AUSTRALIA'S DATA-ENABLED RESEARCH FUTURE: SCIENCE

Australian Academy of Science

JUNE 2022

A collaboration between ARDC, ACOLA and Australia's Five Learned Academies

Sponsored by

ARDC

Australian Research Data Commons

Acknowledgement of Country

The Australian Academy of Science acknowledges and pays respect to the Ngunnawal people, the Traditional Owners of the lands on which the Academy office is located. The Academy also acknowledges and pays respects to the Traditional Owners and Elders past, present and emerging of all the lands on which the Academy operates and its Fellows live and work. They hold the memories, traditions, culture and hopes of Aboriginal and Torres Strait Islander peoples of Australia.

ISBN 978-0-85847-871-8

Cite this report as:
Australian Academy of Science (2022).
Australia's data-enabled research future: Science

Australian Academy of Science
GPO Box 783
Canberra ACT 2601

Tel +61 (0)2 6201 9400
Email aas@science.org.au
www.science.org.au

Australian Learned Academies Future of Research Data project

This project is the result of a partnership between the ARDC, Australia's five Learned Academies and ACOLA to ensure Australia can undertake excellent data-enabled research across all fields of research. Notably, the project sought to help build a more coherent data policy and strategic data planning environment to uplift national data infrastructure. Five domain reports were developed, and a synthesis report focused on common themes and multidisciplinary opportunities and needs. We hope that this project will transition into an ongoing national data policy and strategic planning capability.

Australian Academy of Science

The Australian Academy of Science provides independent, authoritative and influential scientific advice, promotes international scientific engagement, builds public awareness and understanding of science, and champions, celebrates, and supports excellence in Australian science. The Academy is a not-for-profit organisation of individuals elected for their outstanding contributions to science and research.

Acknowledgements

This project received investment from the Australian Research Data Commons (ARDC). The ARDC is supported by the National Collaborative Research Infrastructure Strategy (NCRIS).

The Academy gratefully acknowledges the contribution of the Steering Committee members: Professor Jane Elith, Professor Ginny Barbour, Dr Danny Kingsley, Professor Andy Pitman, Dr Lesley Wyborn, Professor Ian Chubb, Anna-Maria Arabia and Chris Anderson.

Project management, research and writing were provided by the Australian Academy of Science secretariat. Contributing staff members Lauren Sullivan, Alexandra Lucchetti, Chris Anderson, Hayley Teasdale, Stuart Barrow, Robyn Diamond and Leah Albert are gratefully acknowledged.

Acronyms and abbreviations

the Academy	the Australian Academy of Science
ACOLA	Australian Council of Learned Academies
ARC	Australian Research Council
ARDC	Australian Research Data Commons
CARE	Collective benefit, Authority to control, Responsibility and Ethics
FAIR	Findable, Accessible, Interoperable and Reusable
IMOS	Integrated Marine Observing System
NCI	National Computational Infrastructure
NCRIS	National Collaborative Research Infrastructure Strategy
NHMRC	National Health and Medical Research Council
NOAA	National Oceanic and Atmospheric Administration
NRI	National Research Infrastructure
the NRI Roadmap	the 2021 National Research Infrastructure Roadmap
TERN	Terrestrial Ecosystem Research Network
WDS	World Data System
WMO	World Meteorological Organization

Contents

1. Executive summary

Data are fundamental to scientific research. Scientists collect, analyse and interpret data to draw conclusions and make predictions which lead to greater understanding of our world and provide the knowledge to address our biggest domestic and global challenges.

Research and data support governments to make timely, informed decisions on complex issues, such as responses to pandemics and natural disasters like bushfires. Exploring data also enables discovery research that leads to new knowledge and technological innovation.

Research data infrastructure encompasses the facilities, equipment, tools, people and data policies that enable researchers to generate, access, manage and use data. Coordinated, strategic planning for data infrastructure is essential to support research in Australia and international collaboration.

In 2021, the Australian Research Data Commons (ARDC) partnered with the five Australian learned academies and the Australian Council of Learned Academies (ACOLA) to explore data infrastructure needs to facilitate excellent data-enabled research across science, health and medical sciences, humanities, social sciences and engineering and technology.

This report presents data-related needs for science research captured by the Australian Academy of Science through consultations with researchers and other experts across a range of science disciplines. It complements and reinforces findings of the report *Advancing data-intensive research in Australia* published by the Academy in 2021.

The strategic research data needs and challenges identified in this report include:

- **Driving greater national coordination and integration across Australia's research data infrastructure** to improve data access and interoperability between research disciplines, government and industry. This requires careful consideration of existing architecture and repositories, and streamlined access to government data for research in minimally processed, non-proprietary and machine-readable formats. A review and reframe of national research data policy and development of clear, strategic national research data priorities is needed to drive coordination and development of national data infrastructure.
- **Developing consistent and enforceable data policies and standards** to facilitate access, interoperability, responsible use, reuse and sharing of data between research, government and industry.
- **Promoting data sharing by recognising data and software sharing in academic success metrics and investing in the people** and infrastructure to manage data and make data FAIR. Data, code and software sharing enables collaboration and supports research integrity, but requires time and expertise to curate data, provide quality metadata and make them interoperable.
- **Addressing challenges presented by expanding volumes of data and data-intensive research activities**, such as moving, manipulating and analysing large amounts of data, data storage and retention. Data infrastructure development will need to consider these issues to serve the future needs of research.
- **Developing a digitally skilled research workforce as an urgent priority** to underpin effective data infrastructure and data-intensive research. Building this capability will require both investment in data experts and raising the data skills of researchers. Appropriate funding, attractive employment and recognition for data science expertise are needed to attract and retain data experts in the research sector.

