Organisation for Economic Co-operation and Development

DSTI/STP/GSF(2017)5/FINAL

Unclassified **English - Or. English**

21 December 2017

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

OECD Global Science Forum

CO-ORDINATION AND SUPPORT OF INTERNATIONAL RESEARCH DATA NETWORKS

Final draft report

This paper was approved and declassified by the Committee for Scientific and Technological Policy (CSTP) on 24 October 2017 and prepared for publication by the OECD Secretariat.

This paper is also available as OECD Science Technology and Industry Policy Paper No. 51.

Contact: Carthage Smith (carthage.smith@oecd.org).

JT03425011

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

Foreword

Open science is being advocated by policy makers in many countries and by international organisations as a way of increasing the efficiency and effectiveness of public investment in science. A critical foundation for open science is access to research data (data that is used and generated by public research) that needs to be provided in such a way as to be readily re-useable for further research and analysis. This access needs to be sustainable over the long-term and uninhibited, insofar as is possible, by national, disciplinary, or cultural barriers. It is dependent on individual data repositories at the institutional, national and disciplinary levels and on co-ordinated international networks of these repositories.

In October, 2015, the OECD Global Science Forum (GSF) commissioned two separate projects to inform policies to promote open data for science. The first of these addressed business models for sustainable data repositories and is the subject of a separate dedicated policy report. The second project, which is the subject of this report, focused on internationally co-ordinated data networks. The overall aim of this project was to identify principles and policy actions that can enable the establishment and maintenance of effective international data networks that are necessary to support a global open science enterprise. It was agreed at the outset that a case study approach should be adopted to build on the experiences of existing data networks across different scientific domains and regions. The potential benefits of conducting this project in collaboration with relevant international partner organisations were also explicitly recognised.

This project has been carried out in partnership with the World Data System of the International Council for Science (ICSU-WDS or WDS). It has also benefitted from strong links with the international Research Data Alliance (RDA). The two-day workshop in Brussels in March, 2017 that was a critical part of the project was generously hosted by the European Commission Directorate for Research and Innovation.

In addition to these organisations, many individuals across the World have shared their knowledge and expertise to realise this project. An international Expert Group, co-chaired by Sanna Sorvari and Andrew Treloar (see Appendix 1 for the full membership) has overseen the project and the writing of the report. Other experts from a number of data networks (see Appendices 2 and 3) generously shared their experiences via telephone interviews and many of these and a number of additional stakeholders participated in the workshop in Brussels. The report itself was drafted by Mark Parsons with substantive input from several Expert Group members. The work was supported by members of the OECD Secretariat - Giulia Marsan-Ajmone and Taro Matsubara, who conducted the project interviews - and Carthage Smith.

This publication is a contribution to the OECD Going Digital project, which aims to provide policymakers with the tools they need to help their economies and societies prosper in an increasingly digital and data-driven world.

For more information, visit www.oecd.org/going-digital

#GoingDigital

Table of contents

Foreword **2**

Abstract **4**

Executive summary **5**

Background and context 5
Conclusions and recommendations.................... 6

1. Introduction **9**

1.1. Background and aims.................... 9
1.2. Definitions and scope.................... 9
1.3. Methods 11

2. Observations and challenges **15**

2.1. Diversity and complexity.................... 15
2.2. Common motivations and challenges 16

3. Promoting international co-ordination.................... **19**

3.1. Users, services, and interoperability 19
3.2. Facilitating openness across countries and cultures.................... 22

4. Funding and governance.................... **25**

4.1. A central challenge 25
4.2. Respecting difference.................... 26
4.3. Sustaining co-ordination 26
4.4. Industry involvement 27

5. Guiding questions and recommendations **28**

5.1. Guiding questions 28
5.2. Policy and planning 30
5.3. Building and maintaining useful networks 32
5.4. Funding and long term sustainability.................... 35
5.5. Concluding remarks 37

Endnotes.................... **38**

References **38**

Appendix 1. OECD-GSF/ICSU-WDS Expert Group.................... **40**

Appendix 2. Persons interviewed for case studies **41**

Appendix 3. Workshop on International Co-ordination of Data Infrastructures for Open Science, Brussels, 30-31 March 2017.................... **42**

Appendix 4. Questions for the initial survey.................... **43**

Appendix 5. Case study interview questions.................... **44**

Abstract

International research data networks are critical for progress in many scientific domains and underpin efforts to promote open science. At the same time, many of these networks are fragile and the responsibilities for their support and performance are frequently distributed across a variety of different actors. This report explores the challenges and enablers for the effective functioning of international research data networks. It analyses the diversity and complexity of these networks, and issues such as governance and funding, in a selection of 32 cases. It includes a set of policy recommendations as a basis for building the shared understanding that is necessary to develop effective and sustainable international research data networks.

Keywords: International, Research, Data, Networks, Co-ordination, Interoperability, Open Science, Open data

Executive summary

Background and context

Research data repositories, working together in federated international networks, enable the sharing of data within and between scientific disciplines and countries and thus provide the foundation for open science. Federation in this context means supporting exchange of data between the individual repositories, usually for the purposes of discovery, while allowing each repository to retain its own independence. Developing effective and sustainable international research data networks is critical for progress in many areas of research and for science to realise its potential in addressing complex global societal challenges. In this regard, research data networks can be an important driver of innovation in the digital economy. However, the development and maintenance of effective networks is not always easy, particularly in a context where public resources for science are limited and international co-operation is not a priority for many countries or funding agencies.

The global landscape for data sharing in science is complex. Many international data networks already exist and they are highly variable in their aims and structures. Some are linked to large intergovernmental research infrastructures, have highly developed centralised services and deal mainly with the data needs of single disciplines. Some are highly distributed, have much less rigid governance structures, and provide access to data from many different domains. Most are somewhere between these two extremes and they cover different geographic regions from regional to global. All provide a mix of data and associated data services to meet the needs of the research community, and this provision depends on a mix of hardware, software, standards, protocols, and human skills. These come together, working across national boundaries, in technical and social networks. In all of this, what makes a network function effectively (or the converse) is unclear. Without this understanding, it is hard to see what can usefully be done at the policy level to promote the development of effective and sustainable data networks. Hence the rationale for the present project - to study a variety of currently successful networks, explore the challenges that they are facing and the lessons that can be learned from confronting these challenges, and, where applicable, to translate this analysis into potential policy actions.

This project was conducted with the active oversight of an international group of experts from different countries and fields of data science. With their assistance, detailed descriptive, operational and reflective information was collected on a total of 32 international data networks that to a large extent reflected the variety of the existing landscape. This information gathering was done in two phases - a general survey and structured in-depth interviews, and then the overall findings were considered in a 2-day international workshop that brought together about 50 experts, including funders and policy makers. The overall conclusions and recommendations from the project are summarised below. There are also a number of “lessons learned” embedded in Chapters 2-4 of this report that support the overall conclusions and have policy implications in specific contexts. As suggested at the outset, all of the networks are in their own way unique. Therefore, the recommendations are not designed to be prescriptive. They suggest what needs to be done but are not specific in how it should be done in particular contexts. They provide a basis for building the shared understanding that is necessary, at the policy level and across countries, to develop effective and sustainable international data networks.

The overall goal of this study was to establish principles and policy actions that can accelerate the establishment of effective, open and sustainable international data networks, which are an essential part of the global data infrastructure for science.

Conclusions and recommendations

Despite the great variety in the networks that were included in this study, a number of common challenges and potential solutions were identified. Many of these are intimately related to policy mandates and incentives (including funding) that are beyond the control of network participants. These mandates and incentives will be affected by a greater understanding by policy-makers of the contribution that international data networks currently make to the scientific enterprise and their critical role as a foundation for open science. This includes appreciation of what makes for a successful network and conversely what is likely to lead to failure. Unfortunately, there is no simple one-size-fits-all answer, but effective actions to enable the successful functioning of international data networks can be taken at several scales and by a variety of actors.

The importance of policy

The main barrier to open sharing of curated research data across geographic borders (and scientific domains) is the lack of policy coherence and trust between different communities. This is manifest in different interpretations of openness, different legal regimes for data sharing, and different ethical perspectives. Such differences need to be respected and understood but should not prevent a common understanding and workable international agreements being reached around the sharing of public research data. In this regard, there are many successful examples of sharing data, including what might be considered "sensitive data", internationally, and the lessons from these cases are there to be built on.

Recommendation 1

Responsible national authorities should be identified and work toward common definitions of, and agreements on, open data.

Recommendation 2

Governments need to work toward commonly agreed and enforced legal and ethical frameworks for the sharing of different types of public research data.

Data networks as critical infrastructure

Several international data sharing partnerships have been in existence for many decades and have become an invisible part of the infrastructure of science. As such, the danger is that they are ignored but there is a massive increase in supply and demand for research data in almost all fields of science, and data networks need to be correspondingly up-dated or newly developed and maintained to respond to this. These networks are the business-critical point of weakness in many research areas and for open science as whole.

Recommendation 3

All stakeholders need to recognise international research data networks as a critical part of the generic infrastructure for open science.

Recommendation 4

Responsible national and international authorities must include data networks in long-term strategic planning and support processes for research infrastructure.

Building successful networks

Developing and maintaining a successful international data network is dependent on a number of factors, both technical and social. Individual networks need to be tailored to the needs of specific data providers and users and they also need to evolve over time. They require an appropriate mix of long-term commitment, consistency and flexibility.

Recommendation 5

In establishing, developing, operating, and supporting international data networks the following "organisational" aspects should be taken into account:

- Networks should have a clearly defined user and data provider community. At the same time potential requirements of new users, e.g. in terms of interoperability, should not be ignored.
- Networks should have a clear understanding of how they relate to other networks and how they fit into the global research and data sharing landscape.
- Different governance models can succeed, but there is need to define clearly what is to be governed and to involve the data users and providers.
- Roles and responsibilities across a network, e.g. for data curation versus service provision, must be clearly defined and, where possible, data champions identified.
- The necessary level of standardisation needs to be clearly defined taking into account user requirements and the need to maintain an appropriate balance of autonomy amongst network nodes and central control.
- Objective and transparent mechanisms for network assessment should be based on standard quality management frameworks and certification of participating data repositories.
- Consideration needs to be given to regional differences in capacity - infrastructure and human resources - and in culture, community needs and expectations.

Funding and long term sustainability

The current funding arrangements for international data networks are inadequate. Despite the increasing demand for data curation, data repositories are struggling for support in many countries and open access to valuable data has in some cases been lost (see the companion report on Business Models for Sustainable Research Data Repositories (OECD, forthcoming). There are additional funding and sustainability challenges for international data networks, many of which have no obvious sponsor for their critical central co-ordination functions. Often there is a considerable amount of uncosted "in kind" support involved in networking and the amount of extra funding required for co-ordination is small but this should not mean that it is neglected since it can be critical to the network's success. There can be considerable overall cost benefits from federation of activities. At the same time, many networks do not currently have clearly articulated value propositions that can be used to justify additional investments.

Recommendation 6

Funders and host institutions should view internationally co-ordinated data networks as a long-term strategic investment and support them and engage with them accordingly.

