7 May 2020.

Artic Network. Artic Network: hCoV-2019 (nCoV-2019/SARS-CoV-2). Artic Network. 2020. https://artic.network/ncov-2019. Accessed 24 Apr 2020.

BBMRI-NL. Integrative Omics data set | BBMRI. Bio Banking Netherlands. 2020. https://bbmri.nl/services/samples-images-data/integrative-omics-data-set. Accessed 25 Apr 2020.

Berners-Lee T. Linked Data - Design Issues. Linked Data - Design Issues. 2009. https://www.w3.org/DesignIssues/LinkedData.html. Accessed 25 Apr 2020.

Broad Institute. Genotype-Tissue Expression (GTEx) Portal. The Broad Institute of MIT and Harvard. 2020. https://www.gtexportal.org/home/. Accessed 24 Apr 2020.

CNB, Segura JM. 3DBioNotes: Automated biochemical and biomedical annotations on Covid-19-relevant 3D structures. Centro Nacional de Biotecnología - Biocomputing Unit -. 2020. https://3dbionotes.cnb.csic.es/ws/api. Accessed 23 Apr 2020.

CoreTrustSeal Standards and Certification Board. CoreTrustSeal Trustworthy Data Repositories Requirements: Extended Guidance 2020–2022. 2019. doi:10.5281/zenodo.3632532.

COS. Coronavirus outbreak research collection. Center for Open Science. 2020. https://osf.io/collections/coronavirus/discover. Accessed 1 May 2020.

COVID-19 Biohackathon April 5-11 Participants, University of Manchester, HITS gGmbH. A COVID-19-specific instance for EOSC-Life’s WorkflowHub. The WorkflowHub. 2020. https://covid19.workflowhub.eu/. Accessed 23 Apr 2020.

COVID19-hg. COVID-19 Host Genetics Initiative. 2020. https://www.covid19hg.org/.

Creative Commons. CC BY 4.0 - Attribution 4.0 International. https://creativecommons.org/licenses/by/4.0/. Accessed 6 May 2020.

Creative Commons. CC0 1.0 - Universal. https://creativecommons.org/publicdomain/zero/1.0/. Accessed 6 May 2020.

DDBJ. Bioinformation and DDBJ Center. Bioinformation and DDBJ Center. 2020. https://www.ddbj.nig.ac.jp/index-e.html. Accessed 24 Apr 2020.

DNAstack. COVID-19 Beacon. COVID-19 Beacon. 2020. https://covid-19.dnastack.com/_/discovery?position=3840&referenceBases=A&alternateBases=G. Accessed 24 Apr 2020.

DTL. Personal Health Train. Dutch Techcentre for Life Sciences. 2018. https://www.dtls.nl/fair-data/personal-health-train/. Accessed 25 Apr 2020.

ELIXIR. COVID-19: The bio.tools COVID-19 Coronavirus tools list. bio.tools · Bioinformatics Tools and Services Discovery Portal. 2020. https://bio.tools/t?domain=covid-19. Accessed 23 Apr 2020.

EMBL-EBI. Expression Atlas: Gene Expression across species and biological conditions. European Molecular Biology Laboratory European Bioinformatics Institute. 2020. https://www.ebi.ac.uk/gxa/home. Accessed 25 Apr 2020.

EMBL-EBI. Pathogens: Surveillance, Identification Investigation. European Molecular Biology Laboratory - European Bioinformatics Institute. 2020. https://www.ebi.ac.uk/ena/pathogens/covid-19. Accessed 25 Apr 2020.

ENA-Docs. ENA Documentation. 2020. https://ena-docs.readthedocs.io/en/latest/. Accessed 6 May 2020.

European Molecular Biology Laboratory European Bioinformatics Institute (EMBL-EBI). EMBL-EBI COVID-19 Data Portal. 2020. https://www.covid19dataportal.org/. Accessed 30 Apr 2020.

European Nucleotide Archive. SARS-CoV-2 (Severe acute respiratory syndrome coronavirus 2) Submissions — ena-browser-docs latest documentation. 2020. https://ena-browser-docs.readthedocs.io/en/latest/help_and_guides/sars-cov-2-submissions.html. Accessed 27 Apr 2020.

European Nucleotide Archive. ENA virus pathogen reporting standard checklist. 2020. https://www.ebi.ac.uk/ena/data/view/ERC000033. Accessed 6 May 2020.

FAIRsharing. Binary Alignment Map Format. 2015. doi:10.25504/FAIRSHARING.HZA1EC.

FAIRsharing. Browser Extensible Data Format. 2015. doi:10.25504/FAIRSHARING.MWMBPQ.

FAIRsharing. Database of Genotypes and Phenotypes. 2015. doi:10.25504/FAIRSHARING.88V2K0.

FAIRsharing. European Genome-phenome Archive. 2015. doi:10.25504/FAIRSHARING.MYA1FF.