This report presents recommendations for action and leadership to address these urgent research data issues. Acting on these recommendations is essential to ensure Australia's ability to predict and rapidly respond to societal challenges and crises, to leverage opportunities such as digital health and advanced manufacturing, and to collaborate internationally.

2. Introduction: Australia's science research data ecosystem

Data are transforming science research, enabling new, complex analyses on large-scale datasets and exploration of research questions that were not previously possible. Generating, accessing, analysing and interpreting, and managing data is becoming central to the practice of science and scientific collaboration.

The ability to access and use quality, reliable data is essential for timely, informed decision-making on critical issues in the national interest such as public health, disaster response, energy and food security, infrastructure planning and environmental management. Data are an important strategic national asset, but investment in a cohesive data infrastructure system to support researchers and other stakeholders to use data is essential to realise its value.

In 2021, the Australian Academy of Science (the Academy) published the report *Advancing data-intensive research in Australia,* which summarises and discusses the issues facing Australia's research community as research across all disciplines becomes more data-intensive. It includes an international perspective on data policies and practices, and presents actionable recommendations.[1]

Advancing data-intensive research in Australia found that the nation's current science research data ecosystem is fragmented, with a complex network of data infrastructures, data assets and policies across disciplines and institutions. The report determines that to enable the opportunities afforded by data-intensive research, Australia's research data policies require greater coordination and enforcement, an integrated eResearch infrastructure, universal adoption of data sharing principles and practices, access to public government data for research, improved interoperability, and a digitally skilled research workforce.

This report reinforces these findings and provides additional insights into the coordination and integration, data management, policies and standards and skills needed to maximise the use of existing research and data infrastructure and inform planning and development to serve future scientific research. Section 3 of this report illustrates the benefits and opportunities that could flow from transformative national data infrastructure.

Recent policy developments reflect the importance of data in policymaking, research, the economy and society, including the Australian Government's *Digital Economy Strategy*,[2] the *Australian Data Strategy*[3] and the Australian Chief Scientist's work on a national open-access strategy[4]. Data are recognised as a key part of open scientific knowledge in the UNESCO Open Science Recommendation, which Australia has adopted.[5] These developments present an opportunity for data-related discussions between government, research and industry to guide strategic planning for coordinated national research data infrastructure.

The critical need for coordinated and integrated national research data infrastructure to support excellent research is acknowledged in the *2021 National Research Infrastructure Roadmap* (the NRI Roadmap), which recommends a National Digital Research Infrastructure Strategy to coordinate existing infrastructure, guide investment and prepare for future data-related opportunities.[6] Section 4 of this report presents current and emerging research data needs, gaps, weaknesses, risks and opportunities in the natural and physical sciences, informed by discussions with researchers from a range of science disciplines. These issues should be considered in the strategy and in national research infrastructure development.

2.1 Background and objectives

This report is part of a collaborative project between the Australian Research Data Commons, Australia's five learned academies and the Australian Council of Learned Academies, which seeks to understand Australia's data infrastructure, assets, policy, and skills needs for research. Through this collaboration the learned academies are developing a cohesive data agenda to provide leadership, advocacy and planning for research and establish a network of data policy and planning capability to support strategic planning for national research data infrastructure.

This report from the Australian Academy of Science provides an overview of national strategic data infrastructure, capabilities, policies and skills needed to support science research in Australia. The aims of the report are to identify:

- strategic data-related needs and requirements of research domains
- existing capability, gaps, challenges and opportunities
- potential opportunities for leadership, advocacy and planning for national data infrastructure.

What is research data infrastructure?

Research data infrastructure refers to the 'facilities, equipment and tools that serve research through data generation, data manipulation and data access.'[19]

In this report, 'data infrastructure' is used broadly to include data, repositories and other digital research infrastructure, tools and services, hardware and software, skills and data policies.

'Research data' refers to data generated by, collected or accessed for research activities.

2.2 Project framework and approach

The gaps, challenges and opportunities outlined in this report were identified through consultations with researchers from a range of disciplines. Three types of research – Discovery, Understanding and Prediction – were described to guide the discussions and reflect the different scales, complexity, requirements, aims and stakeholders of different research domains and projects.

Definitions of research types used in consultation workshops

DISCOVERY

Discovery in research is concerned with targeted projects where quantitative or qualitative data may be collected or collated then analysed to describe a particular phenomenon. Data may also be explored to uncover new research questions and form new hypotheses.

UNDERSTANDING

Understanding in research refers to research and analyses that deepens understanding of or explains systems or processes. The focus is on larger scales, greater complexity and broader data analyses than may be considered for research aimed at 'discovery'. Research projects may involve working with quantitative or qualitative data.

PREDICTION

Prediction in research covers predictions which are made from various types of models. Predictions may be made once or continually updated, they might be spatial or temporal, they may have one or millions of stakeholders and users. Predictions might be discarded or kept and managed so they are insulated against changes in technologies or software.

Three workshops were held, each focusing on one of the defined research types. Participants were presented with narratives that provided examples of relevant research scenarios (Appendix 1). The narratives were intended to stimulate discussion of data needs, gaps, weaknesses, risks and opportunities in science research, with participants asked to reflect on relevant data issues within their field of research. The outcomes of these discussions informed this report, and reinforce many of the findings and recommendations from *Advancing data-intensive research in Australia*. The consultation methodology is detailed in section 7 Process and Consultation. A summary of key ideas, gaps and opportunities raised in the discussions are provided in Appendices 2 and 3.

3. Why do data matter?

Data are fundamental to scientific research. Scientists collect, analyse and interpret data to draw conclusions and make predictions which lead to greater understanding of our world and provide the knowledge to address our biggest domestic and global challenges.

Modern societies rely on scientific research data for informed, evidence-based decision-making, better connected services, to de-risk industry investment and for the discovery that leads to technological innovation.