Recommendation 7

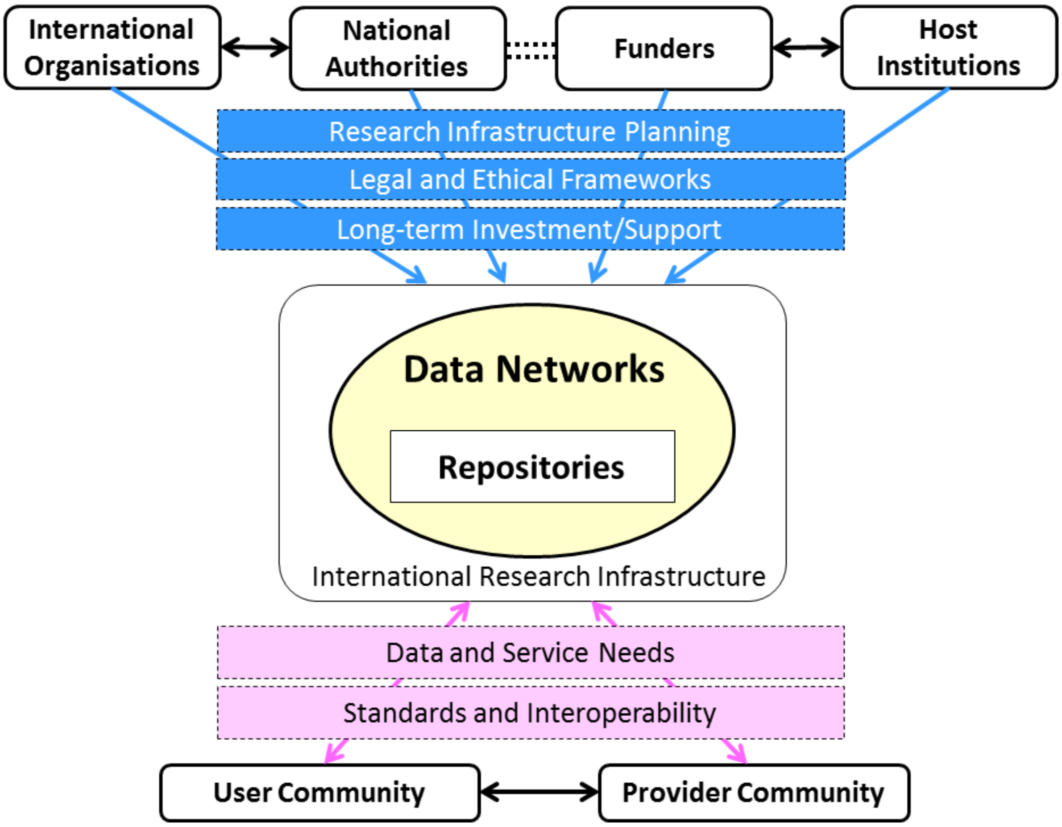
Networks should have clear business models, including value propositions and measures of success that are relevant to their different stakeholders and these measures should be monitored.

Recommendation 8

Funders should actively participate in relevant international discussions and forums to improve long-term functioning, support and co-ordination of data networks.

The main actors and responsibilities associated with these recommendations are illustrated below in Figure ES.1 and the recommendations are expanded upon in Chapter 5 of this report.

Figure ES.1. Key actors and responsibilities for international research data networks


International Organisations
National Authorities
Funders
Host Institutions
Research Infrastructure Planning
Legal and Ethical Frameworks
Long-term Investment/Support
Data Networks
Repositories
International Research Infrastructure
Data and Service Needs
Standards and Interoperability
User Community
Provider Community

Source: Authors' analysis.

1. Introduction

1.1. Background and aims

Open data sharing networks that operate across national boundaries have a long history. For example, meteorological data have been shared among nations dating back to agreements established in 1873 (Le Treut et al., 2007). Such sharing is enabled by a number of factors such as an understanding of how data might be useful once acquired, and appropriate data acquisition mechanisms, but sharing and reuse can only achieve its promise when the necessary standards and national policies are implemented to allow the data to be meaningfully aggregated, interpreted and sustained (Edwards, 2006).

As noted in the report from the G8+05 Global Research Infrastructure Sub Group on Data (2011), the emergence of "data driven science" reflects the increasing value of a range of observational, sensor, streaming, and experimental data in every field of science. Information and communication technology infrastructures for scientific data are emerging world-wide, but, often these data cannot be shared nor are they interoperable across countries and disciplines; moreover, they are unsustainable due to lack of commonly agreed governance, legal frameworks, and funding models.

Over the past sixty years, multiple internationally co-ordinated data networks have developed in different research domains. These networks each have their own histories, specificities, and rationales. They have different missions. In a minority of cases the mission is embedded in an intergovernmental agreement but otherwise a variety of more or less formal governance arrangements have also been adopted. Despite their heterogeneity, these networks share a number of common drivers and goals as well as common challenges.

This study examined 32 internationally co-ordinated data networks in order to better understand these common benefits and challenges and to learn what makes successful networks. The overall goal of this study was to establish principles and policy actions that can accelerate the establishment of effective, open and sustainable international data networks, which are an essential part of the global data infrastructure for science.

1.2. Definitions and scope

Early in the study it became apparent that there are many varying conceptions and usages of the term "infrastructure" when it relates to research, development, and information sharing. A number of formal definitions exist, all quite broad in their scope, but these vary by country, research domain, and profession (see Box 1).

For example, the geospatial community began discussing a "spatial data infrastructure" in the early 1990s (NRC, 1993). About the same time, the concept of an "Information infrastructure" was emerging in reference to the Internet and WWW. In the 2000s, there was a growing focus on the needs of data more broadly than geospatial information but more specifically than the general Web. The terms Cyberinfrastructure in the US (NSF Blue Ribbon Advisory Panel, 2003), e-Infrastructure in Europe, and e-Research Infrastructure in Australia, came into use to help describe new requirements around integrating, data, networks, high-performance computing, and other components to take advantage of growing data and capabilities. In Europe, the term "research infrastructure" (RI) has a specific definition and RIs have received a particular focus with the development of the European Strategy Forum on Research Infrastructures (http://www.esfri.eu).

An earlier OECD report (OECD, 2014) focused on International Distributed Research Infrastructures (IDRI), which are increasingly recognised as a particular type of RI (See Box 1) and can be interpreted as including some federated research data structures. The current study focuses specifically on the data stewardship, interoperability, and reuse aspects of the scientific enterprise, which were not covered in the IDRI report. While it included some of the same organisations that were considered earlier, they were categorised for the current work as **internationally co-ordinated data networks** in order to avoid confusion.

Box 1. Some definitions of infrastructure

Cyberinfrastructure

The comprehensive infrastructure needed to capitalise on dramatic advances in information technology has been termed cyberinfrastructure (CI). Cyberinfrastructure integrates hardware for computing, data and networks, digitally-enabled sensors, observatories and experimental facilities, and an interoperable suite of software and middleware services and tools. (NSF Cyberinfrastructure Council, 2007).

e-Infrastructure

e-Infrastructure refers to a combination and interworking of digitally-based technology (hardware and software), resources (data, services, digital libraries), communications (protocols, access rights and networks), and the people and organisational structures needed to support modern, internationally leading collaborative research be it in the arts and humanities or the sciences.(Research Councils UK, 2014).

E-Research infrastructure

Comprises the ICT assets, facilities and services that support research within institutions and across national innovation systems, and that enable researchers to undertake excellent research and deliver innovation outcomes. (CASRAI, 2015).

Internationally distributed research infrastructure

An International Distributed Research Infrastructure (IDRIS) is a multi-national association of geographically separated distinct entities that jointly perform, facilitate or sponsor basic or applied scientific research. (OECD, 2014).

Research data infrastructure

Research Data Infrastructures can be defined as managed networked environments for digital research data consisting of services and tools that support: (i) the whole research cycle, (ii) the movement of research data across scientific disciplines, (iii) the creation of open linked data spaces by connecting data sets from diverse disciplines, (iv) the management of scientific workflows, (v) the interoperation between research data and literature and (vi) an integrated Science Policy Framework. (GRDI2020 Consortium, 2010).

Research infrastructure

Facilities, resources and services that are used by the research communities to conduct research and foster innovation in their fields. They include: major scientific equipment (or sets of instruments), knowledge-based resources such as collections, archives and scientific data, e-infrastructures, such as data and computing systems and communication networks and any other tools that are essential to achieve excellence in research and innovation. (ESFRI, 2016).

Furthermore, Star and Ruhleder (1996) argue that infrastructure is best understood as a body of relationships, so we chose to focus on those relationships or networks that enable broad and distributed data reuse. This report was also developed in parallel with a companion report, *Business Models for Sustainable Research Data Repositories* (OECD, forthcoming). That report focused specifically on how to resource and sustain individual research data repositories. This report focuses on the collaborative network aspect necessary to connect those repositories and other services in a global infrastructure.

While there were varying views expressed on terminology in the questionnaires, interviews, and workshops conducted for this study, there was strong consensus on the shared value of internationally co-ordinated data networks in providing services for sharing and reusing data. It is recognised that data be must well connected to high-bandwidth networks and closely connected to high-performance computation for some analysis tasks, but this report focuses on the reuse and stewardship of data in an international and often interdisciplinary context. Overall, the networks studied here seek to contribute to and fit into a broader, global data sharing infrastructure.

1.3. Methods

An international Expert Group (EG, Appendix 1), with members from different countries and disciplines reflecting the diversity of the data network landscape, oversaw this project. The expertise and active involvement of this EG was critical to the conduct of the project, which used a case study approach. Cases for inclusion were identified by EG members, who also then played a key role in collection and analysis of information from these cases, with the support of the GSF secretariat (GSF members were also invited to propose cases for inclusion).

The identification of case-study candidates was based on the following criteria:

- Data networks and data for open science were the main focus（journal publishing was not a major focus but could be included when associated with data).
- Networks had to have an international remit but could be concentrated on one country or region (integration of national solutions into global networks was an important theme).
- The emphasis was on identifying “good examples” that others could learn from.
- Include different levels and scales of network.
- Ensure a mix of disciplinary and domain specific or inter-disciplinary networks.

This led to the identification of 32 data networks (see Table 1). It should be noted that although the included cases are to some extent representative of the variety of existing networks, there was no attempt made to try and select a statistically significant sample of networks. The analysis and findings have been treated accordingly

Table 1. List of data networks surveyed and interviewed

	Acronym	Name	Geographic coverage	Discipline
1		AddNeuroMed	Regional (Europe)	Dementia research
2	ACTRIS	Aerosols, Clouds, and Trace gases Research Infra-Structure	Regional (Europe)	Environmental data etc.
3	**ADNI**	**Alzheimer's Disease Neuroimaging Initiative***	**US (International links)**	**Dementia research**
4	Argo	Argo	Global	Oceanography and meteorology
5	CBRAIN	Canadian Brain Research Imaging Platform	Canada (International links)	Neuroimaging, cognitive neuroscience etc.
6	CLARIN	Common Language Resources and Technology Infrastructure	Regional (Europe)	Humanities and social sciences
7	CESSDA	Consortium of European Social Science Data Archives	Regional (Europe)	Social Sciences
8	**DARIAH**	**Digital Research Infrastructure for the Arts and Humanities***	**Regional (Europe)**	**Arts and Humanities**
9	**ELIXIR**	**ELIXIR***	**Regional (Europe)**	**Life Sciences**
10	**EMBL - EBI**	**European Molecular Biology Laboratory - European Bioinformatics Institute***	**Regional (Europe)**	**Life sciences,**
11	**GBIF**	**Global Biodiversity Information Facility***	**Global**	**Life sciences, Biodiversity**
12	GIRO	Global Ionospheric Radio Observatory	Global	Ionospheric physics etc.
13	GODAN	Global Open Data for Agriculture and Nutrition	Global	Agriculture and nutrition
14	**GEO**	**Group on Earth Observations***	**Global**	**Earth Observations (Multidisciplinary)**
15		Helix Nebula	Regional (Europe)	Various disciplines, mainly focusing on large sciences
16	**ICSU-WDS**	**ICSU World Data System ***	**Global**	**multidisciplinary**
17	IPCC - DDC	Intergovernmental Panel on Climate Change - Data Distribution Center	International	Climate change
18	INCF	International Neuroinformatics Co-ordinating Facility	International	Neuroscience
19	IODP	International Ocean Discovery Program	Global	Seafloor drilling, ocean samples/observations
20	**IVOA**	**International Virtual Observatory Alliance ***	**Global**	**Astronomy**
21	ICPSR	Interuniversity Consortium for Political and Social Research	Global	Social sciences,
22	**IUGONET**	**Interuniversity Upper atmosphere Global Observation NETwork ***	**Japan (international links)**	**Space physics**
23	LIGO	Laser Interferometer Gravitational-Wave Observatory	International	Gravitational physics
24	NITRC	Neuroimaging Informatics Tools and Resources Clearinghouse	USA (international links)	Neuroimaging etc.
25		OpenAIRE	Regional (European)	All disciplines
26	**H3ABioNet**	**Pan African Bioinformatics Network***	**Regional (Africa)**	**Bioinformatics and genomics**
27	PaNdata	Photon and Neutron data infrastructure initiative	Regional (Europe)	Neutron and photon laboratories (Multidiscipline)
28		SeaDataNet	Regional (Europe, Mediterranean and Baltic regions)	Oceanography
29	SBP studies	Swedish Brain Power studies	Sweden (international links)	Brain research
30		UK Biobank	UK (international links)	Universal health data
31	**WDCM-GCM**	**World Data Centre for Microorganisms - Global Catalogue of Microorganisms***	**Global**	**Biodiversity**
32	WLCG	Worldwide LHC Computing Grid	Regional (Europe)	Astrophysics, Nuclear & Particle Physics

*Networks highlighted in **bold** were subject to in depth interviews (in most cases this was in addition to participation in the initial survey).