FAIRsharing. European Nucleotide Archive. 2015. doi:10.25504/FAIRSHARING.DJ8NT8.
FAIRsharing. FASTA Sequence Format. 2015. doi:10.25504/FAIRSHARING.RZ4VFG.
FAIRsharing. FASTQ Sequence and Sequence Quality Format. 2015. doi:10.25504/FAIRSHARING.R2TS5T.
FAIRsharing. Flow Cytometry Data File Standard. 2015. doi:10.25504/FAIRSHARING.QRR33Y.
FAIRsharing. Gating-ML. 2015. doi:10.25504/FAIRSHARING.QPYP5G.
FAIRsharing. Gene Transfer Format. 2015. doi:10.25504/FAIRSHARING.SGGB1N.
FAIRsharing. Generic Feature Format Version 3. 2015. doi:10.25504/FAIRSHARING.DNK0F6.
FAIRsharing. GWAS Central. 2015. doi:10.25504/FAIRSHARING.VKR57K.
FAIRsharing. Minimal Information about a high throughput SEQuencing Experiment. 2015. doi:10.25504/FAIRSHARING.A55Z32.
38. FAIRsharing. Minimum Information about Flow Cytometry. 2015. doi:10.25504/FAIRSHARING.KCNJJ2.
FAIRsharing. NCBI GenBank. 2015. doi:10.25504/FAIRSHARING.9KAHY4.
FAIRsharing. NCBI Sequence Read Archive. 2015. doi:10.25504/FAIRSHARING.G7T2HV.
FAIRsharing. NCBI Viral genomes. 2015. doi:10.25504/FAIRSHARING.QT5KY7.
FAIRsharing. Sequence Alignment Map. 2015. doi:10.25504/FAIRSHARING.K97XZH.
FAIRsharing. Statistics Ontology. 2015. doi:10.25504/FAIRSHARING.NA5XP.
FAIRsharing. Variant Call Format. 2015. doi:10.25504/FAIRSHARING.CFZZ0H.
FAIRsharing. Genome-Wide Association Studies Catalog. 2018. doi:10.25504/FAIRSHARING.BLUMRX.
FAIRsharing. Japanese Genotype-phenotype Archive. 2018. doi:10.25504/FAIRSHARING.PWGF4P.
FAIRsharing. Minimal information about Adaptive Immune Receptor Repertoire. 2018. doi:10.25504/FAIRSHARING.31HEC1.
FAIRsharing. VDJServer. 2018. doi:10.25504/FAIRSHARING.NZDQ0F.
FAIRsharing. AIRR Rearrangement File Format. 2020. https://fairsharing.org/bsg-s001474. Accessed 7 May 2020.
FAIRsharing. DNA Data Bank of Japan. 2020. doi:10.25504/FAIRsharing.k337f0.
FAIRsharing. FAIRsharing - Adaptive Immune Receptor Repertoire Resources. 2020. https://fairsharing.org/search/?q=AIRR. Accessed 7 May 2020.
FAIRsharing. FAIRsharing - Genomics Databases. 2020. https://fairsharing.org/search/?q=genomics&content=biodbcore. Accessed 15 Apr 2020.
FAIRsharing. FAIRsharing - Genomics Standards. 2020. https://fairsharing.org/search/?q=genomics&content=standards. Accessed 6 May 2020.
FAIRsharing. FAIRsharing - Transcriptomics Databases. 2020. https://fairsharing.org/search/?q=transcriptomics&content=biodbcore. Accessed 7 May 2020.
FAIRsharing. FAIRsharing - Transcriptomics Standards. 2020. https://fairsharing.org/search/?q=transcriptomics&content=standards. Accessed 7 May 2020.
FAIRsharing. FAIRsharing Collection: COVID-19 Resources. FAIRsharing Collection: COVID-19 Resources. 2020. https://fairsharing.org/collection/COVID19Resources. Accessed 22 Apr 2020.
FAIRsharing. iReceptor Public Archive. FAIRsharing. 2020. doi:10.25504/FAIRSHARING.EKDQE5.
Freunde von GISAID e.V. (“GISAID”). GISAID: Genomic epidemiology of hCoV-19. GISAID - Next hCoV-19 App. 2020. https://www.gisaid.org/epiflu-applications/next-hcov-19-app/. Accessed 24 Apr 2020.
GA4GH. Enabling responsible genomic data sharing for the benefit of human health. Global Alliance for Genomics and Health. 2020. https://www.ga4gh.org/. Accessed 25 Apr 2020.
GA4GH. GA4GH: Data Security Toolkit. Global Alliance for Genomics and Health. 2020. https://www.ga4gh.org/genomic-data-toolkit/data-security-toolkit/. Accessed 25 Apr 2020.

GA4GH. GA4GH: Genomic Data Toolkit. Global Alliance for Genomics and Health. 2020. https://www.ga4gh.org/genomic-data-toolkit/. Accessed 25 Apr 2020.

GA4GH. GA4GH: Regulatory & Ethics Toolkit. Global Alliance for Genomics and Health. 2020. https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/. Accessed 25 Apr 2020.

Galaxy Project. Best practices for the analysis of SARS-CoV-2 data: Genomics, Evolution, and Cheminformatics. COVID-19 analysis on usegalaxy. 2020. https://covid19.galaxyproject.org/. Accessed 23 Apr 2020.

GenBank. SARS-CoV-2 (Severe acute respiratory syndrome coronavirus 2) Sequences. U.S. Center for Disease Control. 2020. https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/. Accessed 15 Apr 2020.

GLOPID, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. Principles of data sharing in public health emergencies. 2018. https://www.glopid-r.org/wp-content/uploads/2018/06/glopid-r-principles-of-data-sharing-in-public-health-emergencies.pdf.

Hatcher EL, Zhdanov SA, Bao Y, Blinkova O, Nawrocki EP, Ostapchuck Y, et al. Virus Variation Resource - improved response to emergent viral outbreaks. Nucleic Acids Res. 2017;45:D482–90. doi:10.1093/nar/gkw1065.

HLA COVID-19 Consortium. HLA COVID-19. 2020. http://hlacovid19.org/. Accessed 7 May 2020.

INSDC. International Nucleotide Sequence Database Collaboration. http://www.insdc.org/. Accessed 6 May 2020.

Kovalsky A. COVID-19 workspaces, data and tools in Terra. COVID-19 workspaces, data and tools in Terra - Terra Support. 2020. http://support.terra.bio/hc/en-us/articles/360041068771. Accessed 24 Apr 2020.

Lapp H. Minimum Information About a Phylogenetic Analysis. GitHub. 2017. https://github.com/evoinfo/miapa. Accessed 7 May 2020.

LSRI. LSRI Response to COVID-19. European Life Science Research Infrastructure. 2020. https://lifescience-ri.eu/ls-ri-response-to-covid-19.html. Accessed 23 Apr 2020.

Majovski R. Broad scientists release COVID-19 best-practices workflows and analysis tools in Terra. Terra Support. 2020. http://support.terra.bio/hc/en-us/articles/360040613432. Accessed 24 Apr 2020.

Martinez-Martin N, Magnus D. Privacy and ethical challenges in next-generation sequencing. Expert Review of Precision Medicine and Drug Development. 2019;4:95–104. doi:10.1080/23808993.2019.1599685.

MPEG, the Moving Picture Experts Group., ISO/IEC JTC1/SC29/WG11. White paper on the objectives and benefits of the MPEG-G standard. 2018. https://mpeg.chiariglione.org/sites/default/files/files/standards/docs/w15047-v2-w15047_GenomeCompressionStorage.zip. Accessed 25 Apr 2020.

National Genomics Data Center. 2019nCovR - China National Center for Bioinformation. 2020. https://bigd.big.ac.cn/ncov?lang=en.

National Institute of Allergy and Infectious Disease (NIAID). Data Sharing and Release Guidelines. 2013. https://www.niaid.nih.gov/research/data-sharing-and-release-guidelines.

Nextstrain Team, Bedford T, Neher R. Nextstrain Genomic epidemiology of novel coronavirus. 2020. https://nextstrain.org/ncov/global. Accessed 15 Apr 2020.

NIAID DAIT. ImmPort Shared Data. ImmPort Shared Data. 2018. https://immport.org/shared/home. Accessed 15 Apr 2020.

NIH-NCBI. NCBI Virus: Severe acute respiratory syndrome-related coronavirus, taxid:694009. National Institutes of Health - National Center for Biotechnology Information. 2020. https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=Severe%20acute%20respiratory%20syndrome-related%20coronavirus,%20taxid:694009. Accessed 24 Apr 2020.

NIH-NCBI. NCBI Virus: Submit Sequences. National Institutes of Health - National Center for Biotechnology Information. 2020. https://www.ncbi.nlm.nih.gov/labs/virus/vssi/docs/submit/. Accessed 27 Apr 2020.