Research and data enable governments to make critical decisions on complex issues, for example, observational data from the Bureau of Meteorology inform responses to natural disasters like bushfires and floods and planning for drought. Data exploration also enables blue-sky research leading to new knowledge, such as data collected by Australia's astronomy facilities allowing research into the origins of our universe.

Research data are also reused to answer new questions, but must be curated, stored and maintained, made findable, and accompanied by metadata to enable use. Once collected, particularly in longitudinal studies, data need to be migrated as inevitable changes in hardware, software and data standards evolve.

The following examples illustrate the value of data and efforts to develop a more strategic, coordinated data infrastructure ecosystem for Australia.

Data infrastructure for biomedical research

The interrogation of large databases is ushering in a new era of biomedical and health research discovery and enhanced healthcare. The appropriate data infrastructure could enable integration of large numbers of genome sequences with clinical information, pharmaceutical records and physiological data from smart sensors (and other sources). This would allow machine learning to detect correlations between genes, lifestyle and disease. Such correlations will have immediate clinical applications in identifying preventable conditions and disease risk, enhancing accuracy in diagnoses and treatments, and developing new therapies.

With transformative data infrastructure, it will be possible to compare environmental and genetic parameters against disease incidence and progression and understand factors that contribute to disease. Such advances are dependent on data availability. Presently, relevant data are held in various places and are not easily accessible. What is required is essential infrastructure to collect, house and make these data available, and the ethical, privacy and security frameworks to underpin it.

Providing this integrated evidence-based information could revolutionise the quality and effectiveness of healthcare, leading to improved community wellness, quality of life and national productivity, as well as greater efficiency and cost-effectiveness of healthcare systems.

Digital geoscience opportunities

We rely on Earth's solid interior to survive and prosper. It is a geologically complex system that controls the availability of groundwater, energy, and minerals resources. Geoscientists collect large amounts of digital observational data to understand the Earth and its systems, yet often these datasets are inaccessible, not fully integrated and specific to geoscience subdomains or geographical location. Improved data management practices would allow holistic analysis of an integrated, entire Earth system and enable integration with data from broader disciplines, for example from the social, biological, environmental or atmospheric sciences.

The majority of Australia's identified and exploited mineral resources have been discovered through surface exploration, but future discoveries will require detailed understanding of processes beneath Earth's surface. Development of an integrated predictive geoscience capacity in the next decade would allow us to understand the evolution and predict the behaviour of Earth's complex systems through time.[7,8] This capability could enable identification of significant mineral resources and address current shortages of critical minerals, including copper, lithium and cobalt.[7]

A predictive understanding of how geohazards (such as earthquakes) occur, and how activities such as human-driven sedimentary basin extraction and storage activities might affect other natural resources (including groundwater) or trigger geohazards, would enable sustainable management of our complex and interacting natural resource systems. Integrated geoscience data systems that are FAIR, with datasets from across geoscience sub-disciplines, geographical locations and from government, academia and industry will enable this predictive capacity in Australia.

A unified national biosecurity data system

Australia has a unique environment and native species found nowhere else in the world. Rapid and accurate identification of potential pests and diseases is critical to protect the environment, our communities and industries from serious ecological, health and economic impacts. Invasive species are a serious burden to Australia, costing an estimated $24.5 billion a year.[9,10]

A coordinated national biosecurity information system bringing together data from across government, industry, researchers, and community stakeholders would strengthen and transform Australia's biosecurity, allowing rapid response to threats, and effective management of both emerging risks and established pests and diseases. The system would provide up-to-date, integrated and accessible data from sources such as biological reference collections and databases, data generated through new diagnostic techniques such as eDNA analysis, surveillance data, border security information, geospatial and climate data, and transport networks.

Such a system would be underpinned by improved data access, transparency and interoperability across jurisdictions, sectors (agricultural, environmental, marine, health and research) and industries (tourism, freight and farming)[11], with agreed data standards and sharing protocols that address concerns associated with privacy, commercial sensitivity and trade implications. It would support timely and informed decision-making though accurate compliance information, enhanced monitoring, risk assessment, modelling and improved understanding of the impacts of invasive species on the environment.

4. National, strategic research data needs for science

Five themes emerged from consultations as areas of need to enhance Australia's data infrastructure and support scientific research: national coordination and data integration; data policy and governance; data sharing; data storage, computing and architecture; and data skills and expertise. The consultation workshops were framed around Discovery, Understanding and Prediction which have different requirements around data, software and models. However, the discussions clearly identified cross-cutting strategic data issues. This report focusses on these common needs for data infrastructure.

While many of the data-related needs and challenges presented in this section are shared across science disciplines, national data infrastructure development must consider discipline-specific variations in requirements.

4.1 National coordination and integration

SUMMARY

Realising the full potential of data-intensive research will require greater national coordination and integrated data infrastructure.

Nationally integrated research data repositories for individual science disciplines that can also support multidisciplinary research are a high priority.

Australia's critical databases, data assets and infrastructure for different disciplines should be identified and then mapped to help determine steps to link important datasets together.

Developing a national data strategy with strategic data priorities would be a valuable nucleation point for the coordination and development of a future-proofed national data infrastructure.

Data linkage and integration provide important opportunities to explore complex research questions and develop necessary knowledge to understand and solve societal problems. *Advancing data-intensive research in Australia* found that Australia's research data ecosystem is fragmented, with a complex network of data policies, infrastructure and assets.[1] Valuable science research data are held across a range of entities including government and government agencies, universities and research organisations, NCRIS facilities and private entities, as well as in discipline-specific local and international repositories.

Different institutions have different data management policies and systems, storage capacity and data requirements, presenting challenges for discoverability, interoperability and integration. Additionally, in some instances researchers deposit data in overseas repositories as there are no Australian alternatives, which presents a risk to Australia's national interests if these data become dependent on infrastructure designed and owned by overseas corporate entities. Greater national coordination and integration across Australia's research data infrastructure is required to reduce inefficiencies, preserve data in our national interest and enable multidisciplinary research.