Initial information was compiled on 29 of these networks using a structured on-line template (Appendix 4). The template requested basic descriptive information as well as reflective input on challenges and obstacles, past and future milestones, and overall network development. The templates were completed by EG members, using their own knowledge and expertise and supplementing this, as necessary for roughly half of the cases, with interviews with network managers.

Following an initial analysis of the completed templates, in depth interviews were conducted for 11 of the networks (highlighted in bold in Table 1 and also see Appendix 2). Using a standardised questionnaire (Appendix 5), the GSF Secretariat interviewed individuals with a good working knowledge of the networks to collect more reflective perspectives and to pursue particular issues relevant to policy. In several cases, multiple perspectives were collected on a particular network. Each interview took 1-2 hours, and the interviewees reviewed the completed questionnaire for accuracy. The interview questions ranged across the following areas, which were considered as being potentially policy-relevant:

- Implications, advantages and operational challenges of being an international network
- Steering of the network: the role of funders and the research community
- Limits and challenges related to data openness
- Governance-related issues
- Sustainability and private-sector involvement

OECD staff and the EG analysed the collected information. Pairs of EG members then developed thematic briefing papers on key policy relevant issues, based on the interview themes above. These briefing papers provided the basis for a dedicated international workshop (Brussels, March, 2017) that brought together about 50 data network operators, funders, and policy makers (Appendix 3). The workshop was designed to enrich the preliminary analysis and test emerging recommendations with these different communities.

During the course of the study a number of overarching policy questions were refined and these provided a framework for developing policy recommendations as discussed in Chapter 5:

- When is an internationally co-ordinated data network needed? How does it fit into the existing landscape? How does one assess the performance of the network?
- What interoperability arrangements are necessary for the effective operation of an international network and how are they constructed?
- How can governments maximise data openness and reuse within an agreed international framework?
- What is the best governance model for a particular network and how can it evolve over time? How are decisions made? What are the roles of the different institutions and people involved in a network and how are they best co-ordinated?
- What business model is appropriate to sustain a network over time?

The observations, findings, analysis, and recommendations that are presented in the following chapters of this report take into account all the collected information from the initial survey, in depth interviews, and the international workshop. This is supplemented by the rich experience and knowledge of EG members and with references to the published literature where relevant.

2. Observations and challenges

2.1. Diversity and complexity

A first-level observation of this study was simply how diverse the different networks are. They have widely varying stakeholder communities, research missions, funding and governance models, and network topologies. This diversity must be a primary consideration in any strategic policy analysis or development.

The networks presented a variety of different governance models and covered different geographical regions. Some, (IVOA, WDS) are global in extent, while others (ELIXIR, EMBL, H3ABioNet) are more regional. Some of the data resources are centralised (EMBL) others adopt a hub and spoke network model (GBIF) with some centralised funding and co-ordination and a series of nodes. Others operate as a decentralised network where neither co-ordination of infrastructure nor funding is centralised, but co-ordination of standards is (IVOA.) In some cases, networks may follow a multi-hub and nodes model, where multiple national bodies contribute data to a national hub that is part of an international network, as is the case in Spain and Mexico with regards to the GBIF network. In the USA, on the other hand, GBIF data management is de-centralised and individual institutions can deposit directly into GBIF without going through a central interface.

Many networks are focused on a particular discipline or domain (GBIF, IVOA, EMBL) others are explicitly interdisciplinary (GEO, WDS). Some of the networks could be sorted on the types of data they provide in accordance with the US National Science Board (NSB, 2005) classification of research, resource, and reference collections, but several networks handle many types of data ranging from operational satellite data to lab-generated research collections. Some networks are tightly coupled to specific instruments or research infrastructures (CERN), whereas others deal with a variety of research infrastructures (IVOA) or are very multi-platform and interdisciplinary (WDS). This latter distinction may be the most significant with regard to implications for policies for network development.

The governance and funding model adopted by an international data network often depends on (and determines) the funding sources and operating mechanism of the network. There can be a significant difference in the governance model depending on whether the network is “owned” by science communities and their research performing organisations or if its “owners” are governments (countries). The latter often operate under the framework of established international organisations such as the World Meteorological Organisation (WMO) or the Group on Earth Observations (GEO) ministerial summit. A lot depends on what part of a data network’s activities need co-ordination and policy development at the central level. In large consortia the approach tends to be that the broader issues requiring overarching policies for the network are mainly governed centrally. Activities such as data curation and publishing cycles tend to be managed more locally.

There are myriad reasons for the different types of networks, and they are usually in response to a particular mission or vision, but different types may face different challenges. For example, more interdisciplinary or scientifically generic networks have greater challenges identifying and responding to defined user communities (WDS) than networks with a very specifically defined and focused user community (IVOA) who may then customise their products for other communities. On the other hand, networks with broader user communities may be more innovative in the development of interdisciplinary services. As another example, networks that are tightly associated with a particular large instrument (CERN, LIGO) tend to have less funding complexity and uncertainty, but are typically focussed on a single discipline and may be less agile or adaptive.

Cross-disciplinary (and cross-network) developments require pragmatic and agile governance. Cross-disciplinary work is not an abstract notion and developments should be based on identifying and supporting real research and societal questions. Sometimes, disciplinary focused networks may not see the full potential of their data or the challenges of making it interoperable with data from other disciplines. Often these challenges are because of the different sociologies of the communities more than the structure and description of the data itself. Data "champions" within a network can often play a useful role in making data more useful across disciplines and cultures. These champions may be respected researchers in a specific discipline, but they are often individuals with experience in different disciplines or research domains and IT or library science.

This complexity of network types makes it hard to understand how they all interrelate. There was significant discussion throughout the present project on the need to have a better understanding of the overall scientific data landscape. It is beyond the scope of this report to provide a full map of the global data landscape, and it would only be of marginal use, because the landscape is continually changing. However, it was clear that the most effective networks are well aware of and define their place well within the relevant part of global data landscape, and they continually respond as this landscape evolves

There was also a lot of discussion within the Expert Group and at the workshop on how and whether these diverse networks could be classified in some way. Of course, any classification scheme depends on perspective, and no classification scheme perfectly captures all the networks, but some distinctions or contrasts can be useful when making policy choices. This report will to highlight how these contrasts are relevant to policy and operational decisions as appropriate.

2.2. Common motivations and challenges

Despite the diversity, there are common motivations for people and organisations to establish and maintain the various networks. Similarly there were also a number of common challenges that most of the networks faced.

2.2.1. Common motivators

- Mutualisation of resources and increased resilience: partners of an internationally networked data infrastructure can access more data and resources as part of the network than in isolation. In some cases, they can obtain more funding nationally or internationally by being part of the network. Centralised functions of a network ensure cohesion and promote collaborations and improved efficiency. Distributed functions build robustness and provide back-up mechanisms, for example, to ensure long-term preservation of datasets.

- Increased visibility, reputation, and influence: participating in data networks allows all partners—especially the "small" players—to participate in defining international standards and practices. It can provide better international and national exposure and even a form of accreditation, which improves reputation and has positive feedbacks in terms of overall organisational efficiency and sustainability.

- Improved trustworthiness and quality assurance: data networks in most cases establish “rules of engagement” with various levels of requirements for participating data repositories including certification and data quality standards.
- Global reach, improved agility, and quality: international networks allow scientific communities to build and maintain global datasets and can help bridge the digital divide between rich and poor countries. The “friendly competition” between partners and the richness of their shared experience allows better adaptability, awareness, problem solving capacity, and generally increases the quality of the data and the network.
- Inclusiveness and consolidation: distributed functions within international data networks allow multiple partners in different countries to play a role depending on their capacity. At the same time, duplication and redundancy can be reduced and efficiencies can be gained.

2.2.2. Common challenges

- Variable data sharing policies, unclear or missing data usage licences, and legal regimes across countries and regions create obstacles and difficulties for establishing internationally co-ordinated data networks. The general trend in many countries regardless of their economic development is for openness as a default policy for public research data but the situation on the ground is still highly variable, including in regions such as Europe that from a policy standpoint might seem relatively homogenous.
- Privacy concerns and data restrictions around human subject and other sensitive data. Data restrictions are made even more challenging when trying to share data internationally. There is inconsistent agreement on how sensitive data should be handled, and marginalised communities may be concerned with how their data will be shared and used.
- Cultural and linguistic differences. It takes time and a lot of energy to build sustainable trust relationships between international partners. Differences in regional cultures and in scientific cultures and maturity levels of scientific communities often result in different expectations about services offered and needed.
- Global connectivity is not homogenous across the world. In the context of increased volumes of research data, storage is becoming more affordable thanks to cheaper technologies but connectivity is still very expensive and poses a serious challenge for internationally networked data infrastructure, where the least-connected countries and communities are often disadvantaged.
- Divergence between the needs of the target scientific community and the objectives or governance of the data network providers: a top-down approach developed without an engaged target community is generally less successful than a bottom- up international network driven by the target community.

- Variable funding regimes and reward systems either hinder the building and maintenance of international data networks or create a misalignment between network partners. Differences in levels of public investment for data infrastructures between countries make it difficult to co-ordinate international data networks. Sustainability of the networks is extremely challenging to achieve if key elements are dependent on research-based funding. On the one hand, research funders in rich countries are generally keen to fund the establishment of data networks as projects but less keen on funding their co-ordination (human resources) and maintenance in the long-term. On the other hand, research data networks tend to be a low priority in the less developed and less-connected countries.

3. Promoting international co-ordination

3.1. Users, services, and interoperability

Users experience a network through the services it offers. No matter how sophisticated the architecture and complex its components, it is the service that impacts the user community. Most networks offer additional services on top of the basic data collection and provision. These include cataloguing of metadata, reformatting, quality controlling, clustering and aggregating data, offering tools to access, and explore the data inventory and training users in the use of these specific tools. The most successful networks are those that clearly identify and strongly engage their user community in defining data and service needs.

3.1.1. User engagement

Several approaches are being used in order to engage scientists directly. At the level of network governance, giving scientists leading and active roles in steering and advisory committees has successfully engaged scientific users and is widely practiced. Another successful strategy is the establishment of working groups to address scientific topics or questions requiring data or services from the network. This ensures that scientists inform the development of the network according to their needs, which naturally attracts more users from the scientific community.