NIH-NCBI. Sequence Read Archive (SRA) Submission Quick Start. National Instutes of Health - National Center for Biotechnology Information. 2020. https://www.ncbi.nlm.nih.gov/sra/docs/submit/. Accessed 27 Apr 2020.

O'Donnell V, Wakelam M, Subramaniam S, Dennis E. LIPIDMAPS. 2020. http://www.lipidmaps.org.

Rambaut A. Virological: Novel 2019 coronavirus discussion forum. Virological. 2020. http://virological.org/c/novel-2019-coronavirus. Accessed 24 Apr 2020.

Rambaut A. Phylogenetic analysis of nCoV-2019 genomes. Edinburgh UK: University of Edinburgh; 2020. http://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356. Accessed 24 Apr 2020.

RCSB Protein Data Bank. RCSB Protein Data Bank SARS-CoV-2 Resources. 2020. https://www.rcsb.org/news?year=2020&article=5e74d55d2d410731e9944f52&feature=true. Accessed 23 Apr 2020.

RDA-CODATA Legal Interoperability Interest Group. Legal Interoperability of Research Data: Principles and Implementation Guidelines. Zenodo; 2016. doi:10.5281/zenodo.162241.

RDA-COVID19-Omics Subgroup. RDA-COVID19-Omics. RDA. 2020. https://www.rd-alliance.org/groups/rda-covid19-omics. Accessed 23 Apr 2020.

Renieri A. GEN-COVID: Impact of Host Genome on COVID-19 Clinical Variability. GEN-COVID. 2020. https://sites.google.com/dbm.unisi.it/gen-covid. Accessed 25 Apr 2020.

Shanghai Public Health Clinical Center & School of Public Health. Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome. Shanghai, China: Shanghai Public Health Clinical Center & School of Public Health; 2020. http://www.ncbi.nlm.nih.gov/nuccore/MN908947.3. Accessed 24 Apr 2020.

SOLID. Understanding Linked Data. https://solid.github.io/understanding-linked-data/#1. Accessed 25 Apr 2020.

Stoltzfus A, O'Meara B, Whitacre J, Mounce R, Gillespie EL, Kumar S, et al. Sharing and re-use of phylogenetic trees (and associated data) to facilitate synthesis. BMC Res Notes. 2012;5:574. doi:10.1186/1756-0500-5-574.

Swiss Institute of Bioinformatics (SIB). SARS-COV-2, COVID-19 Coronavirus Resource: SARS coronavirus 2 (SARS-CoV-2) proteome. SARS coronavirus 2 ~ ViralZone page. 2020. https://viralzone.expasy.org/8996. Accessed 23 Apr 2020.

Technical Committee : ISO/TC 215/SC 1 Genomics Informatics. ISO/TS 20428:2017 Health informatics — Data elements and their metadata for describing structured clinical genomic sequence information in electronic health records. ISO. 2017. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/79/67981.html. Accessed 25 Apr 2020.

Technical Committee : ISO/TC 276 Biotechnology. ISO/AWI 20688-2: Biotechnology — Nucleic acid synthesis — Part 2: General definitions and requirements for the production and quality control of synthesized gene fragment, gene, and genome. 2013. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/58/75852.html. Accessed 25 Apr 2020.

The Human Protein Atlas consortium. SARS-CoV-2 related proteins - The Human Protein Atlas. 2020. https://www.proteinatlas.org/humanproteome/sars-cov-2. Accessed 23 Apr 2020.

UniProt. COVID-19 UniProtKB. UniProt. 2020. https://covid-19.uniprot.org/uniprotkb?query=*. Accessed 23 Apr 2020.

Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:1–9. doi:10.1038/sdata.2016.18.

wwPDB consortium, Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, et al. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Research. 2019;47:D520–8. doi:10.1093/nar/gky949.

Zhang K-Y, Gao Y-Z, Du M-Z, Liu S, Dong C, Guo F-B. Vgas: A Viral Genome Annotation System. Front Microbiol. 2019;10. doi:10.3389/fmicb.2019.00184.

# 8. Social Sciences Sub-Group Guidelines

## 8.1 Sub-Group Focus and Description

Data from the social sciences is essential for all domains (including omics, clinical, epidemiology) seeking to better plan for effective management of COVID-19 and understanding its impact. Social scientists are collecting new information and reusing existing data sources to better inform leaders and policymakers about pressing social issues regarding COVID-19, to enable evidence-based decision-making. Key data types in the social sciences include qualitative; quantitative; geospatial; audio, image, and video; and non-designed data (also referred to as digital trace data). Recommendations made in these guidelines will help ensure that data contributions from the social sciences are shared in ways that allow them to be leveraged for the broadest impact and reused across all domains.

## 8.2 Initial Sub-Group Guidelines

The overall principle appropriate in times of public crises like COVID-19 is to allow the sharing of as much data as openly as possible and in a timely fashion, maintaining the public trust. This requires appropriate ethical and legal considerations. The following recommendations in relation to metadata, storage, sharing and ethical and legal issues should be referenced in making decisions which necessarily balance individual and public rights and benefits.

### 8.2.1 Data Management Responsibilities and Resources

1. Researchers should create a Data Management Plan (DMP) at the beginning of the research process when it can be included in the work plan and the budget and subsequently guide the handling of the data and help all disciplines understand the data. The DMP is a "living" document, which may change over the course of a project, and helps to document data for reuse and findability. Projects already underway that might contribute data to address COVID-19 should update their DMPs to ensure alignment with current recommendations. See examples of Data Management Plans (general and discipline-specific, including the Social Sciences); Data Management Plan Exemplar: Mixed Methods; ICPSR Guidelines for Effective Data Management Plans; CESSDA: Adapt your Data Management Plan. A list of Data Management Questions based on the Expert Tour Guide to Data Management (+ Word template); DMPonline, DMPTool, Portage DMP assistant and ARGOS for creating data management plans that meet institutional and funder requirements. Consult the European Commission Guidelines for open access to publications, data and other research outputs for Horizon 2020 projects working on the 2019 coronavirus disease (COVID-19), the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), and related topics, and address the relevant aspects of making the data FAIR in a DMP.
2. When writing the DMP, researchers should contact the repository of your choice which may offer guidelines for the DMPs in advance of deposit.
3. Researchers should aim to register their DMP as an openly accessible, public deliverable.

## 8.2.2 Documentation, Standards, and Data Quality

4. Researchers and statistical agencies should provide thorough documentation about the research context, methods used to collect data, and quality-assurance steps taken, as well as consider the minimal number of metadata variables shared that will allow linking the different types of data produced around COVID-19.
5. Given the multidisciplinary nature of pandemic research, social scientists should provide sufficiently detailed and explicit documentation, including metadata, such that data can be understood by researchers and be machine-actionable to allow for broad and interdisciplinary use. To do so, researchers should utilise community-endorsed metadata standards, controlled vocabularies and ontologies, and recommended file formats.
6. To encourage interdisciplinary research, social scientists should be mindful of commonly accepted professional codes or norms for documentation needs when producing documentation according to their own particular disciplinary norms. This allows for all domains to be able to ensure the research integrity of social science data it accesses or reuses.
7. To ensure that data is more generalizable, researchers should provide access to information that can be used to address selection bias (e.g., demographic characteristics, geographic variables).