TOWARDS NATIONALLY INTEGRATED REPOSITORIES

Serious consideration should be given to designing and transitioning to a nationally coordinated and integrated research repository ecosystem that enables interoperability and supports multi-disciplinary research. This could be a network of subject-specific repositories supplemented by integrative hubs that support the needs of broader science disciplines. These subject-specific repositories could allow for specialised data quality management, curation and metadata attributes that general repositories cannot cater for.[12] The specifics of the architecture of such repositories and hubs, and the steps required to reach it, needs focussed national-scale consultation and planning.

Such repositories will need to be sustainably resourced to maintain and manage critical datasets for future research. This ecosystem will also need to operate on consistent standards based on international best practice for repositories, like those being developed by the Confederation of Open Access Repositories, to facilitate international collaboration.[13]

As illustrated by the examples in Section 3, such data infrastructure would be highly beneficial to many science disciplines and bring datasets from different projects together to enable future exploration and discovery, and access to data for decision-making. High-level leadership and commitment from relevant federal government agencies is required to drive it.

Identifying and mapping critical datasets and infrastructure for future research would be a first step towards developing this ecosystem, to determine steps to link important existing datasets and repositories together and identify gaps. This will avoid duplicating effort and help to break down silos to create a more strategic, integrated system to serve the strategic needs of different research disciplines. The data inventories pilot program mentioned in the Australian Data Strategy may be a starting point for government-managed data assets.[3] The decadal plans for science developed by the Academy's National Committees for Science can also play an important role by identifying critical data issues, data assets and infrastructure for disciplines required to support Australian research in the longer term.

POWER OF COORDINATION, DATA LINKAGE AND INTEGRATION

Datasets need to represent the diversity of the Australian population to achieve the best research outcomes. There are examples, such as for predictions in human genetics (e.g. disease traits), where the data needed to develop health solutions tailored to individuals is not accessible in Australia and researchers rely on data from repositories overseas. This limits solutions for the Australian context because diversity in the data being used is not representative of diversity in the Australian population. The power of linking data to identify diversity is illustrated by the Multi-Agency Data Integration Project, led by the Australian Bureau of Statistics. This enabled Commonwealth demographic data to be integrated with immunisation records to determine COVID vaccine uptake among different culturally and linguistically diverse populations, then target ongoing communication and action appropriately.[14]

A COLLABORATIVE, COORDINATED AND STRATEGIC DATA ECOSYSTEM

Advancing data-intensive research in Australia highlighted improved access to government (public) data for research as a major priority and the need for coordination between initiatives aimed at improving access to government data (section 3.1.3 and recommendation 3.2 of that report).[1] Consultations for this project reinforced this, with a lack of consistent agreements and coordination across jurisdictions presenting a major barrier to accessing datasets for research and impeding collaboration with industry.

Data sits in Commonwealth, state and industry silos, with government agencies often unable to bring datasets together or integrate data on a larger scale. Also, some private entities are reluctant to release data. This issue appears to be pervasive across many science disciplines, with individual researchers spending much time negotiating access to data. This is highly inefficient and prevents fast, coordinated response to situations requiring timely information, such as biosecurity and disaster response. Inability to access data also leads to large data gaps, resulting in insufficient data for quality studies. It also prevents data linkage and leads to missed opportunities for impact. Inability to share data between state jurisdictions also creates a problem for creating national, federated data platforms.

Consistent data sharing agreements and shared data priorities across government, research and industry would help break down silos and advance progress towards open data. The *Data Availability and Transparency Act 2022*[15] provides the impetus for improving access to public sector data for accredited institutions and users. In light of these reforms, there is an opportunity to develop data sharing arrangements with appropriate privacy and security frameworks to coordinate data access and integration between research, governments and industry.[6]

Early and sustained engagement with relevant industries to ensure that data repositories and platforms are designed with all users' needs in mind would also support coordination for national data assets, creating a meaningful connection between research and industry for research translation. Additionally, communicating the value proposition for national coordination and data integration to practitioners, end-users, and funders (e.g. via practical use cases) is critical to provide motivation and rationale and achieve buy-in from stakeholders.

A culture of competition in Australia's research system also impedes coordination and collaboration. Part of the problem is that funding is allocated in small, short-term amounts via competitive processes. Taking public health and clinical sciences as an example, a large, centralised, multi-institutional structure to share resources and data would enable large scale studies to be implemented quickly.

Developing a national research data strategy with strategic data priorities for research would be a valuable nucleation point for the coordination and development of future-focussed national data infrastructure. The learned academies could provide leadership and a unified voice for research data priorities and bridge discussions in the research sector with current parallel policy conversations around making government data open, reflected in the *Australian Data Strategy*.[3]

4.2 Data policies and governance

SUMMARY

Data infrastructure needs to be grounded in consistent, standardised and enforceable data policies based on the FAIR and CARE principles, as well as legal and ethical frameworks for the responsible collection and use of data.

Areas in urgent need of reform include streamlining ethics and data access processes and supporting long-term data management beyond the life of research projects.

A national research data strategy with research data priorities could help drive coherent national data policy and create a shared vision between government, the NCRIS facilities, research institutions and research discipline communities.

Developing ambitious, future-focused data infrastructure necessitates consistent, standardised data policies that facilitate responsible data use and reuse, data sharing and integration. These policies should be based on the FAIR (Findable, Accessible, Interoperable and Reusable) principles and CARE (Collective benefit, Authority to control, Responsibility and Ethics) principles for Indigenous Data Governance.[16,17] As the ability to collect, link and integrate data advances, the development of legal and ethical frameworks for the responsible collection and use of data is an ongoing challenge for science research. There is a strong need to balance privacy, security and IP with open access to data and software to enable collaboration and research translation.