By involving the right combination of people in its governance structures, a network can become very influential, benefit from the selected expertise, achieve community trust, and become widely known. This depends on the personal commitment of the individual committee members, though. It can be difficult to engage scientists in the specifics of data sharing because it is tangential to their main interests and offers no direct reward. Again, data champions can help connect research communities into the larger data sharing network. This sort of liaison work can be critical, and should be encouraged. It is also important to find the right balance of stakeholders and to balance the expertise of the researchers with data, IT, and other experts in governance and decision-making processes.

More direct actions for engaging users include training courses and workshops, where users are directly introduced and trained in the services of a respective network. Such training and workshop activities can have high impact and are important for both the data networks and the user community. User needs are directly revealed and trained users are more likely to use a certain system again and become loyal customers. At the same time, the network operators receive direct feedback on their performance and services and can rapidly respond. However, training activities are costly and time consuming and can only be offered to a subset of potential users. Many data networks lack funding to involve their users in such a direct way. What may be more feasible in some cases is promoting the use of data from networks for education purposes, which can introduce data holdings and services to a future user community.

At a broader level, many data networks use surveys to obtain information about their performance and usability but it can be challenging to get the necessary feedback. Often a combination of training and surveys is used. Users can also be engaged via professional outreach activities such as scientific publications, dissemination of best practice documents or other guidelines that directly inform and guide the user community. Another option to involve scientists directly is to promote personal exchange and interactions. Some networks open collaborations, where individual scientists can directly submit data or they have a helpdesk where scientists get support and provide feedback.

Compulsory data submission into a community-endorsed repository, prior to scientific publication, is practiced in some fields and can help ensure data availability and also network engagement and usage. In addition, some data networks ensure that data contributors are credited if their submitted data is used for secondary analysis and publication. Without such mandates and incentives, data networks can be significantly weakened since data is used without recognition of its source (both the network that supplied it and the contributing researchers). Instituting a rigorous accreditation and data provenance tracking system is not a trivial task, though, and can hinder real-time access to data.

In developing networks it is important to gain and maintain the trust of the data providers – the source of the data - who may or may not be users. This helps ensure that data is not being manipulated, is well preserved, and is made available with enriched metadata and in line with the agreed intellectual property right conditions. Good practices and quality assurance encourage other data providers to contribute. So by encouraging and promoting the sharing and reuse of data, networks can get access to more data, extend their reach, and influence other networks.

Data curators and the data “champions” mentioned earlier can be central to developing the necessary trust between communities to develop data sharing agreements and to act as interpreters between different community languages. This emerged quite strongly in the humanities, where researchers have difficulties related to technology driven networks, but it was apparent in all networks. There is a need for specially trained people that can act as mediators within and across disciplines and especially between computer and information science and domain science. These data professionals are an emerging and important profession.

3.1.2. Network services

The topology (distributed or centralised) of data networks often directly affects their fitness of purpose. Within a typical, highly centralised and controlled network, metadata and data are made available by network nodes or individual providers to a centralised facility or hub where metadata and data are enriched, standardised, catalogued, structured, aggregated, quality controlled, and directly served from the centralised facility to the user. User services are also often primarily the responsibility of the centralised facility. Within a centralised network, aggregation and detailed quality control can be performed centrally, which assures consistency and documentation. It requires strong agreement amongst the network partners, though, and takes significant effort. It risks failure if the central control is too heavy handed. When done well, core services and flagship activities can be defined which can increase visibility and make networks more competitive.

In loosely co-ordinated, distributed networks, individual repositories normally take more responsibility for the stewardship and curation of their data. This can be useful because the local nodes often have a closer working relationship with their data providers and users, but the level of quality control and assurance often varies across nodes. Sometimes the same data may be hosted by more than one repository or service in the network. Duplicates and version conflicts can occur and are sometimes impossible to resolve. On the other hand, the flexibility of loose networks can often lead to more rapid and innovative adaption to changing needs because of the ready availability of a broad body of expertise and perspectives.

In general, broad or generic issues are better dealt with centrally and more disciplinary-specific issues managed in a more distributed way—closer to the actual user and discipline. Different nodes may also have different specialties and provide particular services for the whole network. Successful networks exchange good practices between their members and build overall network capacity. Some disciplinary-specific or instrument-based networks require a level of homogeneity that can only be provided by a strongly centralised network, but there are no hard and fast rules. For example, the IVOA is one of the most distributed networks we examined, but it is very focused on and well supported by the astronomy community. It did, however, take ten years for this network to be fully established and adopted. Network development takes time, and, as with everything, there are trade-offs.

The more generic and heterogeneous the data inventories of a network, the less useful they tend to be for the specialist user. It can be challenging to identify a key user community that will support a multi-disciplinary data infrastructure. On the other hand, the grand challenges of society all require interdisciplinary data integration and reuse. A central challenge for all the networks is to not only identify a relevant primary user community but also to be responsive to the ever changing and sometimes unanticipated needs of that community. This might include tailored services that are often offered for various user levels (e.g. experienced vs. beginning or heavy vs. low volume data users accessing a programmable interface (API) vs. a simple point and click interface). Even individualised services may be provided to meet diverse needs.

In some networks, it can be challenging to define core services or expertise. It should be recognised that, unfortunately, networks sometimes overstate their services and/are over-ambitious as to what they can provide. The maintenance of user services is often limited due to lack of personnel. This can, in extreme cases, alienate users and is an area where all stakeholders need to ensure realistic planning, development, monitoring and accountability.

3.1.3. Standards and interoperability

The challenges related to changing user needs can be at least partially overcome by ensuring some level of standardisation and quality assurance across the network as a whole and making judicious choices on what services are appropriate for different user communities. Many networks are involved in implementing international standards and defining specific requirements for different research fields in order to improve the quality and usability of data submitted by their various providers. In some cases, archives are certified by international bodies, such as the Data Seal of Approval or World Data System. This builds confidence and gives the network a quality stamp that can help attract more users.

The international value of data services is strongly influenced by their alignment with established standards, and their alignment with users' needs, recognising that many users integrate data form different sources. Networks deal with this in different ways. Sometimes mandatory standards are imposed at the network layer or through a central node. These standards may be quite rigorous and specify detailed metadata content, specific vocabularies, and precise transfer protocols. More typically, mandated standards are fairly lightweight and only cover basic citation and access information. Rigorous and detailed standards are most effectively developed and successfully adopted when there is a close association with an expensive research facility like a telescope or when there is a clear and explicit research demand that unites different facilities. For example, Earth observation data and model output used in the large climate model inter-comparisons for the Intergovernmental Panel on Climate Change (IPCC) are very well standardised in their format, documentation, and protocols. In this case, there is not a mandate, per se, but there is strong incentive for data sources to agree on rigorous standards because they want to contribute to the periodic IPCC climate assessments.

Another approach is to translate or broker across multiple defined standards to enable interoperability across diverse systems. GEO had particular success with this approach. They substantially increased the number of resources available through their systems when they replaced specific deposition requirements with a system that was able to broker across multiple standards. This approach is subject to mistranslation, but can be a powerful way to lower the overhead of belonging to the network for the participants and can more easily adapt to change. Similarly, several networks reported that taking a “network of networks” approach was useful in addressing cross-disciplinary or cross-network development, although this does add additional complexity.

Regardless of the approach to interoperability, it works best when there is a good understanding of the different roles of the different players in the network and there is a negotiated balance between distributed autonomy and central control. Unfortunately, these sorts of negotiations are rarely supported explicitly. There is continuous pressure from users and funders for data networks to increase their data inventories and keep developing new services despite limited funding. There is less attention on the mundane mechanics of standardisation and interoperability. Successful networks define their scope well and avoid the “not-invented-here” or “try-to-solve-it-all” mentality.

There is a similar tension in the adoption of commercial technologies, but as networks grow and become more interconnected, there will be more opportunities to optimise the “build-vs-buy” decision. We can expect to see more specialist service providers emerge and networks renegotiating their service arrangements and developing new value propositions. This changing dynamic is discussed in the report on Business Models for Sustainable Research Data Repositories (OECD, forthcoming). It is likely that different data repositories or providers will take different “build-vs-buy” decisions that may have implications for their participation in co-ordinated international networks and this needs to be openly discussed and monitored.

3.2. Facilitating openness across countries and cultures

A major goal of open data for science is to ensure that data is available to anyone who needs it, regardless of location or affiliation, with minimal barriers for access and reuse. International data networks inherit all of the “traditional” challenges of sharing data and data exchange within a research community, e.g. reluctance of researchers to share their data, lack of incentives, privacy issues and informed consent for personal data, and long-term sustainability. These challenges are multiplied in the context of international networks, in that attitudes as well as policies to address these challenges may differ across the partners of a network. As noted in Section 2.2, there are also the added challenges inherent in working across national boundaries: different cultures, languages, technological capabilities, skill levels, and legal frameworks. Networks that involve under-represented communities or have certain types of sensitive data must also grapple with mistrust from both local researchers and human research subjects engendered by past experience and unfair exploitation of natural and intellectual resources (Harmon, 2010). This has reinforced requirements from subjects, whose data is being used for research, that open sharing should benefit their communities. Fostering appropriate policies and governance mechanisms are imperative for achieving a congenial open, data-sharing ethos.

3.2.1. Attitudes and trust

The top challenge raised by international networks for open sharing of data across borders was the differences in attitudes and policies between countries. Within a country, different research communities may have different perspectives, and it is important to appreciate that any characterisations of particular nations or regions may only apply to certain types of data. Nevertheless, several networks (EMBL, WDS, H3BioNET) highlighted the lack of a culture of open sharing in South America, Africa, and Asia for certain types of data, particularly biological data. Even in mature networks there are challenges. IVOA had broad international backing, no privacy concerns, and the ability to develop standards across a multi-national federation, but there were different views, at the outset, on how best to query data. In this case, it was reported that cultural differences within international working groups had to be carefully managed to ensure that consensus was reached. In essence, the central issue is the need to establish and maintain trust across a network as well as with data providers and users in different countries.

In some countries and communities there is a mistrust of international networks because of past exploitation of resources and researchers. Researchers may come into a country from other nations, conduct their studies, and then withhold important data from local researchers until they publish. And the publications are in a foreign language and provide no benefit to the local community. So-called "parachute research" has attracted attention in the medical field where profit-driven motives have led to treatments being developed with data collected from a developing country, which then cannot afford to pay for the new treatments for its own citizens. Regional problems can quickly become international problems as was evidenced by recent Zika and Ebola outbreaks, where restricted access to data slowed response to the disease. Mistrust may explain why the culture of open sharing is less established in many countries: "They only share when they absolutely have to. They fear that they do not get credit and others steal the data." (H3ABioNet presentation at the project workshop in Brussels, March, 2017).

Equitable partnerships, based on mutual understanding and respect, are a prerequisite for promoting a global culture of open research data. More explicit guidance on expectations and norms for working across cultures need to be developed, and there must be some indication from funders that relevant data and associated assets will stay local at the end of a project. There is nascent but growing recognition of the concept of indigenous data sovereignty – "the right of a nation to govern the collection, ownership, and application of its own data"[1]. International networks are a way of promoting this concept while ensuring that data can be shared and used internationally in a way that is sensitive to, and fully acknowledges, its origins.

Some of these trust-related issues are because incentives and mandates for sharing differ widely across countries. Some countries, e.g., the UK, have funding agencies that are really pushing requirements for open data, while others do not. Several countries have national policies that require certain types of data not be moved out of the countries of origin. Some networks, e.g. ELIXIR, have addressed this problem, at least partially, by negotiating agreements for the trans-national sharing of metadata while the primary data stays within the country of origin.

