Organisation for Economic Co-operation and Development

DSTI/STP/GSF(2021)13/FINAL

Unclassified **English - Or. English**

25 August 2021

DIRECTORATE FOR SCIENCE, TECHNOLOGY AND INNOVATION
COMMITTEE FOR SCIENTIFIC AND TECHNOLOGICAL POLICY

OECD Global Science Forum

Enhancing access to research data during crises: lessons learned from the COVID-19 pandemic

Summary note of a GSF-RDA workshop

This meeting was held on 23 April 2021 via Zoom.

This report provides a summary of the virtual GSF workshop co-organised with the Research Data Alliance (RDA). This workshop was a contribution to the *GSF project on Mobilising science in response to crises: lessons learned from COVID-19*.

The workshop presentations and video-recording are also available.

This report was declassified by CSTP on 14 July 2021.

Contact: Science, Technology and Policy Division (DSTI/STP), Carthage Smith (carthage.smith@oecd.org).

JT03480202

Enhancing access to research data during crises: lessons learned from the COVID-19 pandemic.

Summary note of a GSF-RDA workshop, held on 23 April 2021

1. Background: COVID-19 and data

1. Data access and re-use is critical for ensuring the efficiency and effectiveness of the global scientific research effort to address the COVID-19 pandemic. Access to trusted, well described data, together with the software, models and workflows that are necessary for the production and analysis of this data, is necessary across the many different domains of science that are being mobilised to understand and combat COVID-19. Transparent and timely access to research data and models underpins the science advice that is informing public health strategies. Such data and analytical tools have been critical to the rapid development and testing of vaccines and therapeutics. They are also essential for designing, monitoring and assessing the impact of the socio-economic policies that are being implemented in response to the pandemic.

2. The unprecedented spread of the virus has prompted a rapid and massive research response and this has been greatly facilitated by established international data sharing initiatives. However, such initiatives remain restricted to certain research domains and in many fields there are no universally adopted systems or standards, for collecting, documenting and disseminating COVID-19 research data and associated code and software. Many data are not reusable by, or useful to, different communities if they have not been sufficiently documented and contextualised or appropriately licensed. This is not a new challenge for but, in the context of COVID-19, it is a challenge that needs to be urgently addressed and for which, in many cases, solutions exist but have not been fully adopted at scale.

3. The responsible, FAIR and timely sharing of data is an essential element of the Open Science approach that the world needs to effectively combat pandemics like COVID-19 and other complex crises. Unnecessarily restricting or delaying access not only slows-up scientific progress but, as demonstrated in several countries, it can also undermine public trust in science and science-based decision-making. At the same time, much of the data related to the pandemic is personal or sensitive, and it is critically important to protect privacy and ensure the security, ethical management and use of this resource. There are important data gaps and biases and collecting and accessing representative data for certain population groups is very difficult.

4. Many international and national bodies have developed frameworks and guidelines to promote access to data and information relating to COVID-19. There was a commitment at the outset of the pandemic from a large number of research funding agencies, scientific associations and publishers on Sharing research data and findings rlevant to the novel coronavirus (COVID-19) outbreak. At the intergovernmental level, a revised OECD Recommendation on Access to Research Data from Public Funding was adopted by OECD Council in January 2021. At a more operational level the RDA-COVID-19 WG Recommendations and Guidelines for Data Sharing were published six months into the pandemic and provide detailed advice for different research domains. Policymakers, research funders, and research institutions around the world have been working together to implement these frameworks and guidelines, to mandate, incentivise and support actions

that harmonise and streamline the responsible and timely provision and exchange of data locally, nationally and internationally.

2. Workshop Description

5. This workshop was jointly organised by the OECD Global Science Forum (GSF) with the Research Data Alliance (RDA) and was recognised as co-located event with the RDA Plenary 17. It was a contribution to a broader GSF project on *Mobilising Science in Times of Crisis: lessons learned from COVID-19*. Presenters and panellists were selected in consultation with RDA and the Expert Group that was overseeing the GSF project. Over 150 people, including science policy makers, funders and data practitioners, attended the event

6. The workshop was structured around case examples from different research domains that are directly relevant to the pandemic. These cases were selected to be a mix of project level initiatives, national and international initiatives. A final moderated panel discussion brought together a group of policy makers and funders to reflect on what they had heard and provide their perspectives.