Advancing data-intensive research in Australia discusses FAIR and CARE principles and open science (section 3.1.1 and 3.1.2) and provides an international perspective on data standards, policies and funding arrangements that support data management and facilitate access and interoperability (section 3.2). There are existing international standards that Australia could adapt for a coordinated national approach to research data management that meets world standards, with the benefit of supporting international sharing, creation of international data assets and collaborative research. The report identifies Finland, Ireland, the UK, Sweden and the European Commission as examples of jurisdictions with open science, research guidelines and funding models which enable coordinated approaches to research data management and data sharing.[1]

Australia should also align with international standards to support collaboration, which include the Core Trust Seal certification and the World Data System (WDS) membership for research data repositories.[1] The Confederation of Open Access Repositories is also currently developing a global Community Framework for Good Practices in Repositories.[13]

The ARDC's Institutional Underpinnings program is a partnership between 25 Australian universities that aims to create a jointly-agreed framework for the management, sharing, retention and disposal of research data.[18] This institutional-level coordination should form part of a broader strategic and coordinated national research data policy framework for research, government and industry. The following issues emerged from consultations relating to national research data policy reform.

REFRAMING NATIONAL DATA POLICY

A national review and reframing of research data policy could establish data as a public good and ensure alignment between research, government and industry, and jurisdictions across Australia and internationally. Such a review should address implementation of the FAIR and CARE principles across the research sector and government agencies holding data, as well as data management and sharing and data infrastructure coordination. This aligns with recommendations 3.1 and 3.2 of *Advancing data-intensive research in Australia* to universally adopt the FAIR and CARE principles and lead a national reform of research data policies, involving the Chief Scientist, the ARDC and learned academies, and other relevant bodies.[1]

The draft *Australian Data Strategy* recognises data as a national asset and outlines the Australian Government's vision for the national data ecosystem.[3] It acknowledges the need for robust data management practices and structured data sharing partnerships between sectors for the purposes of research. It also aims to optimise use and access to public data and improve data integration and consistency of data standards. The *Australian Data Strategy*[3], *Digital Economy Strategy*[19] and the recommended National Digital Research Infrastructure Strategy[6] present opportunities to align data principles between research and government. As the research sector is a major user of government data, alignment of the research sector's needs and government data policy is critical to enhance use of public data for research.

As mentioned previously in Section 4.1, development of a national research data strategy with research data priorities would help drive coherent data policy and ensure that it will meet future research needs. These priorities should be a shared vision between government, the NCRIS facilities, research institutions and research discipline communities.

STREAMLINING ETHICS PROCESSES AND DATA ACCESS

Streamlining ethics processes and data access requirements is an area in urgent need of reform. Lengthy, complex ethics applications to access datasets, which are not suitable for the level of risk, are a significant barrier which delays research or leads researchers to change or abandon projects. Often researchers must apply for ethics approvals by multiple organisations for the same project. These processes need to be updated to reduce administrative burden while mitigating ethics and privacy risks.

4.4 Data sharing

SUMMARY

Data sharing practices enable discovery, drive innovation and support research integrity.

Impediments to data sharing include resource and expertise constraints, differences in data sharing cultures across different disciplines, and data volume and data quality.

Incentives are needed to encourage data sharing in Australia. However, alone they will not be enough. Greater investment in the infrastructure and people required to manage and curate data is also essential.

This section discusses the relevance of data sharing and offers suggestions to promote data sharing practices across the research community, and between government and industry stakeholders. As modern science becomes increasingly data-intensive and collaborative, data sharing practices are important to enable discovery, drive innovation and support research integrity. Advantages of data sharing include linking diverse datasets to answer new research questions, making efficient use of resources by reducing duplication, enabling reuse of data for new research projects, supporting reproducibility and transparency and enabling collaboration to accelerate research.[20,21] The importance of data sharing is clearly illustrated in the case of the SARS-CoV-2 genome sequence, which was freely and openly shared internationally, allowing the rapid development of tests and vaccines. There are also unique advantages to sharing and linking Australian data, for example mitigating data gaps in medical research to improve health outcomes for Aboriginal and Torres Strait Islander peoples.

DATA SHARING CHALLENGES

Data sharing requires researchers to spend valuable time curating and sharing datasets, providing quality metadata and making data interoperable. Researchers may lack the data curation and metadata expertise required to make their datasets useful to others.[22] Researchers may also choose to withhold data for reasons including concern about legal issues and protection of their ability to publish, or possible scrutiny.[20,23,24] A survey conducted by *Springer Nature* found that the main challenges to data sharing among researchers at the publication stage of the research cycle were 'organising data in a presentable way', 'unsure about copyright and licensing', 'not knowing which repository to use', 'lack of time to deposit data' and 'cost of sharing data'.[25] *The State of Open Data 2021* survey found that misuse of data, not receiving appropriate credit or acknowledgement and being unsure about copyright and data licensing were respondents' main concerns about sharing data.[26]Even when published research includes a data availability statement, there is evidence that in practice researchers do not always share the data.[27]

Data sharing cultures vary across different disciplines, influenced by the state of disciplines in recognising the value of aggregating data into repositories, their level of capacity to work with large datasets and the nature of the data the discipline works with (e.g. privacy concerns with sharing and concerns that shared data may be misused).[21] Organisational practices and data sharing culture within organisations also impacts data sharing.[28,29,30] Different stakeholders' risk perceptions are also influential.

