RESEARCH DATA ALLIANCE
rd-alliance.org

RDA COVID-19

Recommendations and Guidelines for data sharing

Executive Summary

Final version, 30 June 2020 –available at https://doi.org/10.15497/rda00052

Background

Data holds the potential to drive rapid response and informed decision-making during public health emergencies. There is a need for timely and accurate collection, reporting and sharing of data within and between research communities, public health practitioners, clinicians and policymakers. Accurate and rapid availability of data will inform assessment of the severity, spread and impact of a pandemic to implement efficient and effective response strategies.

The availability of efficient information and communication technology has improved the global capacity to implement systems to share data during a pandemic. However, the harmonisation across these sophisticated yet diverse systems combined with the timeliness of accessing data across information systems are currently major roadblocks. The World Health Organization'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 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 Research Data Alliance (RDA) COVID-19 Working Group (CWG) members bring varied global expertise to develop a body of work that comprises how data from multiple disciplines inform response to a pandemic combined with guidelines and recommendations on data sharing under the present COVID-19 circumstances. This extends to research software sharing, in recognition of the key role played by software in analysing data. The work has been divided into four research areas (Clinical, Omics, Epidemiology, Social Sciences) with four cross-cutting themes (Community Participation, Indigenous Data, Legal and Ethical Considerations, Research Software), as a way to focus the conversations and provide an initial set of guidelines in a tight timeframe. The detailed guidelines are aimed to help **stakeholders follow best practices to maximise the efficiency of their work, and to act as a blueprint for future emergencies**. The recommendations in the document are aimed at **helping policymakers and funders to maximise timely, quality data sharing and appropriate responses in such health emergencies**.

The CWG addressed 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 policymaking, including an overarching framework, common tools and processes, and principles that can be embedded in research practice**. The guidelines address general aspects of data practice, for example the FAIR principles (that research outputs should be Findable, Accessible, Interoperable, and Reusable), or the adoption of research-domain community standards.

There are foundational overarching challenges and recommendations that appear across the four research sub-groups as well as the cross-cutting themes. These foundational elements are presented in the summary before the area-specific challenges, recommendations and guidelines are articulated.

Challenges

The unprecedented spread of the virus has prompted **a rapid and massive research** response with a diversity of outputs that pose a challenge to interoperability. 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. 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.

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. Furthermore, many outputs are not reusable by, or useful to, different communities if they have not been sufficiently documented and contextualised, or appropriately licensed. Correspondingly, research software is developed and maintained in an *ad hoc* fashion. Access information for the software developed for analysis is not noted consistently in papers and, if the software is available, it is often placed in arbitrary locations with no guarantee of its persistence.

Recommendations

Governments, research funders, and research or research-supporting institutions around the world must coordinate with one another, and support and promote **Open Science** through policy and investment to streamline the flow of data between local entities, and across international jurisdictions.

There are motivational barriers to making data outputs available rapidly. There is a need for **incentivising the early publication/release of data outputs and the software used to produce them** during a public health emergency. The early publication/release of data outputs and the tools used to create them should be encouraged by building trust, providing incentives for sharing data and providing appropriate governance.

There is a need to **invest** in state-of-the-art information technology (IT) and data management systems infrastructure. The investment should also be directed towards people and skills to fully utilise the potential of large-scale infrastructure. The **minimum required infrastructure for pandemic response** in terms of technology, skills, people and frameworks should be accessible to all jurisdictions/sectors.

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. A balance between achieving 'perfectly' **FAIR outputs** and timely sharing is necessary with the key goal of immediate and open sharing as a driver. Data management plans (DMPs) should be created early in the research process and updated regularly to prepare for data deposit and reuse.

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.

Research outputs need to be documented, which includes **documentation** of methodologies used to collect, define and construct data, data cleaning, data imputation, data provenance and so on. 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.

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. By providing persistent identifiers, requiring preferred formats, rich metadata, etc., certified trustworthy repositories already guarantee a baseline FAIRness of and sustained access to the data, as well as citation.

Pre-print journals should undergo **an expedited review process** to balance the need to publish findings rapidly with the requirement to publish relevant and reliable findings. Full reports should be made available immediately upon communication of results, e.g. through a press release. Peer-reviewed data articles should be treated as first-class research outputs equal in value to traditional peer-reviewed articles. In order to **expedite reuse**, 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.

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.

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 related legislations** activated during a pandemic need to clearly outline data custodianship/ownership, publication rights and arrangements, consent models, and permissions around sharing data and exemptions.

The sub-groups and cross-cutting themes have each articulated the challenges facing researchers working on COVID-19, as well as recommendations/guidelines for improving data sharing (Table 1). These sub-group guidelines and recommendations should be considered directly depending on the relevant area of COVID-19 research as well as policy/decision-making.

Table 1 - Summary of challenges, guidelines and recommendations

Sub-groups/cross cutting themes	Challenges	Guidelines for researchers	Recommendations for funders/policymakers
Clinical	Promotion of clinical data sharing is important due to many studies and trials being performed under enormous time pressure	Standardised clinical terminologies should be used and a fair balance achieved between timely data sharing and protecting privacy and confidentiality	Measures should be taken in order to organise the sharing of data and trial documents in a suitable, trustworthy and secure data repository
Omics	An increased need of rapid openness for omics data to gain early insights into molecular biology of the processes at cellular level	Omics research should be a collaborative effort to learn the genetic determinants of COVID-19 susceptibility, severity and outcomes	Promote use of domain-specific repositories to enable standardisation of terms and enforce metadata standards

Sub-groups/cross cutting themes	Challenges	Guidelines for researchers	Recommendations for funders/policymakers
Epidemiology	Data and models are frequently incomplete, provisional, and subject to correction under changing conditions	Data models must include clinical data, disease milestones, indicators and reporting data, contact tracing and personal risk factors	Incentivise the publication of situational data, analytical models, scientific findings, and reports used in decision-making
Social Sciences	Require equal inclusion of social and economic context with health-related information to enable evidence-based decision-making	Enable interoperable cross-disciplinary and cross-cultural data collection, data use and collaboration for managing social sciences data during pandemics	Ensure robust funding streams for social sciences research for understanding and managing the human aspects of pandemics
Community	Need specific guidelines for enabling citizen scientists undertaking research to contribute to a common body of knowledge	Encourage public and patient involvement (PPI) throughout the data management lifecycle from research question to final data sharing and usage	Balance between timely testing and contact tracing, emergency response, community safety and individual privacy concerns
Indigenous Data Guidelines	Indigenous data rights, priorities and interests must be recognised in all COVID-19 research and surveillance activities	Indigenous governance of data collection, ownership, sharing and use priorities is the central principle of Indigenous data sovereignty	CARE Principles of Indigenous Data Governance set minimum standards for collectors, users and stewards of data
Legal and Ethical Considerations	Achieve a balance between rights of people and interests of researchers and policymakers	Ethical instruments should be interpreted with the law, and can guide the interpretation of the law if the law does not address a particular issue	During a pandemic, ethical review and approval for legally sharing data should be expedited
Research Software	Need systems in place for sharing of research software and accelerated and reproducible research during a pandemic	It is critical for software that is used in data analysis to produce results that can, if necessary, be reproduced	Funders must allocate financial resources to support the development and maintenance of new research software