Aim: to explore key challenges and good practices to enhance research data access in response to the COVID-19 pandemic and identify lessons for science policy in the future.

7. Each of the three main workshop sessions and the final panel discussion were focused around 3 key questions:

1. What are the main challenges for access to research data in these fields in relation to COVID-19?
2. How are these challenges being addressed?
3. What is needed to be better prepared for future crises and what are the implications for science policy?

3. Introduction - Policy frameworks and guidelines

- **Carthage Smith, OECD Global Science Forum**
- **Natalie Harrower, Research Data Alliance**

8. The OECD Council approved a revised Recommendation on Access to Research Data from Public Funding in January 2021 as the world entered the second year of the COVID pandemic. This recommendation reaffirms the relevance and importance of several key principles that were originally set out in 2006: openness, flexibility, transparency, legal conformity, protection of intellectual property, formal responsibility, professionalism, interoperability, quality, security, efficiency, accountability, and sustainability. In addition, it provides updated policy guidance on seven areas that have emerged as crucial for enhancing access to research data in recent years: 1. Data governance for trust; 2. Technical standards and practices; 3. Incentives and rewards; 4. Responsibility, ownership and stewardship; 5. Sustainable infrastructures; 6. Human capital; and, 7. International cooperation on data access. Importantly, the scope of the recommendation has been extended to include access to algorithms, workflows, models and software (including code). Although the recommendation is not specifically targeted towards crisis response or public health emergencies, these seven focus areas are highly relevant to the pandemic.

9. Under the umbrella of the Research Data Alliance, a large international consortium of scientists, research managers and data specialists have been actively working together since the start of the pandemic to address the issues that are hindering data access across different scientific domains. The RDA-COVID-19 WG Recommendations and Guidelines for Data Sharing were published in June 2020 and provide detailed advice for different research domains. This includes guidance on data and metadata standards, controlled vocabularies, trustworthy repositories, data licensing and data documentation. It also provides concrete guidance on participant consent and protocols for managing personal data, as well as advice on overarching legal and ethical considerations.

10. The RDA guidelines identify a number of actions that are required throughout the research data lifecycle in order to enhance data access and re-usage and these can be mapped onto the seven main focus areas in the OECD recommendation. Building on its guidelines, RDA has joined with a number of other international bodies, in the Data together initiative and is working closely with research funders in the Global Research Collaboration for Infectious Disease Preparedness. It has also partnered with CODATA the World Data System and GO-FAIR, to develop the Virus Outbreak Data Network (VODAN), which focuses on making SARS-CoV-2 virus data FAIR and machine-readable.

4. Session 1 - Access to Biomedical and Clinical data

- Nevine Zariffa, Scientific Project Lead, International COVID-19 Data Alliance (ICODA), UK: *ICODA harnessing the power of health data*.
- Michael Brudno, Chief data scientist, University Health Network, Canada.
- Marie Paule Kieny, Director of Research, INSERM, France: *COVID-19 vaccine trials in France and the move to open science*

4.1. Nevine Zariffa

11. ICODA is a non-commercial initiative looking to promote data reuse for a greater societal impact, with a focus on randomised clinical trials (RCTs). It uses a standardised format to compare trials side by side and enable investigation of subpopulations, which is important to identify specific treatments for at risk subgroups (Precision public health approach). Thus far it has worked mainly on publicly available trial data sources and making those machine readable so that they can be used for meta-analyses. ICODA works with willing data partners. Thus far, there is commitment from 10 pharmaceutical companies but progress is slow, fewer than half a dozen datasets have so far been included in the enriched ICODA directory.