Consultations highlighted opportunities for engagement between research, industry and government stakeholders on data infrastructure planning and alignment of policies and approaches to enable effective data sharing and data applications.[31] The Australian Agrifood Data Exchange (OzAg Data Exchange) was highlighted as an example initiative that is exploring frameworks for research and agriculture industry stakeholders share data across the agrifood sector.[32]

Some datasets collected and held by private entities (e.g. indoor air quality data, data from agriculture) have potentially high value if shared and analysed alongside other data. Incentives are needed to encourage sharing, with options including offering industry access to data from research and government. National datasets that are managed by government agencies are often only made available as a data product, rather than in minimally processed, non-proprietary and machine-readable formats that have greater utility for research.[33]

Data volume and data quality are significant challenges for data sharing. Researchers can share large amounts of data, but not necessarily in forms that are usable to others, limiting use and interpretation. Verifying data quality and providing sufficient metadata to trace provenance and make data transparent and reusable are ongoing challenges in science and have implications for accountability of policy decision-making based on data. National standards based on international best practices will help to ensure the veracity of data and metadata contributed to repositories.

A lack of knowledge and established practice among researchers regarding data licencing, IP rights, and privacy is another barrier to data sharing. Consistent data and IP policies within and between institutions, training for researchers, and provision of data expertise can address this issue.

INCENTIVISING DATA SHARING

Incentives are needed to encourage data sharing in Australia, with the aim to lead a cultural shift to make data sharing and open data, code and software practices the norm. Incentives include data sharing requirements in grant agreements and publications, recognising data and software sharing in academic success metrics for hiring and promotion and subsidising institutions data storage costs if they make data available.

Research organisations should recognise open science practices such as data sharing and open source software in criteria for academic promotion. Recognising and rewarding data sharing in terms of academic success metrics and career advancement is critical. *The State of Open Data 2021* found that 65% of respondents had never received credit or acknowledgement for sharing data, yet the main motivations respondents indicated for sharing their data included citation of research papers, co-authorship on papers, increased visibility of their research and public benefit.[26] Data citation can be used to attribute data and to count data as a research output. Additionally, linking data to publications and other research outputs facilitates sharing and enables further investigation.

Research funders have an important role to play in incentivising data sharing. *Advancing data-intensive research in Australia* notes that the National Health and Medical Research Council (NHMRC) and Australian Research Council (ARC) policies are currently based on recommendations rather than regulation and compliance.[1] The report highlights the UK Research and Innovation's Common Principles on Data Policy, which has greater compliance monitoring, and the US National Science Foundation's Dissemination and Sharing of Research Results policy which require preparation of data management plans and reporting on the dissemination and sharing of research results, putting research data 'on equal footing to the publications that arise from a research project.'[1] The report suggest that the ARC and NHMRC adopt similar practices.[1]

Advancing data-intensive research in Australia also notes that policies to make data open and FAIR are inconsistent across Australia's research institutions.[1] Consultations for this project found that the responsibility for sharing data is largely placed on individual researchers. While incentives are an important element to encouraging data sharing, greater investment in the infrastructure and people to manage data and make data FAIR, readable by both humans and machines, is needed to enable data sharing and avoid placing an unfair burden on individual researchers.

4.5 Data storage, computing and architecture needs

SUMMARY

Data needs to be delivered in a way that is discoverable, usable, and flexible to cater for the range of stakeholders who will use it.

Investment is required to build Australia's data capacity to move and manipulate large amounts of data, including a strategy that leads to exascale data capacity.

Long-term funding to support the continuity of data storage and management is critical to preserve datasets that support long-term monitoring or may be useful in the future.

New technologies are offering opportunities to collect a wealth of data. For example, next-generation sequencing and digitising analogue works are transforming biological and natural history collections. Data are constantly changing, so databases and infrastructures need to be future-ready and adaptable to new forms and sources of data. Additionally, data needs to be maintained and updated so that it remains usable as hardware and software evolve.

DISCOVERABILITY, USABILITY AND FLEXIBILITY

The way data are delivered, discoverability and usability needs to cater for the range of stakeholders who will use it. Currently, there are useful data sources available, but the ability to exploit them is low as they are not available in a platform or interface that is easy for typical stakeholders to use, or data are available across multiple different interfaces rather than a unified, single source. Data architectures and systems need to serve end-user requirements, such as being able to de-identify data or export it in the volumes or formats required, to enable full utilisation of data. Research commercialisation and end-user needs should be considered when establishing systems and critical data repositories to ensure that industry and other stakeholders can interact with and exploit data to solve problems and make decisions.

Systems with integrated computing resources also need to be flexible to avoid unintended consequences of limiting people to a certain set of tools as data and methods evolve. Balancing standards and flexibility will enable ease of use without compromising ability to innovate.

LARGE VOLUMES OF DATA

Moving and manipulating large volumes of data is a consideration, particularly for computationally intensive disciplines such as astronomy, climate, Earth observation and computational biology. Distributed computing, cloud computing and 'moving code to data' are options for data integration and analysis of large datasets and multivariate data for some fields. Investment that sets us on a path to exascale data capacity in Australia's national research infrastructure is important to support computationally intensive disciplines into the future such as bioinformatics, climate, solid earth and astronomy.[31]

CONTINUITY OF DATA MANAGEMENT, STORAGE AND DATA RETENTION

Data storage, retention and management beyond the life of a research project is another urgent challenge. For example, institutional requirements to destroy health research data after five years (an arbitrary number lacking methodological basis) limits research possibilities. Management of data from long-term monitoring projects and support for large multi-institutional big data projects that do not fit within regular grant funding cycles were also highlighted as challenges.

Long-term funding, continuity for data storage and appropriate data management plans are needed to preserve datasets that are important for long-term monitoring or may be useful to answer future research questions. Additionally, data kept for long periods of time require periodic upgrading due to evolution of supporting software, hardware, standards and vocabularies. For disciplines such as geoscience, where instruments in the field are consistently collecting large amounts of data, addressing issues of storage capacity, location in relation to computational infrastructure, and making data available in minimally processed, high resolution forms as FAIR data is vital to maintain and use important datasets long term.