## 8.2.3 Storage and Backup

1. Research institutions should provide researchers with robust and secure data storage facilities that follow recommendations regarding areas such as regular backup in multiple locations and data protection. Where possible, researchers should use the official storage provisions available from their institution, including when working remotely.
2. Data may have particular requirements as to how it can be stored and accessed, based on laws and regulations, research ethics protocols, or secondary data licenses. Sensitive data and human subject data containing personally identifiable information (PII) or protected health information (PHI) should be adequately protected and encrypted when at rest or in transit.
3. Where possible, best practice is to store data (including participant consent files) without direct identifiers and replace personal identifiers with a randomly assigned identifier. Researchers should create a separate file, to be kept apart from the rest of the data, which provides the linking relationship between any personal identifiers and the randomly assigned unique identifiers.

## 8.2.4 Legal and Ethical Requirements

1. Find a balance that takes into account individual, community and societal interests and benefits whilst addressing public health concerns and objectives to enable access to data and their reuse, and maximise the research potential.
2. It is recommended to establish rigorous approval mechanisms for sharing data (via consent, regulation, institutional agreements and other systematic data governance mechanisms). Researchers have a responsibility for ensuring research participants understand that there may be a risk of re-identification when data are shared.
3. Ethics review during a crisis like the COVID-19 pandemic is critical to protect highly vulnerable populations from potential harm. Therefore this report endorses guidance such as the Statement of the African Academy of Sciences' Biospecimens and Data Governance Committee On COVID-19: Ethics, Governance and Community engagement in times of crises.

4. Where possible, provide immediate open access to all relevant research data. Open data should be licensed under Creative Commons Attribution 4.0 International License (CC BY 4.0) or a Creative Commons Public Domain Dedication (CC0 1.0) or equivalent. If immediate open access is not possible, researchers should make data available as soon as possible. Researchers whose data have legal, privacy, or other restrictions should seek out appropriate alternative avenues for data sharing including restricted access conditions.
5. Ensure licenses and agreements in data acquisition enable downstream data sharing and preservation. If working with commercial partners, seek opportunities to negotiate data sharing mechanisms agreeable and equitable to all parties.

### 8.2.5 Data Sharing and Long-term Preservation

1. Ensure data shared is FAIR: Findable, Accessible, Interoperable and Reusable. In the current emergency context, it is a moral imperative to share the data and preserve it in the most open way possible for each case.
2. Select data for long term preservation; researchers should retain data that underpin published findings, data that allow for validation and replication of results, and the broader set of data with long-term value.
3. Deposit quality-controlled research data in a data repository, whenever possible in a trustworthy digital repository committed to preservation, such as one having undergone formal certification. As the first choice, disciplinary repositories are recommended for maximum visibility, followed by general or institutional repositories.
4. In order to expedite re-use, data that could be used to advance research on pandemics should be given top priority in the data publication process, fast-tracked by repositories, institutions, and other data publishers.
5. To ensure social sciences data can be linked with data being produced by other entities, consider preserving information that enables data linkages to be made, under appropriate security frameworks by creating a separate file, to be kept apart from the rest of the data, which provides the linking relationship between any personal identifiers and the randomly assigned unique identifiers.
6. Repositories should provide key metadata associated with its datasets, optimally utilising a metadata standard that allows for interoperability. They also should employ tools such as persistent identifiers for discovering and citing the data, as well as mechanisms for linking data and other research objects.
7. Researchers should make available and deposit with data in a repository all documentation--such as codebooks, lab journals, informed consent form templates--which are important for understanding the data and combining them with other data sources. Researchers should also make available information regarding the computing context relevant for using the data (e.g., software, hardware configurations, syntax queries) and deposit it with the data where possible.

## 8.3 Additional Working Documents & Links

For the complete guidelines please see RDA COVID-19 WG Guidelines - Social Sciences

Additional materials can be found at: RDA-COVID19-Social-Sciences

# 9. Overarching Research Software Guidelines

## 9.1 Focus and Description

It is important to put forward some key practices for the development and (re)use of research software, as these facilitate code sharing and accelerated results in responses to the COVID-19 pandemic. This section will be relevant to audiences ranging from researchers and research software engineers with comparatively high levels of knowledge about software development to experimentalists, such as wet-lab researchers, with almost no background in software development.

Seven clear, practical recommendations around basic software principles and practices are provided here, in order to facilitate the open and clear collaborations that can contribute to resolving current challenges. These recommendations aim to enable relatively small points of improvement across all aspects of software that will allow its swift (re)use, facilitating the accelerated and reproducible research needed during this crisis. These recommendations highlight key points derived from a wide range of work on how to improve your research software right now, to achieve better research (Wilson et al. 2017, Jiménez et al. 2017, Lamprecht et al. 2019; Akhmerov et al. 2019; Clément-Fontaine et al. 2019).

## 9.2 Initial Guidelines for Researchers

It is important to put forward some key practices for the development and (re)use of research software, as these facilitate code sharing and accelerated results in responses to the COVID-19 pandemic. This section will be relevant to audiences ranging from researchers and research software engineers with comparatively high levels of knowledge about software development to experimentalists, such as wet-lab researchers, with almost no background in software development writing scripts or macros.

Six clear, practical recommendations around basic software principles and practices are provided here, in order to facilitate the open and clear collaborations that can contribute to resolving current challenges. These recommendations aim to enable relatively small points of improvement across all aspects of software that will allow its swift (re)use, facilitating the accelerated and reproducible research needed during this crisis. These recommendations highlight key points derived from a wide range of work on how to improve your research software right now, to achieve better research (Wilson et al. 2017, Jiménez et al. 2017, Lamprecht et al. 2019; Akhmerov et al. 2019; Clément-Fontaine et al. 2019).

1. **Make your software available**
   Making software that has been developed available is essential for understanding your work, allowing others to check if there are errors in the software, be able to reproduce your work, and ultimately, build upon your work. The key point here is to ensure that the code itself is shared and freely available (see information about licenses below), through a platform that supports access to it  and allows you to effectively track development with versioning (e.g. code repositories such as GitHub, Bitbucket, GitLab, etc.).
   1.1. Resources:
      1.1.1. Four Simple Recommendations to Encourage Best Practices in Research Software

https://doi.org/10/gbp2wh

1.1.2. FAIR Software guidelines on code repositories https://fair-software.nl/recommendations/repository

2. **Reference your software with Persistent Identifiers (PIDs)**

Equally important to making the source code available is providing a means of referring to it (Cosmo et al. 2018). For this reason, software should be deposited within a repository that supports persistent identifiers (PIDs - a specific example being DOIs) such as Zenodo, Figshare or Software Heritage which provides more persistent storage than the above code repositories in R1.

