RDA COVID-19 WG
Recommendations and Guidelines on Data Sharing

Second LAC Scientific Data Management Workshop

10 February 2021

Ingrid Dillo
RDA Council co-chair
Deputy Director, DANS Netherlands

CC BY SA

RDA COVID-19 Response

- Call to action European Commission using RDA as a forum to bring an international community of experts together
- fast track Working Group aimed at developing a system for **data sharing in public health emergencies**
- Recommendations and guidelines:
  - detailed **guidelines** to help **researchers** and **data stewards** follow best practices to maximise the efficiency of their work, and to act as a blueprint for future emergencies;
  - **recommendations** to help **policymakers** and **funders** to maximise timely, quality data sharing and appropriate responses in such health emergencies.

Community Effort

- Around 600 RDA members and newcomers registered
- 160+ active contributors to the documents
- Experts in the different fields as group moderators
- Wider engagement: weekly webinars, requests for comments

RDA's guiding principles:

- Openness
- Consensus
- Inclusive
- Harmonization
- Community-driven
- Non-profit and technology-neutral

A Collaborative Cross-Disciplinary Effort

CLINICAL

- Standardise terminologies, and find balance between timely data sharing and protecting privacy, confidentiality
- Organise data sharing and trial documents in trustworthy repositories

OMICS

- Select the best data formats and standards to fit the sub-discipline
- Promote use of domain-specific repositories to enable standardisation

EPIDEMIOLOGY

- Data models must include clinical data, disease milestones, indicators, reporting data, contact tracing and personal risk factors
- Incentivise publication of situational data, analytical models, scientific findings and reports

SOCIAL SCIENCES

- Enable interoperable cross-disciplinary, cross-cultural data use and collaboration
- Ensure robust funding streams for research aimed at understanding and managing the human aspects of the pandemic

COMMUNITY

- Encourage public and patient involvement throughout data management lifecycle
- Balance between timely testing and contact tracing, emergency response, community safety, and individual privacy concerns

RESEARCH SOFTWARE

- Software used in data analysis must be able to reproduce results, if necessary
- Allocate financial resources to support development and maintenance of new research software

INDIGENOUS DATA GUIDELINES

- Indigenous governance of data collection, ownership, and sharing and use priorities is the central principle of Indigenous data sovereignty
- CARE Principles set minimum standards for collectors, users, and stewards of Indigenous data.

LEGAL AND ETHICAL CONSIDERATIONS

- Although the law provides the foundation for data handling, ethical frameworks should also inform expedited approval to maximise data use and sharing
- Expedite ethical review and approval for legal data sharing during a pandemic

Output

- Structured through a series of teams - chairs & moderators, research areas, editorial, visualisation team, Zotero library team
- April 1 - 30 June continual sprints, consultations, webinars, 6 releases
- 143 pages with 4 page Executive Summary and many more outputs
- see: The Value of RDA for COVID-19 https://www.rd-alliance.org/value-rda-covid-19-0

Policy brief for Funders

https://www.rd-alliance.org/system/files/RDA%20COVID-19%20WG%20Recommendations%20to%20Funders.pdf

- RDA COVID-19 Recommendations and Guidelines for Data Sharing, final release, published 30 June 2020
- RDA COVID-19 Recommendations and Guidelines for Data Sharing Infographic
- RDA COVID-19 Guidelines and Recommendations - the prior 5 releases
- RDA-COVID19 WG Zotero Library
- DS Wizard Navigation Tool for the RDA COVID-19 Recommendations
- MindMap Navigation Tool for the RDA COVID-19 Recommendations

CC BY SA

What are the Challenges Being Faced?

Critical Need for Rapid Data Sharing

Rapid massive research response with diverse outputs challenges **interoperability of data.**

A trade off between...

Timeliness **Precision**

Lack of Harmonised Universal Standards and Context

Lack of pre-approved sharing agreements and archaic information systems hinder rapid threat detection and evidence-based response.

No universally adopted system or standard for

COVID-19 research outputs.

Lack of documentation, context, and appropriate licensing challenges **reusability.**

What are the Objectives?

1.0
Clearly define detailed guidelines on data and software sharing for COVID-19 research.

1.1
Help stakeholders follow best practices to **maximise efficiency.**

1.2
Act as a **blueprint** for future emergencies to maximise the efficiency of their work.

2
Develop **recommendations** for funders and policymakers to maximise timely, quality data and software sharing and appropriate responses in health emergencies.

3
Address interests of researchers, policymakers, funders, publishers, and providers of data sharing infrastructures.

Global Effort to Raise the Bar for Data Sharing

- 117 cross-sectoral signatories to the Wellcome Trust statement in January 2020.
- Agreement by 30 leading publishers on immediate open access to COVID-19 publications and underlying data.

What are the Key Recommendations?

The RDA COVID-19 Recommendations and Guidelines are aimed at developing a systematic approach 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.