As highlighted in recommendation 3.4 of *Advancing data-intensive research in Australia*, conversations within disciplines about what data should be retained, the appropriate length of retention of data, ongoing use, who data are available to, and long-term, sustainable data storage, curation and management are needed. On this, *Advancing data-intensive research in Australia* highlights work in the UK.[1] A 2019 study by a UK-based organisation investigated what research data should be kept, identifying two primary drivers for keeping research data: research integrity and reproducibility; and the potential for reuse.[1,33] Also in the UK, the Digital Curation Centre provides a five-step checklist for assessing research data.[34] The checklist encourages researchers to consider whether the data have potential reuse purposes, whether the data must be kept under laws or policies, whether the data have long-term value, and what costs may be incurred by retaining the data.[34]

Consideration of what data should be stored should be part of research data management plans and embedded in the research workflow. Specific consideration should be given to the costs associated with the collection, retention, and preservation of data. Decisions around what data should be kept long term need to be informed by the strategic needs of disciplines for future research and require careful consideration to balance cost and benefits of storing and maintaining data.

4.6 Data skills and expertise

SUMMARY

There is an urgent need to build data skills in Australia's research sector. Data skills are needed through a specialised data science workforce available for research, and also more generally across all researchers.

Appropriate funding, attractive employment conditions and recognition for data science expertise are all critical for attracting and retaining people with data skills in the research sector.

People are an essential part of data infrastructure. Data infrastructure development and efforts to integrate data need to be complemented by adequate training and access to expertise to generate, curate, analyse and manage data for research. Both access to experts and raising the digital skills of researchers were emphasised in consultations.

As highlighted in chapter 6 of *Advancing data-intensive research in Australia*, data-intensive research requires data scientists, research software engineers and other technical experts, alongside raising the data literacy of researchers.[1] Developing a digitally-skilled research workforce is urgently needed to enable data, models and software to meet the FAIR principles[5,35,36] and avoid compromising the pace and scale of research.[1]

The NRI Roadmap acknowledges that 'People and expertise are an intrinsic and essential part of the NRI.' (p. 60). As noted in the NRI Roadmap, the NRI Workforce Strategy and National Digital Research Infrastructure Strategy should consider investment and a long-term plan to develop the specialised workforce to support researchers to use, reuse and manage data, to develop software and to utilise digital research infrastructure. This section covers issues highlighted in consultations that need to be addressed to build data skills and expertise.

Data-intensive research presents challenges with generating, analysing, storing, processing, sharing and managing data. These challenges require data curators and managers, as well as a cohort of experts who can navigate workflows, optimise code, develop software and enable efficient computation. This expertise is particularly important at the interface between discipline experts and research infrastructure such as large-scale computing centres.

Full-time data professionals cannot be funded from individual research grants to curate and manage data, making it challenging to attract and retain expertise and build capacity within research organisations. Lack of attractive employment conditions and recognition for data science expertise in academia is also likely to impact data-intensive research, with data scientists being drawn to industry.[1] Opportunities to retain skilled people in research include developing metrics to recognise the contributions of data experts who collaborate on research, and providing stable positions and clear career paths for advancement.[37]
An example is the Melbourne Data Analytics Platform (MDAP), which provides a centralised source of data support for researchers.[38] The platform's staff are 'academic specialists', a career path where data analytics and software development metrics are recognised as research outputs for promotion.

Raising the data skills of researchers is also urgently needed to build the capacity of discipline experts to create, store, use and manage data throughout a project lifecycle. A baseline level of data literacy would be valuable to equip researchers with knowledge of how to handle large datasets and work with data specialists. Training needs vary in terms of previous experience and the skills needed by different disciplines.[1] With the growing necessity of data skills, and variation in the level of data, statistical and computational skills required between research disciplines, individual science disciplines need to address the key competencies required (e.g. by the end of a PhD). *Advancing data-intensive research in Australia* recommended that research organisations review their staff development programs and provide courses for students to develop data skills.[1]

While some initiatives and grassroots programs to raise the data skills of researchers exist, *Advancing data-intensive research in Australia* did not identify instances of national coordination for programs specifically aimed at upskilling researchers in Australia.[1] Strategic planning for a digitally skilled workforce and coordinated investment in training programs will support data capability in the research sector. The Australian Government has provided some investment to raise Australia's data-related capability, outlined in the *Australian Data Strategy*. For example, CSIRO's Next Generation Graduates Programs offer scholarships to train graduates in AI and data science as part of the Australian Government's Digital Economy Strategy and Artificial Intelligence Action Plan.[39]

Early- and mid-career researchers were identified as a potential target cohort for initiatives to build data science skills in the research workforce while addressing job insecurity.[40] As an example, the German Centre for Integrative Biodiversity Research (iDiv) every competitively funded data synthesis project is supported by a dedicated postdoctoral position.

Creating human linkages and working in teams with complementary expertise is becoming more common and enables access to data skills to support research and is critical to enable multidisciplinary approaches. This requires a shared understanding between those collecting the data and others who will be analysing, using and interpreting the data to ensure that the right data are collected and that ways of working are complementary and efficient.

5. Opportunities for Academy leadership, advocacy and planning on research data issues

Through its role as a provider of independent, authoritative and influential science advice and engagement with research communities nationally and internationally, the Australian Academy of Science is positioned to provide leadership on research data policy issues and support strategic planning for national research data infrastructure.