2.1. Resources:

2.1.1. FAIR software guidelines on citing software https://fair-software.nl/recommendations/citation

2.1.2. List of software registries https://github.com/NLeSC/awesome-research-software-registries

2.1.3. Making your code citable through GitHub and Zenodo https://guides.github.com/activities/citable-code/

3. **Provide metadata/documentation for others to use your software**

(Re)using code/software requires knowledge of two main aspects at minimum: environment and expected input/output. The goal is to provide sufficient information that computational results can be reproduced and may require a minimum working example.

3.1. Resource:

3.1.1. Ten simple rules for documenting scientific software https://doi.org/10.1371/journal.pcbi.1006561

4. **Ensure portability and reproducibility of results**

It is critical, especially in a crisis, for software that is used in data analysis to produce results that can, if necessary, be reproduced. This requires automatic logging of all parameter values (including setting random seeds to predetermined values), as well as establishing the requirements in the environment (dependencies, etc). Container systems such as Docker or Singularity can replicate the exact environment for others to run software/code in.

4.1. Resource:

4.1.1. Ten Simple Rules for Writing Dockerfiles for Reproducible Data Science https://doi.org/10.31219/osf.io/fsd7t.

5. **Release your software under a licence**

Software code is typically protected by copyright in most countries, with copyright often held by the institution that does the work rather than the developer themself. By providing a licence for your software, you grant others certain freedoms, i.e. you define what they are allowed to do with your code. Free and Open Software licenses typically allow the user to use, study, improve and share your code. You can licence all the software you write, including scripts and macros you develop on proprietary platforms.

5.1. Resource:

5.1.1. Choose an Open Source License https://choosealicense.com/.

6. **Cite the software you use**

It is good practice to acknowledge and cite the software you use in the same fashion as you cite papers to both identify the software and to give credit to its developers. For software developed in an academic session, this is the most effective way of supporting its continued development and maintenance because it matches the current incentives of that system.

6.1. Resource:

6.1.1. Software Citation Principles https://doi.org/10.7717/peerj-cs.86.

### 9.2.1 References

Awesome list of Research Software Registries, https://github.com/NLeSC/awesome-research-software-registries.

Choose an Open Source License - https://choosealicense.com/.

Clément-Fontaine, Mélanie, Roberto Di Cosmo, Bastien Guerry, Patrick MOREAU, and François Pellegrini. 2019. “Encouraging a Wider Usage of Software Derived from Research.” Research Report. Committee for Open Science’s Free Software and Open Source Project Group. https://hal.archives-ouvertes.fr/hal-02545142.

FAIR Software - Five Recommendations for FAIR Software, https://fair-software.nl/.

GitHub Guides - Making your code citable, https://guides.github.com/activities/citable-code/.

Jiménez, Rafael C., Mateusz Kuzak, Monther Alhamdoosh, Michelle Barker, Bérénice Batut, Mikael Borg, Salvador Capella-Gutierrez, et al. 2017. “Four Simple Recommendations to Encourage Best Practices in Research Software [Version 1; Referees: 3 Approved].” *F1000Research* 6 (June): 876. https://doi.org/10/gbp2wh.

Lamprecht, Anna-Lena, Leyla Garcia, Mateusz Kuzak, Carlos Martinez, Ricardo Arcila, Eva Martin Del Pico, Victoria Dominguez Del Angel, et al. 2019. “Towards FAIR Principles for Research Software.” Edited by Paul Groth. *Data Science*, November, 1–23. https://doi.org/10.3233/DS-190026.

Lee BD Ten simple rules for documenting scientific software. PLoS Comput Biol 14(12): e1006561. 2018. https://doi.org/10.1371/journal.pcbi.1006561

Nüst, Daniel, Vanessa Sochat, Ben Marwick, Stephen Eglen, Tim Head, and Tony Hirst. 2020. “Ten Simple Rules for Writing Dockerfiles for Reproducible Data Science,” April. https://doi.org/10.31219/osf.io/fsd7t.

Smith, Arfon M., Daniel S. Katz, Kyle E. Niemeyer, and FORCE11 Software Citation Working Group. 2016. “Software Citation Principles.” *PeerJ Computer Science* 2 (September): e86. https://doi.org/10.7717/peerj-cs.86.

Wilson, Greg, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K. Teal. 2017. “Good Enough Practices in Scientific Computing.” *PLOS Computational Biology* 13 (6): e1005510. https://doi.org/10.1371/journal.pcbi.1005510.

Wilson, Scott. 4 Tips for Keeping on Top of Project Dependencies, https://osswatch.jiscinvolve.org/wp/2013/05/08/4-tips-for-keeping-on-top-of-project-dependencies/

## 9.3 Initial Guidelines for Policymakers

Research software is essential for research, and this is increasingly recognised globally by researchers. National and international policy changes are now needed to increase this recognition and to increase the impact of the software in important research and policy areas. This section provides recommendations for policy makers on how to support the research software community to respond to COVID-19 challenges, based on existing work (Akhmerov et al. 2019).

1. **Support the funding of development and maintenance of critical research software**
   Policy makers must continue to allocate financial resources to programs that support the development of new research software and the maintenance of research software that has a large user base and/or an important role in a research area.

1.1. Examples: UK Research and Innovation is funding COVID-19 related projects that can include work focussed on evaluation of clinical information and trials, spatial mapping and contact mapping tools (UK Research and Innovation 2020). Mozilla has created a COVID-19 Solutions Fund for open source technology projects (Mozilla 2020). USA's National Institutes of Health (NIH) provides "Administrative Supplements to Support Enhancement of Software Tools for Open Science" (NIH 2020b). The Chan Zuckerberg Initiative is funding open source software projects that are essential to biomedical research (Chan Zuckerberg Initiative 2020).

2. **Encourage research software to be open source and require it to be available**
Policy makers should enact policies that encourage provision of an open source license, or at least require it to be accessible. All research software should be released under a licence to ensure clarity of how it can be used and to protect the developers. The use of open source licences should be seen as the default for research software and policy makers should enact policies to encourage that practice. When software is made available under an open source license it means that its underlying code is made freely accessible, as encouraged by the "A" in FAIR (Findable, Accessible, Interoperable and Reusable) to users to examine and can be modified and redistributed. Through this process, software users can review, understand, improve, and build upon the software. As research outcomes rely on software, if software is not open source it must minimally be available for experimentation, to enable understanding of the software's functionality and properties and to reproduce the research outcomes. Whilst preprints and papers are increasingly openly shared to accelerate COVID-19 responses, the software and code for these papers is often not cited and hard to find, making reproducibility of this research challenging, if not impossible (Smith et al. 2016). Encouraging publishers to make software availability a default condition, together with the usually existing requirement for data availability, is an excellent way to greatly improve this.