12. One of the challenges is the variation in data cycles. Investigators want to publish their findings first and then submit it to relevant regulatory authorities before sharing them more widely. There is a balance to be struck between protecting the intellectual property of those that provide the data and timely access to data for critical analyses to inform policy.

13. Improving data access and reuse is a shared responsibility. Researchers should design their studies from the outset to enable secondary access to data, using common standards and definitions. Funders and policy makers need to prioritise accountability to the public and patient groups and mandate and incentivise timely data access for secondary analyses. Realising the overall vision of an open and transparent research information infrastructure, which links all research outputs to individual contributors, depends on a

combination of trusted and accredited data sharing platforms with inclusive and transparent governance. Progress has been made in both of these domains during the current crisis but there is still a long way to go.

4.2. Michael Brudno

14. The Institute for Clinical Evaluative Sciences (ICES) is a non-profit, research organization that uses population-based health and social data to produce knowledge on a broad range of health care issues. It has headquarters in Ontario with nodes in several other Canadian cities and is linked to the University Health Network Canada and has access to high performance computing facilities. As part of this network, the Ontario Health Data Platform (OHDP) enables the use of artificial intelligence on COVID-19 data. OHDP has supported ~50 projects by providing expedited access to computing resource and data, including clinical and administrative data. These projects are aligned with the COVID-19 health research priorities of the Ministry of Health and provide timely, real world insights to inform policy. For example, an integrated analysis of social data, vaccination rates and clinical data at the district level in Ontario has provided critical insights that have informed the local vaccination strategy.

15. The future plan is to expand the OHDP and allow researchers to bring and connect to external datasets across Canada and with other countries. This will need to build on existing centers of excellence to move forward quickly. The challenge is to get the right datasets together. This is a situation in which researchers should be considered primarily as data consumers and not drivers of data collection; they need to align their research with policy needs and exploit available resources.

16. There are existing policies, infrastructures and data resources that can be used to answer important policy relevant research questions but these need to be better integrated. Currently there is a lot of unnecessary duplication and not enough sharing of data.

4.3. Marie Paule Kieny

17. Many different types of data are required to understand the pandemic and inform policy responses. This includes public health surveillance data, health service data and research data from different domains. Real time, or close to real time, data provision is the ideal, which means that with-holding critical research data until after publication is simply not appropriate. The pandemic does not recognise national borders and international sharing of data is critically important. In this context, the Global Initiative on Sharing Avian Influenza Data (GISAID) enables real time tracking of the virus and variants in its genomic sequence across the globe. Anyone who accepts the terms of the Database Access Agreement (DAA) can register and access GISAID data. According to the DAA, the original source of the data has to be acknowledged and collaboration with data providers is encouraged. Moreover, intellectual propert rights cannot be claimed on any fraction of the genetic sequence data obtained from GISAID, and data cannot be shared outside the GISAID registered community. This ‘ownership’ of data reflects the reality that many countries or institutions would not share the data that they collect unless there is some guarantee that their rights will be protected..

18. Data access and openness are also hot issues in relation to vaccine trials with a number of different initiatives being established with different access regimes and requirements. Here again the issue is not just about protecting IPR but often related to academic credit/prestige, which is sometimes takes precedent over urgent societal needs.

19. Having well accepted standards has been invaluable in building federated data sharing networks in genomics, including, for example, between Canada and EU through projects like CINECA (Horizon 2020).

20. The best way to respond to a pandemic is to have the infrastructure and ways of working already established and operating efficiently and effectively in a non-pandemic setting. This infrastructure must not be exclusive or overly restrictive and should be open and welcoming to new partners - allowing countries or institutions who might previously have decided not to share their data to reconsider and reverse that decision.

5. Session 2 - Omics research and epidemiology

- Niklas Blomberg , Director, Elixir, Europe
- Priyanka Pillai, Public Health, University of Melbourne, Australia
- Xihong Lin, Biostatistics, Harvard University and MIT, USA

5.1. Niklas Blomberg

21. Elixir is a European network for omics data provision and analysis and a critical contributor to the European COVID-19 data platform, which provides open access to different types of data from molecular platforms and the literature. This platform was rapidly created on a backbone of existing distributed infrastructures, which were already using common standards and protocols to provide open data. Linking genomic and protein structure data to generate refined data-sets has been very important in the scientific response to COVID-19. In this context, the importance of reference genomes, with unique identifiers and open licences cannot be overstated – they are essential for unambiguous identification of variants.