1. Coordinate cross-jurisdictional efforts to foster global **Open Science** through policy and investment.
2. Incentivise early publication and release of data and software outputs.
3. Invest in state-of-the-art IT, data management systems **infrastructure, economies of scale, and people**.
4. Data, software and models should be **timely and FAIR**: **F**indable, **A**ccessible, **I**nteroperable, **R**eusable.
5. Require the use of **Data Management Plans**.
6. Use common generic as well as domain-specific **metadata standards**, and **persistent identifiers**.
7. Provide **documentation** of context, methodologies used to define, construct, and compile data, data cleaning and quality checks, data imputation, and data provenance.
8. Use **Trustworthy Data Repositories** committed to the long-term preservation and sustained access to their data holdings.
9. **Expedite** article and data review processes, **prioritising** and **fast-tracking data** at all stages.
10. **Balance ethics and privacy,** taking into account public interests and benefits while addressing the health crisis.
11. Access should be as **open as possible** and as **closed as necessary.**
12. Seek **technical solutions** that ensure anonymisation, encryption, privacy protection, and de-identification to **increase trust** in data sharing.
13. Provide **legal frameworks that promote sharing** of surveillance data across jurisdictions and sectors.

A Collaborative Cross-Disciplinary Effort

CLINICAL

- Standardise terminologies, and find balance between timely data sharing and protecting privacy, confidentiality
- Organise data sharing and trial documents in trustworthy repositories

OMICS

- Select the best data formats and standards to fit the sub-discipline
- Promote use of domain-specific repositories to enable standardisation

EPIDEMIOLOGY

- Data models must include clinical data, disease milestones, indicators, reporting data, contact tracing and personal risk factors
- Incentivise publication of situational data, analytical models, scientific findings and reports

SOCIAL SCIENCES

- Enable interoperable cross-disciplinary, cross-cultural data use and collaboration
- Ensure robust funding streams for research aimed at understanding and managing the human aspects of the pandemic

COMMUNITY

- Encourage public and patient involvement throughout data management lifecycle
- Balance between timely testing and contact tracing, emergency response, community safety, and individual privacy concerns

RESEARCH SOFTWARE

- Software used in data analysis must be able to reproduce results, if necessary
- Allocate financial resources to support development and maintenance of new research software

INDIGENOUS DATA GUIDELINES

- Indigenous governance of data collection, ownership, and sharing and use priorities is the central principle of Indigenous data sovereignty
- CARE Principles set minimum standards for collectors, users, and stewards of Indigenous data.

LEGAL AND ETHICAL CONSIDERATIONS

- Although the law provides the foundation for data handling, ethical frameworks should also inform expedited approval to maximise data use and sharing
- Expedite ethical review and approval for legal data sharing during a pandemic

Omics Example: 4.4.2 Guidelines for Host Genomics Data

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

1. **Gene expression data** should in general be retrieved from or deposited in the repositories listed below (Blaxter et al., 2016). To achieve load balancing, it is recommended to choose the respective regional repository. It should be noted that INSDC resources (i.e., DDBJ, ENA and NCBI) synchronise most of their datasets daily[2].
   - 1.1. Transcriptomics of human subjects (requiring authorised access):
     - 1.1.1. Database of Genotypes and Phenotypes (dbGaP) (Mailman et al., 2007)
     - 1.1.2. European Genome-Phenome Archive (EGA) (Lappalainen et al., 2015); the corresponding non-sensitive metadata will be available through EBI ArrayExpress (Athar et al., 2019)
     - 1.1.3. Japanese Genotype-phenotype Archive (JGA) (Kodama et al., 2015)
   - 1.2. Transcriptomics (from cell lines/animals):
     - 1.2.1. ArrayExpress (Athar et al., 2019)
     - 1.2.2. Gene Expression Omnibus (Barrett et al., 2013)
     - 1.2.3. Genomic Expression Archive
   - 1.3. Underlying reads can be retrieved from/will automatically be deposited to the corresponding read archive:
     - 1.3.1. DDBJ Sequence Read Archive (DRA) (Kodama et al., 2012), for submission documentation see here
     - 1.3.2. European Nucleotide Archive for submission documentation see here
     - 1.3.3. NCBI Sequence Read Archive (SRA) for submission documentation see here
   - 1.4. Microarray-based gene expression data:
     - 1.4.1. ArrayExpress (Athar et al., 2019)
     - 1.4.2. Gene Expression Omnibus (Barrett et al., 2013)

CC BY SA

Impact

- More reusable and reproducible research outputs = better science
- More efficient ways to build on research outcomes = faster solutions
- Better return on investment for funded research = better arguments for funding & more resources
- Research outputs with extensibility into other domains and contexts = advance on other infectious diseases/health emergencies

Actions

- Disseminate these guidelines to researchers and funders
- Build the expectation of FAIR data management aligned with these guidelines into research data management for infectious diseases
- new RDA Infectious Diseases Community of Practice
- Paper on community participation

Thank you for listening!

- Ingrid.dillo@dans.knaw.nl
- Website: www.rd-alliance.org
- Twitter: @resdatall
- Email: secretariat@rda-foundation.org