2.1. Examples: The research community has been increasing access to key software and code, such as the Imperial College epidemic simulation model that is being utilised by government decision-makers. This was made publicly available with support by Microsoft to accelerate the process (Adam 2020).

3. **Encourage the research community's ability to apply best practices for research software, including training in software development concepts**
Policy makers should provide programs and funding opportunities that encourage both researchers and research support professionals (such as Research Software Engineers and Data Stewards) to utilise best practices to develop better software faster**.** In order to make research software understandable and reusable, it must be produced and maintained using standard practices that follow standard concepts, which can be applied to software ranging from researchers writing small scripts and models, to teams developing large, widely-used platforms. As research is becoming data-driven and collaborative in all areas, all researchers would benefit from the development of core software expertise, and research support professionals with these expertise also need to be increased. Policy makers should support inclusive software skills and training programs, including development of communities of learners and trainers.

3.1. Examples: There are various initiatives that link community members with specific digital skills to projects needing additional support, including Open Source Software helpdesk for COVID-19 ("COVID-19 OSS HELP" 2020) and COVID-19 Cognitive City (Grape 2020). Other initiatives aim to increase skills for engaging with software and code, such as the Carpentries, USA's NIH events (NIH 2020a); and the Galaxy Community and ELIXIR's webinar series (ELIXIR 2020).

4. **Support recognition of the role of software in achieving research outcomes** Policy makers should enact policies and programs that recognise the important role of research software in achieving research outcomes. It is important that policy makers encourage the development of research assessment systems that reward software outputs, alongside publications, data and other research outputs; and ensure that data and software management plans are a requirement in funding processes. It is also important that policy makers work to ensure these systems include proactive responses when these are not implemented.
   4.1. Examples: Policy makers need to support initiatives such as the Declaration on Research Assessment (DORA n.d.), which are beginning to be utilised by research agencies including the Wellcome Trust (Wellcome 2020), signatories to the Concordat to Support the Career Development of Researchers (Vitae 2020).

## 9.3.1 References

Adam, David. 2020. "Special Report: The Simulations Driving the World's Response to COVID-19." *Nature* 580 (7803): 316–18. https://doi.org/10.1038/d41586-020-01003-6.

Akhmerov, Anton, Maria Cruz, Niels Drost, Cees Hof, Tomas Knapen, Mateusz Kuzak, Carlos Martinez-Ortiz, Yasemin Turkyilmaz-van der Velden, and Ben van Werkhoven. 2019. "Raising the Profile of Research Software." https://doi.org/10.5281/zenodo.3378572.

Chan Zuckerberg Initiative. 2020. "CZI Launches Funding Opportunity for Open Source Software." *Chan Zuckerberg Initiative* (blog). April 30, 2020. https://chanzuckerberg.com/rfa/essential-open-source-software-for-science/.

"COVID-19 OSS HELP." 2020. April 30, 2020. https://covid-oss-help.org/.

DORA. n.d. "San Francisco Declaration on Research Assessment," n.d. https://sfdora.org/read/.

ELIXIR. 2020. "Galaxy-ELIXIR Webinar Series: FAIR Data and Open Infrastructures to Tackle the COVID-19 Pandemic | ELIXIR." April 30, 2020. https://elixir-europe.org/events/webinar-galaxy-elixir-covid19.

Grape, Derek. 2020. "Exaptive Partners with Bill & Melinda Gates Foundation, Launching The COVID-19 Cognitive City." April 30, 2020. https://www.exaptive.com/news/exaptive-partners-with-bill-melinda-gates-foundation-launching-the-covid-19-cognitive-city.

Mozilla. 2020. "MOSS Launches COVID-19 Solutions Fund." The Mozilla Blog. April 30, 2020. https://blog.mozilla.org/blog/2020/03/31/moss-launches-covid-19-solutions-fund.

NIH. 2020a. "NIH to Host Webinar on Sharing, Discovering, and Citing COVID-19 Data and Code in Generalist Repositories on April 24 | Data Science at NIH." April 30, 2020. https://datascience.nih.gov/news/nih-to-host-webinar-on-sharing-discovering-and-citing-covid-19-data-and-code-in-generalist-repositories-on-april-24.

———. 2020b. "NOT-OD-20-073: Notice of Special Interest (NOSI): Administrative Supplements to Support Enhancement of Software Tools for Open Science." April 30, 2020. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-20-073.html.

Smith, Arfon M., Daniel S. Katz, Kyle E. Niemeyer, and FORCE11 Software Citation

Working Group. 2016. “Software Citation Principles.” *PeerJ Computer Science* 2 (September): e86. https://doi.org/10.7717/peerj-cs.86.

UK Research and Innovation. 2020. “Get Funding for Ideas That Address COVID-19.” April 30, 2020. https://www.ukri.org/funding/funding-opportunities/ukri-open-call-for-research-and-innovation-ideas-to-address-covid-19/.

Vitae. 2020. “Concordat to Support the Career Development of Researchers — Vitae Website.” Landing page. April 30, 2020. https://www.vitae.ac.uk/policy/concordat-to-support-the-career-development-of-researchers.

Wellcome. 2020. “Outputs Management Plan - Grant Funding | Wellcome.” April 30, 2020. https://wellcome.ac.uk/grant-funding/guidance/how-complete-outputs-management-plan.

## 9.4 Initial Guidelines for Publishers

A key component of better research is better software. Publishers can play an important role in changing research culture, and have the ability to make policy changes to facilitate increased recognition of the importance of software in research. This section provides recommendations for publishers on how to support the research software community to respond to COVID-19 challenges.

1. **Require that software citations be included in publications**
   It is essential that the role of software in achieving research outcomes is supported. Treating research software as a first class research object in a journal is a very effective mechanism for implementing this as it increases the visibility and credit to the research software developers (for example by enabling academic and commercial citation databases, such as Google Scholar, Scopus and Microsoft Academic).
   1.1. Examples: The AAS Journals encourage software citation in a several ways (explicit software policy, added the LaTeX \software{} tag to emphasize code used, etc) (https://journals.aas.org/software-citation-suggestions/)
2. **Require that publications citing software have stored deposited the software within a repository that supports PIDs**
   For publishers to ensure that the research that they publish is reproducible, cited software must also be findable. Software being incorporated in a publication should be deposited in a repository that supports PIDs) such as Zenodo, Figshare or Software Heritage. These repositories provide PIDs that can be directly included in the citation and referenced in a publication, supporting research integrity (Cosmo, Gruenpeter, and Zacchiroli 2018 Cosmo, Gruenpeter, and Zacchiroli). If the software is deposited along with data, the selected data repository should provide a PID for the collection. Where a number of versions of a software have been used in a specific research study, having the version of the software used within a publication deposited in repositories that offer PIDs can help facilitate reproducibility of the work.
   2.1. Example: The JOSS review process requires authors to make a tagged release of the software after acceptance, and deposit a copy of the repository with a data-archiving service such as Zenodo or figshare (https://joss.theoj.org/about).