3.2.2. Defining open

It is apparent that there is no uniform definition of open access to data across nations or even within nations. Different user communities understand publicly available data in different ways. For some types of sensitive data or where there are issues of equity these difference are not surprising, but there are still many interpretations of what open data is regardless of data type. In some instances, it is translated as controlled access or restricted access to full datasets (WDCM) rather than the fully open definition promoted by open science advocates. Openness may also be tied to funding streams and business models (for example, charging for value added data services), which are in turn influenced by policy mandates and incentives that are implemented at the national level (OECD, forthcoming).

Fully open sharing is also viewed in some countries as being in competition with developing commercial partnerships that might provide revenue for the researchers or their institutions.

Clearly defined policies on intellectual property rights, data ownership and licensing can be difficult to develop, but can do a lot to enhance the operation of a network and availability of data, as well as the level of trust and usability for users of the data. This was highlighted by GBIF, which went through a major effort to apply consistent Creative Commons licenses on all their data. It was challenging and resisted by some at first but ultimately helped build trust across the network.

3.2.3. Skills and capacity

Finally, multiple networks noted the challenges raised by differing technological capacities and skill levels across countries. The GBIF representative noted: “Developing countries want training and developed countries want infrastructure.” This is not just an issue of developed vs. developing nations, but also bigger vs. smaller nations (ELIXIR), stable vs. unstable nations (H3BioNET), or simply differences in national priorities for infrastructure spending. H3BioNET noted the difficulties in working with nations with disputed borders or where geographic or subject areas are subject to sanctions. And the issue is not constrained to the developing world. For example, EMBL and ADNI both noted that some developed countries lack sufficient infrastructure for curating and sharing biological data. All nations still find transferring very large data sets over networks to be challenging. Several networks were interested in the potential to leverage international development funding into infrastructure activities and in building sustainable and inclusive networks. For some networks, specific national nodes in the network already had a remit from sponsoring organisations to engage in capacity building.

4. Funding and governance

4.1. A central challenge

Data networks provide a way for research data to be shared and reused between different research teams. If researchers are to have access to the best data possible, then this sharing must be done at a global level. Many policy declarations over the past decade have highlighted the benefits of open data (OECD, 2015), and some data networks have a history of decades of service supporting data sharing, but there is no consensus as to the best way to fund these networks and their services. Financial sustainability, both for operation and enhancement, was identified as a central challenge for all the networks, and was a major topic of discussion during the project workshop.

The vast majority of public funding for research is available at a national level for nationally based research teams. Thus, there is an inherent disconnect between how data networks (and infrastructure generally) are funded and how they are best managed and governed for maximum social benefit. There is an intrinsic tension between the structural character of data networks and the time-limited, project-based funding models available for many research programmes. Furthermore, the benefit of the network is often for people other than those who create the data - potentially even in different countries - so there can be limited incentives for data collectors to contribute. Finally it is challenging to assess the cost-benefit of data preservation, because the long-term value of most data is not known at the time the data are created. All this means that developing the value proposition for the different stakeholders and funders who support an international data network is difficult.

It has long been argued that data arising from publicly funded research should be considered as a public good (OECD 2007; Arzberger et al., 2004), and, as such, access should be provided at no more than the marginal cost of distribution. However, this principle does not address the issues of how the cost of provision should be met or of how data creators should be recognised for their contribution. In some domains, data networks provide data that is also relevant to private sector research or provide the basis for commercial sector provision of value-added data services that may be located in a country other than those that provided most of the primary data. This raises questions about the role of private sector users and distribution of benefits arising from the use of internationally available public research data.

The OECD report on Business Models for Sustainable Research Data Repositories (OECD, forthcoming) explores these issues in more detail and discusses the overall economic fundamentals and challenges of data sharing, preservation, and reuse, from the perspective of individual data repositories. It suggests how certain regulations, incentives, and business models can be combined to help address the funding and sustainability challenges. The same challenges and models apply equally for internationally co-ordinated data networks, and networks and repositories are mutually dependent. Nonetheless, there are additional layers of complexity in networking across countries that make the design of appropriate funding and governance models even more challenging.

4.2. Respecting difference

The issues and risks around funding and governance relate to the type of network that is being supported. In a centralised data network, who pays for the central facilities? If it is based on multiple subscriptions, which is often the case, what happens when a member or country doesn't contribute? In a decentralised or hybrid model, which node performs which network function? What happens if a node drops out?

As discussed in Section 3.1, centralised networks can often impose more consistent standards, quality control, and data integration. They can more easily define and control core services and flagship products, which can increase visibility and network value. Centralised structures tend to occur when there is a focus on associated instrument or research infrastructure. Funders are much more likely to invest in data networks and related infrastructure when it is clearly tied to the much larger investment of a spacecraft, telescope, or super-collider.

Loosely co-ordinated networks, have the advantage is the lightweight administrative mechanism. It is relatively low cost and easy for countries or institutions to participate because such networks don't impose many additional requirements. These types of networks tend to be more agile and can have more flexible and responsive governance and management structures. Central control is necessarily limited. In the case of WDS, for example, the pre-requisite to join as a member is to comply with a of set standards for data repositories and go through a certification process, but members are not subject to heavy direction or monitoring from the centre but rather more gentle guidance and encouragement.

Domain specific networks often depend on strong support from an influential user community. If a network has proved its value to a user community, they will work to find ways to support it, often through "in-kind" rather than direct support. For example, in the case of IVOA the engagement of the scientific and data provider community has kept it alive for more than 15 years even though the funding of the contributing local initiatives has not been consistent. But not every data network has a large user community with a strong voice or has been around long enough to prove itself to be indispensable.

Networks closely associated with a particular instrument or research infrastructure generally have a clearer or more explicit funding model, but many of the networks indicated that the single biggest barrier to effective data sharing is the myriad of funding agencies and funding schemes that are required to support large scale networks (EMBL, ADNI, WDS, ELIXIR). In addition, the network itself is often structured to match the funding streams rather than optimising for efficiency. For example, there may be a number of national nodes, each with its own national funding lines, perhaps with a central hub funded through an international agreement. Such distributed models of provision require significant co-ordination and the amount of work required for effective integration that delivers seamless access to the combined resources is often underestimated.

4.3. Sustaining co-ordination

Funding is often piecemeal with different sources providing funding for different purposes, and this fragmentation of funding can undermine efficiency and long-term stability. It was frequently reported that it was difficult to obtain funding specifically for data networks, particularly after a start-up phase, and that data networks were only supported as a by-product of research funding. Some data networks were being encouraged to provide paid-for services to generate income, but there were worries about stability of such funding particularly when customers themselves were under financial pressure and pay-for-use charges often equated to simply a recycling of funds from one public budget to another.

Where long-term sustainability has been achieved, it has often come about through a gradual evolution over several decades, starting with ad-hoc collaborations between projects and eventually leading to co-ordinated national funding and potentially formalisation through an international agreement. Formal intergovernmental agreements can assure funding for a defined period of time and provide a basis for long-term sustainability, but such agreements are increasingly rare, tend to be restricted to very large scale infrastructures and they can also have their drawbacks (OECD, forthcoming)

Many of the efforts to optimise co-operative structures and develop multinational governance models for data networks are happening in Europe. Both domain-based research facilities and networked “e-infrastructures” are being designed and implemented as part of the broader vision for a European Research Area. The European Commission has created a specific legal form, a European Research Infrastructure Consortium (ERIC), to simplify multinational co-ordination and governance of European joint-ventures. The members of an ERIC can include non-EC countries and intergovernmental organisations. Several international data infrastructures have adopted the ERIC framework. This is by no means, however, the only way that co-operation can be organised and there are a variety of MoUs and member agreements, with and without formal legal status being used by different international data networks.

Overall, there is a need to define more efficient and coherent governance structures and support mechanism for internationally co-ordinated data networks. In this regard there are opportunities for learning and exchanging of good practices across networks.

4.4. Industry involvement

The level of industrial or private sector involvement in data networks depends on the research domain. Biomedical data networks tend to be most involved with industry. Most of the networks that provide services to both public and private research made no distinction between public and private use, arguing that stimulating private sector competitiveness and innovation was within their mission. However, some networks were considering mechanisms for collecting private contributions towards costs. Sometimes the requirements or desire for commercial partnerships leads to restrictions in data sharing (WDCM) and this needs to be considered relative to overall mission and mandate of a network.

Some successful Public Private Partnerships were reported to be in place for technology provision but some study participants expressed worries about commercial drivers being contrary to the aims of open research and open innovation. Some observed moves by very large software providers to productise the data and expressed worries about building in dependence on commercially provided services (see discussion of these issues in (OECD, forthcoming)). At the same time, it was recognised that industry involvement in basic services such as data storage and development of communication technology is likely to continue to increase over time and this provides opportunities to improve the efficiency and effectiveness of networks.

5. Guiding questions and recommendations

5.1. Guiding questions

The observations described above, notably the diversity of the various networks and the need to respect that diversity, suggest a series of high-level issues or questions that policy makers, funders, network creators and operators need to consider in developing and maintaining well-functioning networks that can accelerate the development of a global data infrastructure for open science. These key questions, introduced in Section 1.3 and listed below, can help policy makers and other stakeholders to make strategic choices and they provide a basis for the recommendations that follow:

When is an internationally co-ordinated data network needed? How does it fit into the existing landscape? How does one assess the performance of the network?

How can governments maximise data openness and reuse within an agreed international framework?

What is the best governance model for a particular network and how can it evolve over time? How are decisions made? What are the roles of the different institutions and people involved in a network and how are they best co-ordinated?

What interoperability arrangements are necessary for the effective operation of an international network and how are they constructed with the network and with other networks?

What business model is appropriate to sustain a network over time?

Each of these issues needs to be considered carefully in designing and supporting a network, but there are few common answers because of the diversity of research data and corresponding networks. Instead, this study provides advice, options, and ultimately specific recommendations that help address these issues in different situations.

5.1.1. When is a data network needed?

Addressing many scientific problems, including any of the grand challenges facing society, requires diverse, complex data to be shared, integrated, and reused across geographies, cultures, scales, and technologies. At the same time, data are becoming ever more complex and voluminous. Governments have strong incentives to maximise the reuse of these data, to reduce duplication of effort, and to improve the efficiency of research. Internationally co-ordinated data networks are essential to enabling and optimising this data sharing and reuse. Policy makers, therefore, need to know when and why they should support the creation or maintenance of such networks, how they should connect with other relevant networks, and how they can ensure efficient and effective operation.

As discussed in Section 2.2, there are many different motivations for creating or joining an international data network, but the central issue should be the definition of a clear research or application demand for data to be shared internationally. The most successful networks have engaged and supportive users who clearly understand and value the services of the network. In some cases, e.g. particle physics or astronomy, the science itself demands international co-operation in that the use of the data can only be done via a distributed system. In other cases, e.g. with human subject data, there may be clear scientific value in having access to larger sample sizes but legal and ethical considerations make network creation and operation more complex. Sometimes it is a matter of filling a gap in data access from a technical or research perspective. Regardless of the data or network type, however, there needs to be a clear, broadly shared value proposition that is regularly assessed.

5.1.2. How can governments maximise research data openness and reuse?

As argued earlier, open data is central to open science. Researchers must be able to access the data from research teams elsewhere in the world to address many of today's research challenges. But the top issue faced by data networks in open sharing of data is the varying attitudes and policies across countries. As with interoperability, fostering openness is a trust building exercise. It is a long process, but governments can do much to accelerate the process by providing a consistent policy framework around *appropriately* open data. This issue drives the important policy recommendations in Section 5.2 below.

