Research Data Alliance COVID-19 data sharing guidelines

Simon Parker – 9th September 2020

CANCER
RESEARCH
UK

Together we will beat cancer

Background

The Research Data Alliance

- The Research Data Alliance (RDA) was launched as a community-driven initiative in 2013 by the European Commission, the United States Government's National Science Foundation and National Institute of Standards and Technology, and the Australian Government’s Department of Innovation.

- The RDA has over 10k members from 145 countries (April 2020), and aims to provide a neutral space where members can come together to promote data-sharing and data-driven research, and develop a cohesive data community that integrates contributors across domain and national boundaries.

- Website: https://www.rd-alliance.org/
  Twitter: @resdatall

The Project

- At the end of March the Research Data Alliance (RDA) was asked by the EU Commission to produce guidelines on the sharing of COVID-19 data, recognising that access to data is vital to tackle and understand pandemics.

- As the guidelines were needed urgently, the RDA turned to its community to ask members to provide the required expertise.

- The guidelines are designed to be general principles and best practices that could be applied to any local legal context.

- The first draft was released on the 24th April, and the 5th version on the 28th May.

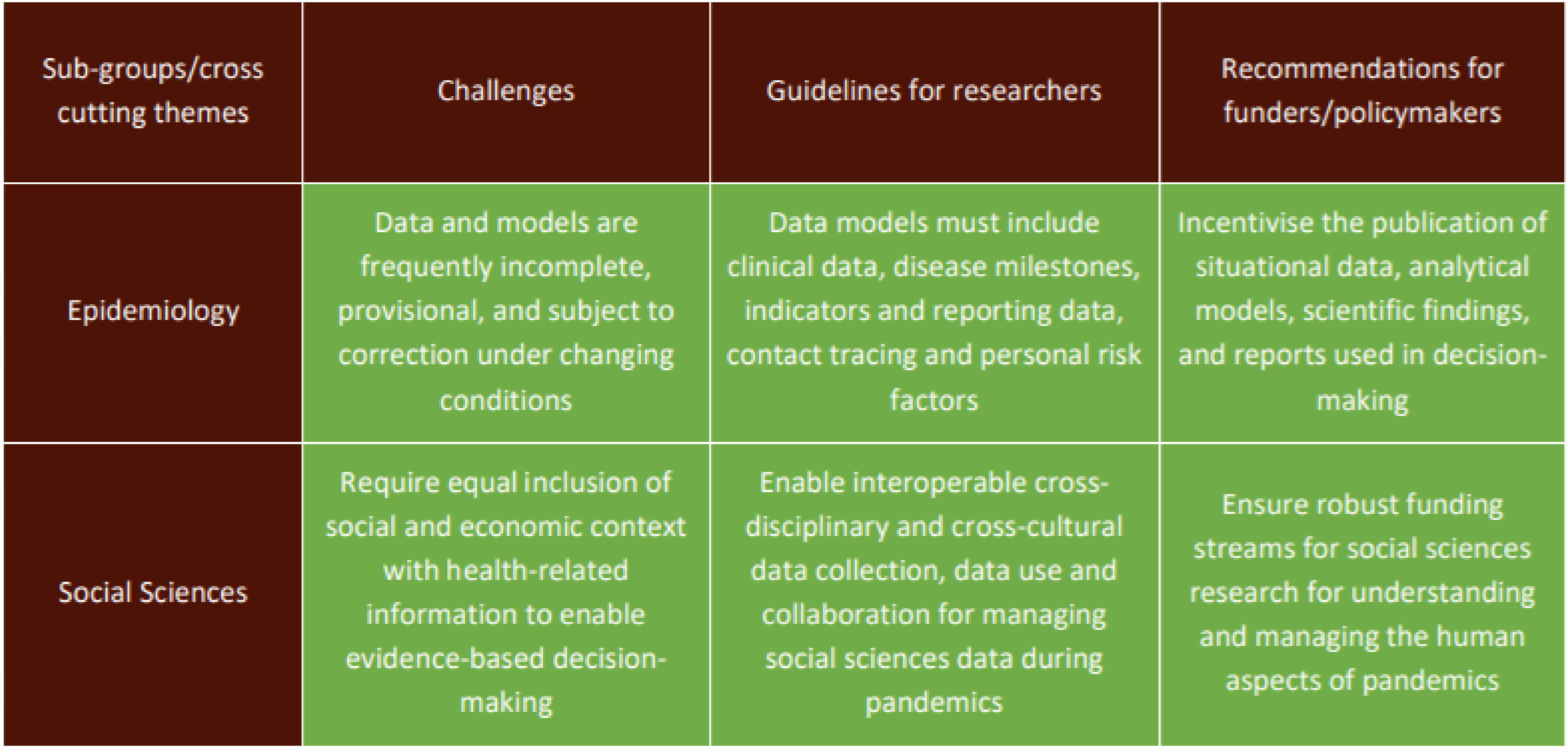
Guidelines

The Sections

- The guidelines contain domain-specific advice for:
  - Clinical Data
  - Omics Data
  - Epidemiology Data
  - Social Science Data

- There are four cross-cutting sections which can be applicable across domains:
  - Community Participation
  - Indigenous Populations
  - Research Software
  - Legal and Ethical Considerations

Executive Summary



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

6. Data Sharing in Social Sciences

6.4.3 Storage and Backup

- Where possible, researchers should avoid using personal storage, and instead use the official storage provisions available from their institution, including when working remotely, as they are more likely to provide robust backup and data protection features.

- Researchers with sensitive data or data with disclosure risk should seek a storage solution for their data which offers flexibility and protection, such as a solution offering remote access work (German Data Forum (RatSWD), 2020).

- Social sciences data, as is true for human subject data in other domains, may have particular requirements as to how it can be stored and accessed, based on laws and regulations, research ethics protocols, or secondary data licences that often vary by country.

- Data access while data collection is active should be limited to those with authorisation to use the data. To speed up access to COVID-related data, we encourage authorising external user groups where possible. 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, and no matter where or how it is stored.

- 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.

- Ensure that data should be backed up in multiple locations all under the same security conditions (See section on Infrastructure Investment).

- Where possible, select a storage solution that allows an easy way to maintain version control.

SDAP Mentioned

- **Statistical Disclosure Control** refers to methods used to reduce the risk of re-identification. They are encouraged when sharing or publishing data, and when publishing research outcomes (Willenborg et al., 2001; Griffiths et al., 2019).

- Griffiths, Emily, Carlotta Greci, Yannis Kotrotsios, Simon Parker, James Scott, Richard Welpton, Arne Wolters, and Christine Woods. "Handbook on Statistical Disclosure Control for Outputs," 2019. https://doi.org/10.6084/m9.figshare.9958520.v1.

Resources

- Research Data Alliance COVID-19 Recommendations and Guidelines Executive Summary
- Research Data Alliance COVID-19 Complete Recommendations and Guidelines and Zenodo version
- Zotero Resource Library

Future Developments

Ongoing Work

- Summary Paper – A summary of the work produced intended for publication.

- Community Participation Paper – This will focus on the process of developing the guidelines.

- Legal and Ethical group
  - Paper on international health data sharing after Schrems II
  - Paper on privacy by design
  - Paper on anonymity