3. **Encourage the research community’s ability to apply best practices for research software, by aligning submission requirements to them**
   In order to make research software understandable and reusable, it must be produced and maintained using standard practices that follow standard concepts. This can be applied to software ranging from researchers writing small scripts and models; to teams

developing large, widely-used platforms. As publishing is an integral part of research, publishers should enact policies and adopt submission procedures that encourage and support these practices, for example through adopting or adapting software management statements similarly to the widely adopted data management statements.

3.1. Example: JOSS requires software to be Open Source and be stored in a repository that can be cloned without registration, is browsable online without registration, has an issue tracker that is readable without registration and permits individuals to create issues/file tickets (https://joss.theoj.org/about); SoftwareX submission process includes two mandatory metadata tables that include license and code availability (https://www.journals.elsevier.com/softwarex).

### 9.4.1 References

Cosmo, Roberto Di, Morane Gruenpeter, and Stefano Zacchiroli. 2018. "Identifiers for Digital Objects: The Case of Software Source Code Preservation." In , 1–9. https://doi.org/10.17605/OSF.IO/KDE56.

## 9.5 Additional Working Documents & Links

Additional materials from the Working Group available: RDA-COVID19-Software

"Good Enough Practices for Software to Help Others to Trust Your Analysis" tree - and a visualisation of this tree

# 10. Overarching Legal and Ethical Guidelines

The intention of these guidelines is to help researchers, practitioners and policy-makers deal ethically and legally with all aspects of pandemic response and in particular with regard to key ethical principles of *equity, utility, efficiency, liberty, reciprocity and solidarity*[1],[2]. In times of public health emergency, it is right to consider how best to respond in terms of increased data and research outcome sharing.  However, it is important that legal and ethical principles are incorporated into research design from the outset. The law supports research and enables data sharing[3]. Knowing compliance with the law protects individual researchers, research more generally and the common good. The rule of law cannot be overlooked, therefore, and needs to be taken into consideration along with respect for overarching concerns related to human rights and dignity[4]. Especially where marginalisation or other forms of stigmatisation are at stake, these rights and values should inform appropriate research practices directed towards the common good.

These guidelines have been produced by the RDA-COVID working group between March and May 2020 during the ongoing coronavirus disease (COVID-19) pandemic. The aim is to identify and collate existing recommendations and guidelines in order to increase the speed of scientific discovery by enabling researchers and practitioners to:

1. Readily identify the guidance and resources they need to support their research work
2. Understand generic and cross-cutting ethical and legal considerations
3. Appreciate country- or region-specific differences in policy or legal instruments
4. Identify the institutional stakeholders best placed to provide relevant ethical and legal guidance

## 10.1 Sub-Group Focus and Description

The COVID-19 pandemic has created significant confusion for researchers in terms of whether, and in which way, existing ethical and legal principles remain relevant.  The COVID pandemic does not serve to remove the basic validity of the rights and interests on which these documents and principles are based. The emergency does, however, mandate a reconsideration of the balance between these rights and interests - in particular between research subject's right to privacy and the public interest in the outcome of research. In some cases, this reconsideration has led to legitimate time limited adaptations of, or derogation from, normally applicable principles.

This document will therefore provide a high-level overview of:

1. Cross cutting principles
2. Hierarchy of obligations
3. Where to seek guidance
4. Initial Recommendations
5. Existing relevant policy statements

The assumption here is that there will be an official statement of when the international community deems the pandemic to have finished. This may then vary by country.
A separate, more detailed report expands on the information here.

## 10.2 Cross-cutting Principles

All activities, especially in times of pandemic or other public emergencies, should be guided by:

6. The FAIR (Findable, Accessible, Interoperable and Re-usable) principles of data to ensure ongoing, beneficial research[5];
7. The CARE principles to ensure ethical treatment of individuals and communities[6];
8. The Global Code of Conduct, specifically Fairness, Respect, Care and Honesty in research activities, to maximise equanimity in research outcome benefit[7];
9. The Five Safes of research data governance[8];
10. Research Integrity guidelines[9].

## 10.3 Hierarchy of Obligations

Ethics and the law exist in a symbiotic, mutually supportive relationship. Ethical and legal considerations related to research are elaborated in four key types of document: ethical guidelines; policy guidance; codes of conduct; and legal instruments. The distinction between these types of instrument is not always obvious. The following principles, therefore, may prove useful for COVID-19 researchers considering the interaction between instruments:

1. Ethical guidelines are often defined and publicised by non-law-making bodies, while legal instruments will be adopted by governments or other legislative bodies.
2. Many ethical instruments are mandatory for researchers or clinicians, such as those imposed by professional associations or bodies, healthcare institutions, or governmental and funding agencies.
3. Instruments exist in a hierarchy, with legal instruments being generally assumed to take precedence over ethical guidance and policy guidance.
4. Jurisprudence and other official guidelines providing authoritative interpretations of legal instruments will often be complementary to related ethical instruments.
5. Both legal and ethical instruments should be consulted together to understand all the pertinent issues which need to be taken into consideration.
6. Ethical instruments are generally interpreted harmoniously with the law, and can guide the interpretation of the law if the law does not address a particular issue.
7. Many ethical instruments are mandatory for researchers or clinicians, such as those imposed by professional associations or bodies, healthcare institutions, or governmental and funding agencies.

Common obligations in using health data that are found in many laws and ethical guidelines include the following:

1. The obligation to respect privacy and confidentiality
2. The obligation to ensure data accuracy
3. The obligation to use anonymised data instead of personal data, or minimise personal data use
4. The obligation to limit the identifiability of personal data as far as possible - including via pseudonymisation techniques
5. The need to process for a specific, authorized, purpose and only to process for secondary purposes provided certain conditions are fulfilled and not processing for purposes beyond scientific research / healthcare e.g. not sharing with employers or other agencies

6. To hold oneself accountable to, and remain transparent towards, the individuals concerned by the data used
7. To provide individuals access to their data, and to rectify errors or biases in the data on request
8. To provide individuals the opportunity to request the deletion or return of their data in certain circumstances if this is possible or required by law[10]
9. The obligation to ensure that data are collected from representative sub-populations and not confined to one group[11]
10. The obligation to ensure equanimity across cohorts to:
    - 10.1. Prevent marginalisation of vulnerable groups
    - 10.2. Encourage trust and engagement from vulnerable groups[12]
11. The obligation to share data and the benefits of research outcomes fairly and without regard to discipline, region or country[13]
12. The obligation to apply legal and ethical practice to all stages of data collection, processing, analysis, reporting and sharing
13. The obligation for data providers as well as data users to validate and verify the provenance of data, and ensure appropriate consent or other legal basis for the data's use
14. The obligation to ensure that de-identified or aggregated data made public does not contain data elements or rich metadata that could easily lead to identify specific persons
15. To include sunset clauses in the retention and exploitation of data collected during a public emergency with a view to future review of the continued use and usefulness of the data.