5.1.3. What is the best governance model for a particular network?

Different data sources require different data networks — e.g. large physical research infrastructures vs. multiple distributed data sources. Different types of data require different data networks — petascale data, sensitive data, rapid data, etc. Different research communities require different data networks because the cultures of data sharing vary. Simply recognising this diversity is a first step toward defining appropriate governance models that actively engage relevant stakeholders. It is critical to think carefully through the various individual and institutional roles involved, ensure data users and providers have a clear voice, and that there is an appropriate balance between local autonomy and central control within a network. The governance must also be adaptable and able to respond to and facilitate the evolution and growth of a network over time.

5.1.4. What interoperability arrangements are necessary for the effective operation of the network, and how are they constructed?

Interoperability can be defined as the ability of different systems to communicate and exchange and use information. It is the basis of a functional network. It is often viewed simply as an exercise in agreeing on and implementing standards for information access, description, and exchange. In practice, it is a complex process of human negotiations and trust building.

The actual creation and operation of efficient interoperable networks is largely the work of technicians and information professionals in partnership with researchers, but policy makers have a responsibility to ensure the work is done effectively, and they need to be mindful of the human effort involved. Interoperability is still largely viewed as a technical problem, but the most difficult aspects of interoperability are rooted in human relationships and trust.

5.1.5. What business model is needed to sustain a network over time?

As noted in Section 4.1, developing a coherent and sustainable business model is a central challenge for virtually all data networks. This is also an area where governments can have a significant impact by establishing appropriate mandates and incentives (including funding). The Business Models for Sustainable Research Data Repositories report (OECD, forthcoming) discusses economic issues in more detail, but this report makes several recommendations especially relevant to internationally co-ordinated data networks.

When considering these 5 questions, interested stakeholders need to take into account broad policy and planning issues, the basics of building and maintaining useful networks, and overall funding and sustainability issues. A set of recommendations is proposed in each of these areas that can help guide policy but can also increase understanding by policy-makers of the contribution that international data networks make to the scientific enterprise and their critical role as a foundation for open science. This includes appreciation of what makes for a successful network and conversely what is likely to lead to failure.

5.2. Policy and planning

The main barrier to open sharing of curated research data across geographic borders (and scientific domains) is the lack of policy coherence and trust between different communities. This is manifest in different interpretations of openness, different legal regimes for data sharing, and different ethical perspectives. Such differences need to be respected and understood but should not prevent a common understanding and workable international agreements being reached around the sharing of public research data. In this regard, there are many successful examples of sharing data, including what might be considered "sensitive data", internationally, and the lessons from these cases are there to be built on.

There is also a need to simply increase awareness of, and plan for, network services. Several international data sharing partnerships have been in existence for many decades and have become an invisible part of the infrastructure of science. As such the danger is that they are ignored but there is a massive increase in supply and demand for research data in almost all fields of science, and data networks need to be correspondingly up-dated or newly developed and maintained to respond to this. These networks are the business critical point of weakness in many research areas and for open science as whole. Policy makers have key roles to play here to harmonise policy, raise awareness, and plan appropriately.

Recommendation 1: Responsible national authorities should be identified and work toward common definitions of, and agreements on, open data.

At a first level, there needs to be a simple agreement on what is meant by "open data". Much policy work has been done in this area, much of it stemming from the OECD's Principles and Guidelines for Access to Research Data from Public Funding (OECD, 2007) and many of the networks have defined data sharing policies or principles. But there is still no commonly shared definition that reaches across networks, disciplines, and countries. It is important that governments work together to establish a baseline. If countries early on are allowed to opt out of certain provisions, e.g., non-restricted access to data, or define public research data and open in different ways, it creates later barriers to the full use of data.

Open Knowledge International has proposed one definition: "'open knowledge' is any content, information or data that people are free to use, re-use and redistribute — without any legal, technological or social restriction.**"**[2] Discussion around this definition or the most relevant alternatives needs to continue. Furthermore, there is growing recognition that "open", is not enough. This is demonstrated in efforts such as the establishment of FAIR Principles (Findable, Accessible, Interoperable, Reusable)[3] and related initiatives.

Governments need to continue to further their agreement on these issues across all jurisdictions.

Recommendation 2: Governments need to work toward commonly agreed and enforced legal and ethical frameworks for the sharing of different types of public research data.

Of course, not all data can be fully open, but restrictions on access to public data should be based on consensus agreements, taking into account relevant legal and ethical considerations. When establishing a network, there should be an assessment of legal requirements across different countries. Any requirements for restrictions on data access, including differences between legal jurisdictions, have to be addressed during the development of international agreements for data co-ordination initiatives so that they do not become a hindrance at a later stage. In some cases (e.g. GBIF) differences can be addressed with relatively simple measures, such as the universal application of Creative Commons user licenses or waivers (CC-BY or CC0)[4]. In other networks, and particularly when one is dealing with human-subject data, finding a universal solution is much more complex and some exceptions for some data in some circumstances may need to be agreed upon. However, such situations are rarely unique; there are a number of international networks handling sensitive data and considerable opportunities for mutual learning and shared good practice within and across networks.

Protocols for sensitive data, subjects, and specimens have been addressed for some data types, e.g., the Nagoya protocol for the “fair and equitable sharing of the benefits arising from the utilisation of genetic resources” (Convention on Biological Diversity, 2011). Similar protocols are lacking in other sensitive areas of mistrust and exploitation in international collaborations. Funders can support efforts to develop more explicit guidance on expectations and norms for working across cultures. There is growing research and action in this area (Alter and Vardigan, 2015; Kukutai and Taylor, 2016), but governments need to work together to define the formal agreements.

Recommendation 3: All stakeholders need to recognise international research data networks as a critical part of the generic infrastructure for open science.

When infrastructure is working well, it is invisible in daily life. The institutions and participants may be very visible in the community, but the routine data movement and underlying protocols and agreements can easily be taken for granted. Basic, operational infrastructure only becomes clearly visible when it breaks. We only notice electricity when the lights go out.

This quality of transparency makes infrastructure difficult to understand and to continually support. It may be an obvious point, but at a first level, governments must simply recognise the value of data sharing networks to optimising open science. No research program should be developed without also considering how the data will be described, shared, and preserved and how that work will be funded. It cannot simply be assumed that existing networks will pick up new data or that the necessary networks will form to enhance the open reuse of data. Funders need to explicitly consider the networking aspect of data sharing and stewardship in addition to the support for individual repositories and data sources.

Recommendation 4: Responsible national and international authorities must include data networks in long-term strategic planning and support processes for research infrastructure.

The funding requirements and assessment processes for data sharing networks are fundamentally different to those of research projects. Networks must be consistent and reliable and well-engineered. They are not meant to be exploratory. They are meant to be routine and operational. As such, they need to be funded on a different basis and according to different criteria to research projects. International Data Networks are part of the shared infrastructure for research. They should not be funded as short-term research projects or as a by-product of research but as valuable assets in their own right.

This does not mean open-ended funding support. Nor does it mean some sort of "tax" on research funding. It means proper lifecycle management. The Business Models for Sustainable Research Data Repositories report (OECD, forthcoming) delves into this in more detail, albeit with a focus on individual repositories. That report notes that repositories cannot rely solely on structural funding, but that structural funding, combined with other funding sources, can provide incentives and a foundation for sustainable business models. Network support is likely to be more complex because there is need for co-ordination and standards and trust building that reach beyond individual repositories. These sort of activities will likely need some sort of structural support, but this needs to be integrated into longer-term national and international research plans. One suggestion at the project workshop was that networks be considered much like large scientific instruments with planned funding for their development, maintenance and operations, and ultimate retirement — with periodic assessment along the way. This large-scale, full-lifecycle, perspective will become increasingly important with growing data volumes and complexity.

5.3. Building and maintaining useful networks

Developing and maintaining a successful international data network is dependent on a number of factors, both technical and personal. Individual networks need to be tailored to the needs of specific data providers and users and they also need to evolve over time. They require an appropriate mix of long-term commitment, consistency and flexibility. There is no one-size fits all solution, but there are guiding considerations in developing and maintaining networks.

Recommendation 5: In establishing, developing, operating and supporting international data networks the following "organisational" aspects should be taken into account:

Networks should have a clearly defined user and data provider community. At the same time potential requirements of new users, e.g. in terms of interoperability, should not be ignored;

Successful networks have clearly defined user communities and a strong working relationship with those communities. That relationship can be through "local" nodes in a distributed network or through a central service, but it is important that the research community has a strong engagement with the network. Chapter 2 describes different methods of engagement and some examples of effective practice. User engagement ensures that data and services are relevant to research needs, but it also leads to an investment in the network by the community that helps sustain it over time. For example, many of the decentralised networks have substantial "in-kind" contributions from their various member nodes.

Similarly, it is important for networks to establish good working relationships with their data providers. This helps ensure well-described, quality data with appropriate attribution. It also improves data availability and interoperability. For example, GEOSS increased its data availability ten-fold once it started working more closely with its provider nodes to establish flexible, brokering relationships instead of imposing a central standard.

Policy makers should require that networks demonstrate a clear user demand and real applications. It is important, also, that a network demonstrates how it will accommodate data use outside its core discipline.

Networks should have a clear understanding of how they relate to other networks and how they fit into the global research and data sharing landscape

Part of understanding user and provider needs is understanding how a network fits into the larger research enterprise. The landscape of data sources, repositories, networks, and co-ordinating organisations is growing increasingly complex. It is challenging for individual networks, let alone their individual partners, to understand how they fit into this complexity, but the successful networks find a way. Many of them actively engage with collaborative cross-disciplinary organisations such as the Research Data Alliance or International Council for Science, as well as working with their discipline-specific research organisations and structures to continually assess their role in the global research landscape.

Different governance models can succeed, but there is need to define clearly what is to be governed and to involve the data users and providers

Different kinds of governance can be successful. There is no one-size-fits-all solution for network governance, and loose co-ordination can work as well as a strongly top-down system depending on the specificities of each case. There are no general rules on what should be governed centrally and what should be decentralised; but it is important to avoid substantial duplication between central and distributed nodes (some redundancy can be beneficial). It is essential to decide what to govern and what decisions are made where. Concrete use cases can help make these decisions, but they do not make a governance system. It is necessary to take a holistic view.

Regardless of the model both data providers and users (and in some cases funders) need to be involved in the governance structures of a network to ensure the network remains relevant and responsive. This can be through direct involvement in Board level governance or through inclusion in various advisory and working groups.

Roles and responsibilities across a network, e.g. for data curation versus service provision, must be clearly defined and, where possible, data champions identified

It is important to designate roles and responsibilities at the inception of a network. As with governance, it is important to clarify what services are provided centrally and what are provided locally. To obtain full collaboration, a network should provide clear accreditation and value for the participants. The governance model must anticipate non-compliance and have a transparent process for to deal with erring members.

It can also be very useful to designate specific co-ordinators and data "champions". A given network may need to interact with different institutions (ministries, funding bodies, research organisations, etc.) in a given country, in which case governments should designate a central contact person who can help the agencies work together and represent them collectively. It is not uncommon for governments to designate a lead agency to co-ordinate national participation in a network. At the same time, some networks have identified their own national contacts. Greater co-ordination, recognition, and empowerment of these data-co-ordinating and ambassadorial positions can greatly increase network effectiveness.

It could similarly be useful for there to be a network of officials in agencies who can communicate around data issues and support the overall cause of open data. See Recommendation 8.

The necessary level of standardisation needs to be clearly defined taking into account user requirements and the need to maintain an appropriate balance of autonomy amongst network nodes and central control

There is no single path to interoperability. As discussed in Chapter 2, different networks will take different approaches that are appropriate to their type of network, data, and community. What is important is that there is clear understanding of the level of standardisation necessary to meet the needs of all network members. A strict, top-down mandate for all aspects of system interoperability rarely works, and most networks try to avoid being too prescriptive. Networks typically need to allow for different types of contributors with varying budgets, capacity, and skills. Often a “network of networks” approach with different types or levels of nodes can provide the necessary flexibility to accommodate diverse contributors.