22. The European COVID-19 data platform was only possible because of the pioneering efforts that had been made over many years in the research community. The challenge was not to put something new in place but to connect existing infrastructure and expertise. In addition, making data available for reuse depends on having the appropriate licensing and governance arrangements in place and again these need to be established over time as part of routine research practice. The research data system needs to be prepared for crisis response before a crisis happens.

5.2. Priyanka Pillai

23. Priyanka presented the challenges, from a researcher perspective, of trying to ‘clean’ and link routine health surveillance data with administrative data and research data from different domains in order to provide an evidence base for policy development. Harmonisation across sophisticated yet diverse systems is a major challenge. Starting within individual research domains there is a need for standardised tools methods and technologies and then, on top of this, there is need to ensure at least minimal compatibility across domains. Cataloguing what already exists is more important than developing major new initiatives.

24. The integration of epidemiological data and omics data is critical for the management of infectious disease pandemics and the vision should be for a global system that is inclusive of all countries and population groups and provides open access wherever this is applicable, i.e. ensures that data are “as open as possible, as closed as necessary”.

Achieving this requires national and international networks of public health professionals and researchers working in partnership.

5.3. Xihong Lin

25. Xihong Lin gave her reflections, as someone who has been involved in international research projects from outset of the original outbreak in Wuhan, on COVID-19 data and resources and their usage to provide the best possible evidence on the course of the pandemic and inform public health policy. The scientific response to COVID-19 was a good example of the benefits, and also the current limitations, of open science and data sharing. Infrastructures and international networks, such as GISAID and Nextstrain, have been essential for tracking the evolution of the virus. There are also a number of well organised open repositories, such as Our World in Data, for aggregated data on deaths, cases and vaccination. Open access resources enable secondary analysis and the development of tools, such as real-time maps, to inform media, citizens and policy makers. However, there are persistent problems with sample bias and ensuring timely access. Many ‘volunteers’ are making great efforts to provide information and tools but there is a lack of coordination and considerable heterogeneity in data collection and management. There is also an overall shortage of disaggregated data that has made it difficult to identify and analyse what is happening in some of the most vulnerable population groups.

26. Industry collects a huge amount of data that can give insights into the pandemic and public private partnerships, e.g. to develop ‘apps’ for citizen science projects, can be particularly valuable. However, traditional research funding grants are generally not well suited to support such projects. There is a need for a global alliance for COVID-19 (or pandemic) data, resources and communication that brings together academia, government, industry and the media to co-create frameworks and standards to enable voluntary, responsible, and secure sharing of COVID-19 data.

6. Session 3 - Social sciences and interdisciplinary research

- Stefania Milan, Department of Media Studies, University of Amsterdam, Netherlands
- Katja Mayer, Science and Technology Studies, University of Vienna, Austria: Engaging with pandemic knowledge: open social science and policy
- Yukio Ohsawa, School of Engineering, University of Tokyo, Japan: "Stay with Your Community" - reducing the risk of the infection of COVID-19

6.1. Stefania Milan

27. The key theme of Dr Milan’s presentation was the importance of changing the ways that the sciences, use, create and exchange data for social analysis. The Covid-19 pandemic is an instance in which the main sources of data, and access to it, are inherently discriminatory. Whether one talks about sources or access, there are significant elements and aspects of society that are ignored. According to the WHO, several months after the start of the pandemic, only two countries in Africa had access to Covid-19 information for their population.

28. Significant portions of the international system are ‘excluded’ from data. Hence, there is no data that will help central and local government authorities in many developing countries to initiate essential preparedness and response measures and programmes. Even

in developed countries, data discrimination persists. Migrants, for example, may not have any access to health care in their host countries, which excludes them from health records. This confounds the data that authorities use, and means that 'official data' is often fundamentally flawed.