Such obligations are formalised through ethical guidance[14],[15],[16],[17]. Especially in times of pandemic specific attention to vulnerable groups and guidance on related global justice issues are to be commanded.

## 10.4 Seeking guidance

In times of pandemic or other public emergencies, it is important to be aware of existing and *ad hoc* resources and guidance. For example:

1. For researchers attached to an academic institution,
    - 1.1. the Institutional Review Board (IRB) or Research Ethics Committee (REC) will provide guidance as well as review
    - 1.2. the Information Governance Board will provide support on data management
    - 1.3. the Data Protection Officer will provide support and guidance on data protection issues
    - 1.4. Data and Biospecimen Access Committees will advise on sharing or providing access to data, as well as Intellectual Property issues
    - 1.5. Technology transfer offices provide guidance regarding intellectual property and related issues
2. For those in Low and Middle-Income Countries, if no local support is available, may contact the UN Ethics Office[18]
3. For professionals affiliated to a professional body, the latter will provide guidance on ethical research activities

4. For medical or other clinical staff, the institution (such as a hospital) will provide research integrity support, including ethical approvals required and an *ad hoc* mechanisms to support emergency research efforts; or the appropriate governing body (e.g., the NHS in the UK) will provide training and support both ongoing and in exceptional circumstances.
5. Hospitals, much like academic institutions, are often staffed by a Data Protection Officer, personnel specialized in research ethics including IRBs or REBs, and administrators responsible for authorizing the sharing of health data

Researchers and other professionals should always consult their institutional support personnel as well as professional bodies. Often in cases of health emergencies such as the COVID-19 pandemic fast track procedures are put in place, allowing the approval processes to be accelerated without diminishing the protection of the rights of persons.

## 10.5 Initial Recommendations

A set of recommendations are currently being collected and collated. They include:

1. Access to research and research outcomes should be shared with all
   1.1. in particular, thinking of vulnerable groups
   1.2. in particular, encouraging the engagement and trust of vulnerable groups
2. Ethical guidelines on data collection, analysis, sharing and publication should not be confined to clinical and biological (omic) data. Such guidelines should also extend to all areas of Open Science
3. In the spirit of the Open COVID Pledge[19], organisations with potentially useful datasets outside the research communities should be encouraged to make those data available to those research communities during emergency, pandemic situations
4. Ethical and legal policies should be drawn up to monitor and regulate the impact of algorithmic profiling and data analytics, not least in terms of design and implementation
5. During a pandemic or similar public emergency, ethical review and approval should be expedited, optionally but beneficially involving the public in approval decisions[20]
6. Policy making should be underpinned by empirical research (evidence based) such that decision makers are held to account
7. Provide guidance and support for non-research organisations to make the data they hold available to the research community
8. All stakeholders (researchers, policy-makers, editors, funders and so forth) should encourage communication across all disciplines and all areas in the spirit of Open Science

In the following version of this document, we also intend to identify areas where more research is needed.

## 10.6 Relevant policy and non-policy statements

The RDA Covid-19 Ethical-Legal group endorses and recommends guidance published as follows:

1. The OECD Privacy Principles (http://oecdprivacy.org/)
2. The UNESCO International Bioethics Committee (IBC) and World Commission on the Ethics of Scientific Knowledge and Technology (COMEST) in their STATEMENT ON COVID-19: ETHICAL CONSIDERATIONS FROM A GLOBAL PERSPECTIVE (https://unesdoc.unesco.org/ark:/48223/pf0000373115)
3. The Council of Europe pointers to national resources from national ethics committees or other related to COVID-19: https://www.coe.int/en/web/bioethics/covid-19
4. The Council of Europe statement on bioethics during COVID-19: https://rm.coe.int/inf-2020-2-statement-covid19-e/16809e2785
5. The European Group on Ethics in Science and New Technologies statement on solidarity https://ec.europa.eu/info/sites/info/files/research_and_innovation/ege/ec_rtd_ege-statement-covid-19.pdf
6. The Global Alliance for Genomics and Health (GA4GH) Framework for Responsible Sharing of Genomic and Health-Related Data (https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/framework-for-responsible-sharing-of-genomic-and-health-related-data/)
7. The Statement of the African Academy of Sciences' Biospecimens and Data Governance Committee on COVID-19: Ethics, Governance and Community engagement in times of crisis (https://www.aasciences.africa/sites/default/files/2020-04/COVID-19%20Ethics%2C%20Governance%20and%20Community%20engagement%20in%20times%20of%20crises%2020April2020.pdf)
8. Committee on Economic, Social and Cultural Rights, Statement on the coronavirus disease (COVID-19) pandemic and economic, social and cultural rights (https://www.ohchr.org/en/hrbodies/cescr/pages/cescrindex.aspx)

## 10.7 References

[1]https://apps.who.int/iris/bitstream/handle/10665/70006/WHO_CDS_EPR_GIP_2007.2_eng.pdf;
[2] https://unesdoc.unesco.org/ark:/48223/pf0000373115
[3]https://edpb.europa.eu/our-work-tools/our-documents/other/statement-processing-personal-data-context-covid-19-outbreak_en
[4] https://www.coe.int/en/web/human-rights-rule-of-law/covid19
[5] https://www.force11.org/fairprinciples
[6] http://www.newcarestandards.scot/?page_id=15
[7] https://www.globalcodeofconduct.org/
[8] https://www.ukdataservice.ac.uk/manage-data/legal-ethical/access-control/five-safes
[9]https://ec.europa.eu/research/participants/data/ref/h2020/other/hi/h2020-ethics_code-of-conduct_en.pdf
[10] Some EU Member States, for example, allow for data to be held indefinitely when used for scientific and research purpose
[11] E.g., the traditional white, Caucasian upper-middle-class male.
[12] See, for example, http://trust-project.eu/wp-content/uploads/2017/03/San-Code-of-RESEARCH-Ethics-Booklet-final.pdf
[13] https://unesdoc.unesco.org/ark:/48223/pf0000233230
[14] For example, the Caldicott principles (https://www.igt.hscic.gov.uk/Caldicott2Principles.aspx),
[15] https://www.echr.coe.int/Documents/Convention_ENG.pdf
[16] https://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/164
[17]https://en.unesco.org/themes/ethics-science-and-technology/bioethics-and-human-rights
[18] https://www.un.org/en/ethics/index.shtml
[19] https://opencovidpledge.org
[20] Cf. the Green / Amber / Red system of risk assessment applied in the UK