Inevitably, some (growing) level of central standardisation must be defined, but there needs to be a balance of autonomy and central control. In general, broader or more generic issues are better dealt with centrally and more user-facing and more specific or disciplinary issues managed in a more distributed way.

Regardless of the approach, funders and practitioners should recognise that time and investment is necessary to negotiate an appropriate level of standardisation within a network and with other networks and systems as well. This requires workshops and committees and engagement with users and relevant external bodies such as the Research Data Alliance, CODATA, and global and disciplinary standards bodies. There also need to be dedicated resources and mechanisms to enable skills development and the exchange of good practice within the network.

Objective and transparent mechanisms for network assessment should be based on standard quality management frameworks and certification of participating data repositories

Standard criteria for quality and performance assessment help repositories and other network partners to ensure useful services and good collaboration. For example, WMO has developed formal Quality Management Frameworks for National Oceanographic Data Centres (NODCs)[5] and the National Meteorological Services and Information Services (WMO-IS)[6].

There are also three major data repository certification schemes:

- WDS-DSA Core Trustworthy Data Repositories Requirements[7]
- nestor-Seal DIN 31644[8]
- ISO 16363[9]

Other quality management protocols, such as the ELIXIR Core Data Resources[10] (presented at the Brussels workshop) can also provide a basis for network assessment.

Consideration needs to be given to regional differences in capacity - infrastructure and human resources - and in culture, community needs and expectations.

Different countries and communities have different levels of technical skills and capacity. Networks need to work collaboratively to build skills and capacity where they are most needed. This needs to be viewed as an important function of the network as part of the core purpose of making more data more usefully available.

5.4. Funding and long term sustainability

Intergovernmental agreements (IGAs) can be useful in providing a sustained framework that enables networks to effectively plan and co-ordinate, but there is significant cost in establishing IGAs and they can potentially restrict network flexibility and adaptability. Whilst IGAs underpin the data network operations of a small number of large-scale research infrastructures (eg. CERN or EMBL-EBI) they are not applicable for the large majority of international data networks. Instead a variety of lighter MoUs or co-operation agreements that may enshrine funding commitments have been adopted by different networks.

Overall, the current funding arrangements for international data networks are inadequate. Despite the increasing demand for data curation, data repositories are struggling for support in many countries and open access to valuable data has in some cases been lost (see the companion report of Business Models for Sustainable Research Data Repositories (OECD, forthcoming)). There are additional funding and sustainability challenges for international data networks, many of which have no obvious sponsor for their critical central co-ordination functions. Often there is a considerable amount of uncosted “in kind” support involved in networking and the amount of extra funding required for co-ordination is small but this should not mean that it is neglected. There can be considerable overall cost benefits from federation of activities. At the same time, many networks do not have clearly articulated value propositions that can be used to justify additional investments.

Recommendation 6: Funders and host institutions should view internationally co-ordinated data networks as a long-term investment and support them accordingly.

The current funding model for most data networks is inefficient and subjects those networks to unnecessary risk. Most funding is ad hoc or project based and lacking international co-ordination. Many of the networks began as bottom up efforts with short term resourcing, but now they suffer from the lack of harmonisation or sustained commitment across the funders and government agencies that support their operation. Improving this co-ordination is an area ripe for policy intervention.

Dedicated funding needs to be provided that is sufficient to support the co-ordination functions and core operations of a network without the need to augment it with piecemeal project based funds. As discussed above, under Recommendation 4, sustainable funding models for networks are likely to be different to those for individual repositories. Long-term commitments to data infrastructure need to be made at national level because these form the basis of international infrastructure. These national commitments should include dedicated resources for international co-operation and co-ordination.

Sustainability is not only about hardware and systems, human resources are equally important. This includes funding for organising workshops, training, and reaching out to the users, as well as for committed and supported staff. Support for governance and working group meetings is often difficult to obtain but can be critical in ensuring the good functioning of a data network.

Recommendation 7: Networks should have clear business models, including value propositions and measures of success relevant to their different stakeholders. These measures should be monitored.

The Business Models for Sustainable Research Data Repositories report (OECD, forthcoming) lays out first level requirements for developing sustainable business models for individual repositories. Business models for data networks (federations of repositories) need to address similar requirements. There is also an additional requirement to justify the need for co-ordinated and standardised sharing of data internationally. The value proposition and therefore the business model will be strongest when there is a strong research or application-driven need for networked services. Funders, data providers, and data users must all see shared value in the network.

Networks need to carefully consider exactly what services they are providing and what services are provided at what level within the network. It is not just access to data, but how the data are accessed, how and where they are aggregated and processed, and how the data will be used in the future. Defining the services helps clarify roles and the level of standardisation and provides a clear measure for the funder and other stakeholders on what to expect from their investment in the network. It also helps emphasise what is necessary to provide those services, which can have implications across the full data life-cycle from generation and/or collection to archiving. Monitoring how services are used can show trends or changes in user activity and can help guide future system and network development. The best networks continually evolve with changing user needs

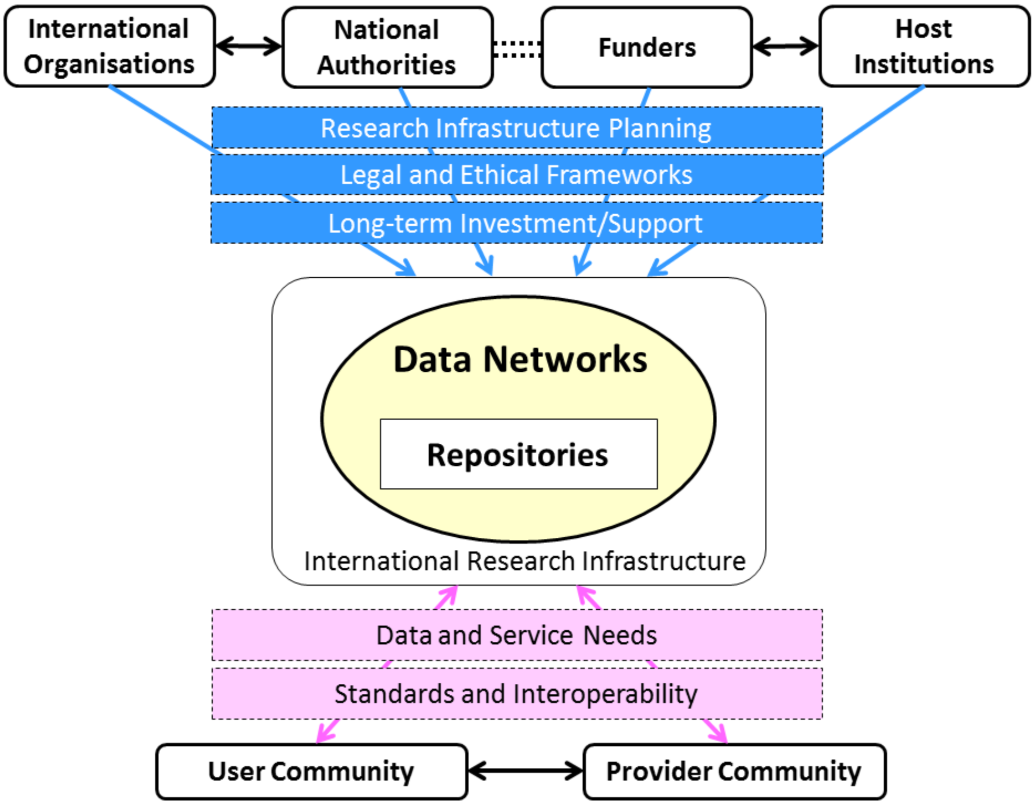
It is perhaps self-evident that government investments should be aligned with explicit performance indicators and other measures of success. Different networks will have different success metrics and they are not all quantitative, but it is important that they are defined and agreed upon by all stakeholders. Problems arise when there is mis-alignment in funder or political and research objectives. It is important that the agreed metrics are routinely monitored and periodically reassessed. Performance assessment of international data networks is an area in which there are considerable opportunities for mutual learning across different networks and agencies, who should continue discussions on what make good measures of success.

Recommendation 8: Funders should actively participate in relevant international discussions and forums to improve long-term functioning, support and co-ordination of data networks.

Funders need to co-ordinate their activities around data networks as much as researchers, technicians, and repositories do. Funders need to work with their beneficiaries and international partners to explore predictable financial support mechanisms for roughly seven- to ten- year timeframes.

Experience in both the RDA and Belmont Forum e-Infrastructure and Data Management program suggest that even relatively informal co-operation arrangements can be very beneficial, especially if they involve people charged with budgetary authority and strategic planning. For example, the “Funders Forum” in RDA provides an opportunity for both public funders of different types and private foundations to have informal discussions every six months on issues ranging from data management plan compliance to exploring cloud technologies or longer term data policy. RDA funders have found value in this forum and actively participate. These informal discussions cannot replace formal arrangements such as IGAs and MoUs for individual networks, but they can do a lot to build trust, creativity, and co-ordination.

Figure 1. Key actors and responsibilities for international research data networks


International Organisations
National Authorities
Funders
Host Institutions
Research Infrastructure Planning
Legal and Ethical Frameworks
Long-term Investment/Support
Data Networks
Repositories
International Research Infrastructure
Data and Service Needs
Standards and Interoperability
User Community
Provider Community

Source: Authors' analysis.

5.5. Concluding remarks

Reflecting the increasing diversity and complexity of research, data sharing networks are very diverse in their governance and operational methods and increasingly complex in their data and service provision. Funders and policy makers need to respect this diversity and complexity and take a multifaceted approach to supporting the operation of networks. The networks themselves need to be increasingly adaptive and entrepreneurial.

There are multiple effective approaches to governing and operating data sharing networks, so it is important to ask the right questions when choosing policy and operational strategies. Policy makers, funders, and network operators all need to take a holistic view and consider the data user and provider perspectives from all angles. Data sharing is an on-going, evolutionary process. The point is not only to support specific research projects but to contribute to a broad and transparent global data infrastructure comprised of many diverse networks.

Endnotes

1. RDA Indigenous Data Sovereignty Interest Group: https://www.rd-alliance.org/groups/international-indigenous-data-sovereignty-ig.
2. http://opendefinition.org.
3. https://www.force11.org/group/fairgroup/fairprinciples.
4. https://creativecommons.org/share-your-work/public-domain/freeworks.

5 https://www.iode.org/index.php?option=com_content&view=article&id=415:04-feb-2014-iode-quality-management-framework-for-nodcs-published&catid=23&Itemid=115.

6. http://www.bom.gov.au/wmo/quality_management/docs/Doc_3_Annex_I_QMF-circ_en.pdf.
7. https://www.icsu-wds.org/news/news-archive/wds-dsa-unified-requirements-for-core-certification-of-trustworthy-data-repositories.
8. http://www.dnb.de/Subsites/nestor/EN/Siegel/siegel.html.
9. https://public.ccsds.org/pubs/652x0m1.pdf.
10. https://www.elixir-europe.org/platforms/data/core-data-resources.

****

References

Alter, C. G. and Vardigan, M. (2015), “Addressing global data sharing challenges”, *Journal of Empirical Research on Human Research Ethics*, Vol. 10 (3) 2015, pp. 317-323, SAGE Publications, http://dx.doi.org/doi:10.1177/1556264615591561.

Arzberger, P. et al. (2004), “An International Framework to Promote Access to Data*”*, *Science*. Vol. 303, AAAS, Washington D.C., pp. 1777-1778. March 19, 2004.

CASRAI (2015), *E-Research infrastructure*, http://dictionary.casrai.org/E-Research_infrastructure.