29. One consequence of data poverty is the way that even unintended data discrimination impoverishes the information produced by the sciences. Those responsible for collecting data need to understand that by allowing large swathes of society to fall outside their research perspectives, their data and subsequent analyses may be flawed.

30. When it comes to the relationship between interdisciplinary research, data and their perceived utility, attitudes of the general public should also be taken into account. Though the outcomes of research and subsequent data are normally highly specialised, social scientists should have capacities to make them more accessible to a general audience. In a related vein, social scientists and policy makers need to understand the contexts and sources of the data that each generates and use.

6.2. Katja Mayer

31. 'What gets counted, counts!' was a theme that in various ways underpinned both this and the previous presentation. The sheer spectrum of social science research presents a formidable challenge in developing a more open social science endeavour. Data in the social sciences reflect a field that is 'super diverse', more often than not contextualised, very fragmented, frequently local and multilingual dependent.

32. At the same time, social scientists have to work with a wide range of data that have not been generated for scientific purposes and over which they have limited control. Much of the data that they use are highly sensitive. These realities underscore the challenges which researchers face when attempting to fulfil their responsibilities for ensuring data privacy and research integrity. In an overarching sense, it is difficult 'to make sense of a world that is already making sense of the world.' That said, there are specific issues that the social sciences have to consider with regards to evidence needs for policy.

33. In the first place, one needs to be fast but not rush. In other words, both researchers and those concerned with policy need to be more anticipatory and adaptive. The mechanics for more transparent peer review of evidence sources need to be in place. Clearly related is the need to ensure that the capacities for dealing with data stewardship and analysis are established, and increasingly these capacities have to bring together private and public sectors.

34. The issue of coordination between policymakers and researchers remains fundamental as does the importance of linking researchers across disciplinary and geo-political boundaries. Combining multiple perspectives is crucial for social science, and there should be greater efforts to create 'sharing infrastructures' that link national policy needs to research and *vice versa*. Similarly, there need to be far better systems and arrangements for sharing social information across international borders. This means addressing power differentials in order to ensure appropriate access to data and data sources.

35. It is important to take a systems perspective. As complexity, global interconnectedness and the prospect of emerging global threats increase, so, too, does the need for systems that will ensure enduring links between data, the social sciences and other disciplines. There already are a variety of national and international social science networks that exist for exchanging data. However, it is evident that the types of data that are exchanged as well as the constructs for doing so need to be developed or in some instances

strengthened. The social sciences and humanities can surely learn from other domains and communities about ways to engage more effectively across disciplines and borders and it is important that they engage in international fora such as the Research Data Alliance.

6.3. Yukio Ohsawa

36. A consistent theme that emerged from this session was how best to maximise the use of data. The integrity of social science data, the interchange between policymakers and researchers as well as data management infrastructures and processes are all important in this regard. A 'data jacket' is digested information on a data-set that includes key meta-data and the subjective expectation of the data owner or potential data users about the utility of the data. Data jackets can play a role in ensuring the multidimensional use of data and enabling data be shared, integrated and analysed in ways that provide contextually-specific information for contextually specific problems. They provide a tool for innovative collaboration, where existing pieces of knowledge are collected that are relevant to a problem domain and which can then be used to address specific questions from different stakeholders

37. Multidisciplinary perspectives and data integration are needed to address complex problems. In the context of integrating data from different sources, there is an important distinction between data owners and data users. The former are generally concerned with 'flow' where the latter are generally concerned with its relevance in particular contexts. The two, however, can come together as 'stakeholders' and problem solvers. In this way multiple stakeholders can share data framed for externalising and solving problems that they potentially share.