Convention on Biological Diversity, United Nations (2011), *Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from Their Utilization to the Convention on Biological Diversity*, https://www.cbd.int/abs/text/.

Edwards P. N. et al. (2007), *Understanding Infrastructure: Dynamics, Tensions, and Design*, http://hdl.handle.net/2027.42/49353.

European Strategy Forum on Research Infrastructures (ESFRI) (2016), *Public Roadmap 2018 Guide*, http://www.esfri.eu/sites/default/files/docs/ESFRI_Roadmap_2018_Public_Guide_f.pdf.

GRDI2020 Consortium (2010), *GRDI 2020, Towards a 10-Year Vision for Global Research Data Infrastructures*, http://www.grdi2020.eu/Repository/FileScaricati/e2b03611-e58f-4242-946a-5b21f17d2947.pdf.

G8+05 Global Research Infrastructure Sub Group on Data (2011), *Draft Report, Global Research Infrastructure Sub-Group on Data*, https://epubs.stfc.ac.uk/work/24111652.

Harmon, A. (2010), "Indian tribe wins fight to limit research on its DNA" in the New York Times. 21 April 2010. http://www.nytimes.com/2010/04/22/us/22dna.html?pagewanted=all.

Kukutai, T. and Taylor, J. (2016), *Indigenous Data Sovereignty Toward An Agenda, Centre for Aboriginal Economic Policy Research (CAEPR),* ANU Press, Canberra, http://www.jstor.org/stable/j.ctt1q1crgf.

Le Treut, H. et al. (2007), "Historical Overview of Climate Change", in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change,* Cambridge University Press, Cambridge and New York, https://www.ipcc.ch/pdf/assessment-report/ar4/wg1/ar4-wg1-chapter1.pdf.

National Research Council (NRC) (1993), *Toward a Co-ordinated Spatial Data Infrastructure for the Nation,* National Academies Press, Washington, D. C., http://books.nap.edu/catalog.php?record_id=2105.

National Science Foundation (NSF) Blue-Ribbon Advisory Panel on Cyberinfrastructure (2003), *Revolutionizing Science and Engineering Through Cyberinfrastructure*. https://www.nsf.gov/cise/sci/reports/atkins.pdf.

NSF Cyberinfrastructure Council, (2007), *Cyberinfrastructure Vision for 21st Century Discovery*, https://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf.

NSB, (2005), L*ong-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*, http://www.nsf.gov/pubs/2005/nsb0540/.

OECD (2017, forthcoming), "Business Models for Sustainable Research Data Repositories," *OECD Science, Technology and Industry Policy Papers*, OECD Publishing, Paris.

OECD (2015), "Making Open Science a Reality, Science", *OECD Science, Technology and Industry Policy Papers*, No. 25, OECD Publishing, Paris, http://dx.doi.org/10.1787/5jrs2f963zs1-en.

OECD (2014), *International Distributed Research Infrastructures: Issues and Options*, OECD Publishing, Paris,

http://www.oecd.org/sti/sci-tech/international-distributed-research-infrastructures.pdf.

OECD (2007), *OECD Principles and Guidelines for Access to Research Data from Public Funding*, OECD Publishing, Paris, http://www.oecd.org/sti/sci-tech/38500813.pdf.

Research Councils UK. (2014), *e-Infrastructure*, http://www.rcuk.ac.uk/research/xrcprogrammes/otherprogs/einfrastructure/.

Star, L. S. and Ruhleder K. (1996), Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, Vol 7 (1), pp. 111.

Appendix 1. OECD-GSF/ICSU-WDS Expert Group

Country/ nominating organisation	Name	Affiliation
UK and WDS/CODATA	Juan Bicarregui	RCUK / Science and Technology Facilities Council
Netherlands and WDS	Ingrid Dillo	Dutch Data Archiving and Networking Services (DANS)
Portugal	João Nuno Ferreira	Technical Director at the Foundation for National Scientific Computing
France and WDS	Francoise Genova	Strasbourg Astronomy Centre
WDS/CODATA (China (People's Republic of))	Li Jianhui	Computer Network Information Centre (CNIC), Beijing, China (People's Republic of)
European Commission	Wainer Lusoli	EC, DG Research and Innovation
RDA (India)	Devika P. Madalli	Professor, Documentation Research and Training Center (DRTC), Indian Statistical Institute (ISI)
INCF (USA)	Maryanne Martone	National Centre for Microscopy and imaging, Health Sciences, UCSD
WDS	Mustapha Mokrane	Director, WDS Programme Office, NICT, Tokyo
Japan and WDS/CODATA	Yasuhiro Murayama	Director of Integrated Science Data System Research Laboratory of National Institute of Information and Communications Technology (NICT)
Korea	Seo-Young Noh	GSDC (Global Science Data Center), KISTI (Korea Institute of S&T Information)
Korea	Sun Kun Oh	Professor of Physics, Konkuk University
South Africa and WDS/CODATA	Dale Peters	Director, UCT eResearch, University of Cape Town
Norway and WDS	Benjamin Pfeil	Ocean carbon data centre, Norway
Switzerland	Thomas Schulthess	Swiss National Supercomputing Center
South Africa	Happy Sithole	Director, Center for High Performance Computing, Council for Scientific and Industrial Research (CSIR)
Finland and WDS/CODATA	Sanna Sorvari **(co-chair)**	Finnish Meteorological Institute
Australia	Andrew Treloar **(co-chair)**	Australia National Data Service and co-chair of RDA Technical Advisory Board

Appendix 2. Persons interviewed for case studies

	Network	Interviewee	Affiliation
1	ADNI	Eric T. Meyer	Professor of Social Informatics and Director of Graduate Studies, Oxford Internet Institute
2	DARIAH	Laurent Romary	Director
3	Elixir	Andrew Smith	External Relations Manager
		Marine Gabory	Trainee
4	EMBL-EBI	Rolf Apweiler	Director
5	GBIF	Donald Hobern	Executive Secretary
6	GEO	Barbara Ryan	Secretariat Director
		Paola De Salvo	Information Technology Officer
7	H3ABioNet	Sumir Panji	Network Manager
		Nicola Mulder	Professor, Computational Biology Group, University of Cape Town
8	ICSU-WDS	Sandy Harrison	Scientific Committee Chair
9	IUGONET	Takuji Nakamura	Steering Committee Chair
		Toshihiko Iyemori	Steering Committee Vice-chair
		Yoshimasa Tanaka	National Institute of Polar Research
10	IVOA	Christophe Arviset	Chair of the IVOA Executive Committee (European Space Agency)
		Robert J. Hanisch	Director, Office of Data and Informatics, NIST, USA (First Chair of the IVOA Executive Committee)
11	WDCM-GCM	Juncai Ma	Professor, Institute of Microbiology, Chinese Academy of Science
		Philippe Desmeth	President of World Federation of Culture Collections (WFCC)
		Hideaki Sugawara	Former WDCM director

Appendix 3. Workshop on International Co-ordination of Data Infrastructures for Open Science, Brussels, 30-31 March 2017

Invited workshop participants*

Name	Affiliation
Carmela Asero	European Comission
Kevin Ashley	Digital Curation Centre (DCC), Scotland
María Guillermina D'Onofrio	Ministry of Science, Technology and Productive Innovation, Argentina
Koenraad De Smedt	Common Language Resources and Technology Infrastructure (CLARIN), Norway
Michael Diepenbroek	Alfred Wegener Institute for Polar and Marine Sciences, Germany
Kylie Emery	Australian Research Council
Robert Hanisch	National Institute of Standards and Technology (NIST), USA
Bjorn Henrichsen	Norwegian Centre for Research Data (NSD)
Tim Hirsch	Global Biodiversity Information Facility (GBIF)
Simon Hodson	ICSU-CODATA
Robert Jones	Helix Nebula, Switzerland
Tibor Kalman	Digital Research Infrastructure for the Arts and Humanities (DARIAH)
Mark Leggott	Research Data Canada
Johanna McEntyre	European Bioinformatics Institute (EMBL-EBI)
William Miller	Office of Advanced Cyberinfrastructure, US National Science Foundation
Cameron Neylon	Curtin University
Sumir Panji	H3ABioNet, South Africa
Mark Parsons	Research Data Alliance (RDA)
Barbara Ryan	Global Earth Observation System of Systems (GEOSS)
Robert Samors	Belmont Forum
Andrew Smith	ELIXIR, UK
Min-Ho Suh	Korea Institute of Science and Technology Information (KISTI)
Jostein Sundet	NordForsk, Norway
Jorge Tezon	CONICET, Argentina
Paul Uhlir	Consultant on information policy and management
Jean-Pierre Vilotte	Belmont Forum
Matthew Woollard	Consortium of European Social Science Data Archives (CESSDA), UK

*Expert Group members and OECD staff (see Appendix 1) participated also in the workshop

Appendix 4. Questions for the initial survey

Introductory

I1: How do you define the role and mission of your data infrastructure network?

I2: How and when did this network come into existence?

I3: What are the aims of the network?

Descriptive

D1: How many nodes are in the network and in what countries? Does it have a central node?

D2: Who are the beneficiaries of the network?

D3: What disciplines is your network related with?

D4: How large is the network community?

D5: Approximately how many staff including data scientists work for the network?

Operational

O1: What is the governance of the network?

O2: How is the network financed?

O3: What is the operating framework (context and standards etc) for the network? Please either summarise or provide a link.

O4: What is the structural model for the network and the role(s) of its respective parts?

O5: How is the network evaluated? (Regular evaluations, key performance indicators, etc)

Reflective

R1: How does the network define open science?

R2: What do you see as the main strengths of the network?

R3: What do you see as the main challenges and obstacles for the network?

R4: What have been the milestones for the evolution of the network and what is its vision for the future?

R5: What have been the key step changes in data stewardship and/or sharing in your domain over the past decade and how do you see this evolving over the next few years?

R6: Have there been any changes to the business model of your network and/or its nodes? If so, what have been the most significant changes?

Appendix 5. Case study interview questions

A - Implications, advantages and operational challenges of being an international network

1. Value added user services versus basic data curation (or levels of service provision, e.g. bronze, silver, gold) - how are different tasks divided across the network? What are the advantages of centralised versus distributed systems?
2. What do you see as the value of providing different service levels? Do your users find these different levels useful?
3. Global network building - what are the trans-national challenges when countries that have v. different regimes and cultures, i.e. beyond Europe? This can include behavioural norms, legal issues, IPR etc.?
4. What are the benefits/motivation for individual repositories to be in a network (noting that most of the case studies have been submitted from the perspective of the network co-ordinator)?
5. How does the issue of trans-border data flows affect your network?

B - Steering the network: funders and researchers community

What is the influence of research traditions culture and historical norms and frameworks in different disciplines and fields of research and how these affect data sharing and openness?

1. Please comment on the role of international, national (and institutional) policies and funder mandates in driving uptake and user behaviour.

C- Limits and challenges related to openness

1. Defining limits on openness - what data to keep and what to make accessible - how is this agreed and implemented across an international network?
2. What are the challenges in providing access to human subject/personal data, including ethical issues, with a focus on how these are dealt with/harmonised across borders?

D – Governance-related issues

1. Governance - particularly for networks whose governance has changed, re what are the challenges and benefits of different systems. If you have changed governance, what were the drivers for this, and were the outcomes of the change as you anticipated?
2. In your view what are the advantages and disadvantages of different structures and legal entities (linked to q10 above). NOTE: This question is intended to allow the respondent to reflect on the whole topic of structures and governance.

E - Sustainability and private sector involvement

1. Discuss your sustainability and funding models, including “glue funding” for central nodes, where applicable, and funding for distributed network components.

2. What is the involvement of the private sector currently and in the future (roles might include: funders, data providers, data users, governance participation, development partners for services etc.)?