7. Session 4 - Roundtable Discussion: National and international policy perspectives

- Camilla Stoltenberg, Director General, Institute of Public Health, Norway.
- Kazuhiro Hayashi, Research Unit for Data Appication, NISTEP, Japan
- Yazdan Yazdanpanah, Director of ANRS Maladies Infectieuses Emergentes, France.
- Claudia Bauzer Medeiros , Institute of Computing, Univerity of Campinas and FAPESP, Brazil.
- Michal Kahn, Centre for Research on Evaluation, Science and Technology, Stellenbosch University, South Africa.
- Kiwon Jang, Korea Bioinformation Center / Korea Research Institute of Bioscience and Biotechnology.
- Steven Kern, Gates Foundation.
- Konstantinos Repanas, Open Science Unit, DG Research and Innovation, European Commission.

38. All discussants made a short introductory presentation, reflecting on the earlier presentations and the challenges, good practices and lessons to be learned from the pandemic. This was followed by a moderated discussion. A number of key issues emerged from this discussion and the workshop as a whole as follows:

1. Research data infrastructures and networks have been critical in the science response to COVID-19. Funding is required to maintain these RIs and networks in a state of readiness. Prior investment is there to be leveraged during crises
2. There are technical obstacles in relation to achieving FAIR data but the major challenge is data governance and this stretches beyond research. All stakeholders – public and private - need to be involved in designing a pandemic data response and management system that is publically monitored, controlled and accountable. This needs to be prepared before the next crisis happens.
3. Citizens and data subjects need to be included in co-designing data governance arrangements for individual projects and for data systems as a whole. Cooperation is required at different scales from local to global. A global system of systems, based on national and regional networks is more likely to be effective than a single rigid global architecture.
4. Real time data access is difficult but is essential for managing a pandemic. From the outset, there is a need to integrate health services and public health surveillance data, with research needs and vice-versa. Scientific research should be integrated into the regular operations of public health agencies and institutes. National statistical agencies can also play an important role.
5. There is a need to support incremental steps towards a global FAIR data system and keep moving forward with ‘a coalition of the willing’. Top down visions, frameworks and direction need to integrated with bottom-up community-led efforts. Cooperative actions are required at different scales from local to global.
6. There are significant gaps and biases in the global data on COVID-19, both in terms of geographic coverage and lack of representation of certain populations groups. This is compounded by a general lack of disaggregated data, which makes it difficult to fully understand the evolution of the pandemic and the effect of policy interventions. Inclusivity and inclusiveness should be an explicit aim for pandemic data response systems, which has important implications for their design and governance.
7. Not all data can, or should, be open but it should be FAIR and all meta-data should be open and FAIR. Open data (and software and workflows) enable many - public and private sector - actors to contribute in the response to crises.
8. There is a balance to be achieved between standardisation and conformity versus diversity and innovation but making data FAIR should be part of routine research practice and the adoption of community recognised standards is critical.
9. The incentive system for researchers needs to change and how we evaluate the contribution of researchers in response to COVID-19 will be a good test of this. Data Management Plans (DMPs) need to be routinely implemented and open science efforts, e.g. provision of good quality data resources or trusted data respositories, should be evaluated and rewarded. This needs to be coupled with appropriate training and education for researchers and research support professionals.
10. International fora, such as the Research Data Alliance, which bring together different actors from different disciplines and different countries play an essential role in building the technical and social bridges that are necessary to ensure the optimal use of data for research and policy-making during crises. International community-led initiatives should be supported accordingly.

39. It was noted that the recently revised OECD Recommendation on Access to Research Data from Public Funding provides a global framework (with commitment from countries) under which many of these issues can be taken forward, via the development of an implementation toolkit which will support the monitoring and the sharing of best practices in this field. The Recommendation of the Council on Health Data Governance (OECD 2016) is also relevant and a new Recommendation on Enhancing Access to and Sharing of Data (EASD) is scheduled for adoption by the OECD Council in the coming months. This covers all data produced by the public and private sectors, and provides recommendations on how to achieve co-operation and partnership between public and private sectors to enhance access to and sharing of data. The COVID-19 pandemic should act as stimulus to fully implement these various recommendations and ensure that research and data systems are able to respond more effectively to societal needs, including future pandemics and other crises.