The Academy can support strategic planning for data infrastructure, policy, skills and capabilities in the following ways:

- Provide input into the 2022 NRI Investment Plan, NRI Digital Research Infrastructure Strategy and NRI Workforce Strategy, guided by the findings of this report, *Advancing Data-Intensive Research in Australia* and National Committees for Science.
- Seek funding to identify and map critical national datasets and infrastructure to guide data integration and coordination.
- Address emerging data-related challenges and opportunities through science discipline decadal planning to reflect research directions and developments that have been endorsed by research communities. Decadal planning should identify key data issues and critical strategic datasets and infrastructure for the discipline.
- Drive cultural change to ensure uptake of FAIR and CARE principles and research data policies through discipline-specific consultation, to provide advice on data retention and curation based on the nature of the research area.
- Advocate for investment in data expertise to curate and manage data and optimise researchers' use of national research infrastructure in the 2022 NRI Investment Plan, NRI Digital Research Infrastructure Strategy and NRI Workforce Strategy.
- Convene expertise and act as an independent, authoritative voice on national strategic data priorities for science and data policy reform.
- Provide science policy advice on scientific data management to the Academy, the Australian Government and Australian organisations with support from the National Committee for Data in Science.
- Engage with governments and other stakeholders to support strategic planning and action to enhance Australia's data infrastructure.
- Support research integrity by promoting responsible data practices and advocating for data policies and processes that strengthen governance and address issues of transparency, reproducibility and replicability.
- Maintain strong engagement and alignment with international data policy dialogue and agendas through the National Committee for Data in Science and engagement with the WDS and the International Science Council's Committee on Data (CODATA).

6. Actions to advance data infrastructure, policies and skills to support science research

The following statements recommend actions to grow research data capability and advance the development of data infrastructure in Australia. Where applicable, recommendations from this report are aligned with recommendations from *Advancing Data Intensive Research in Australia.*

RECOMMENDED ACTIONS FOR RESEARCHERS, RESEARCH ORGANISATIONS AND HIGHER EDUCATION PROVIDERS

Adopt the FAIR and CARE principles and open access policies for research data and include compliance and monitoring measures in data policies. This must be accompanied by providing the resources required (e.g. data infrastructure or data professionals) to support researchers to publish data, code and software and facilitate good sharing practices.

Recognise and incentivise good open science practices for researchers through metrics for academic promotion.

Build data capability through staff development programs, courses for postgraduate research students and undergraduate training offerings to develop data skills.

Acknowledge data generators, curators and stewards in published research outputs, including as authors.

Establish clear academic career pathways for data scientists including data professionals and research software engineers.

Assess, update and streamline ethics and application processes for collecting and accessing data to reduce administrative burden and ensure consistency between research organisations and government.

RECOMMENDED ACTIONS FOR GOVERNMENTS

Lead development of national strategic data priorities for research to facilitate national coordination and guide investment in data infrastructure, skills and capabilities in the national interest. This priority setting should involve research discipline communities, research organisations, government agencies, industry and national research infrastructure facilities.

Recognise the costs of managing data and complying with the FAIR and CARE principles in funding policies for universities and research organisations, research and infrastructure grants and funding councils.

Establish coordinated data access and sharing agreements between state and federal jurisdictions, research and industry to facilitate data movement out of institutional silos and integration of valuable datasets for research and decision making. Government data should be made available for research in minimally processed, non-proprietary and machine-readable formats.

RECOMMENDATIONS FOR THE NATIONAL DIGITAL RESEARCH INFRASTRUCTURE STRATEGY

A National Digital Research Infrastructure Strategy was recommended in the 2021 NRI Roadmap. The Academy recommends that the strategy:

Drives consistent adoption of the FAIR and CARE principles and open access policies across national research infrastructure and presents a plan for Australia's research infrastructure to align with international standards.

Works with researchers and end-users to identify nationally significant data collections and addresses their long-term sustainability to maintain valuable datasets in the national interest.

Drives a strategic, coordinated effort towards a national, integrated data repository ecosystem for science and other disciplines.

Addresses investment in the data expertise required to support national research infrastructure. This expertise is essential to manage, curate and maintain data and assist domain experts with their research through data support such as code optimisation, software development, computation, data curation and management.

Presents a strategy to build data skills among researchers to enable data-intensive research which could include scholarships for early- and mid- career researchers to upskill.

Drives streamlined access to government data for research in minimally processed, non-proprietary and machine-readable formats.

7. Process and consultation

The development of this report was guided by a Steering Committee comprising Fellows of the Australian Academy of Science and members of the Academy's National Committee for Data in Science. The Steering Committee was Chaired by Professor Jane Elith FAA.

Steering Committee

CHAIR

Professor Jane Elith FAA, Honorary Professor, School of Ecosystem and Forest Sciences, The University of Melbourne

COMMITTEE

Professor Ian Chubb AC FAA FTSE, Emeritus Professor, The Australian National University; Secretary Science Policy, Australian Academy of Science

Professor Andy Pitman AO FAA, Director, ARC Centre of Excellence for Climate Extremes, UNSW Sydney

Professor Ginny Barbour, Co-Lead, Office for Scholarly Communication, Queensland University of Technology; Director, Open Access Australasia

Dr Danny Kingsley, Associate Librarian, Content and Digital Strategy Library, Flinders University

Dr Lesley Wyborn, Honorary Professor, National Computational Infrastructure and Research School of Earth Sciences, The Australian National University

Anna Maria Arabia, Chief Executive, Australian Academy of Science

Chris Anderson, Director Science Policy, Australian Academy of Science

Project approach and framework

The approach for this environmental scan was developed in collaboration with the ARDC, ACOLA and the other learned academies.

The Australian Academy of Science developed three scenario narratives on the themes of discovery, understanding and prediction in research. The narratives explained the scope of these three themes and provided illustrative examples of science research scenarios. The three narratives are provided in Appendix 1. These were intended to stimulate reflection and discussion of a range of current and future data challenges that impact science research across different disciplines. The discovery, understanding and prediction framework was used to reflect different scales, complexity, data requirements, aims and stakeholders of different research activities and encourage cross-disciplinary discussion.

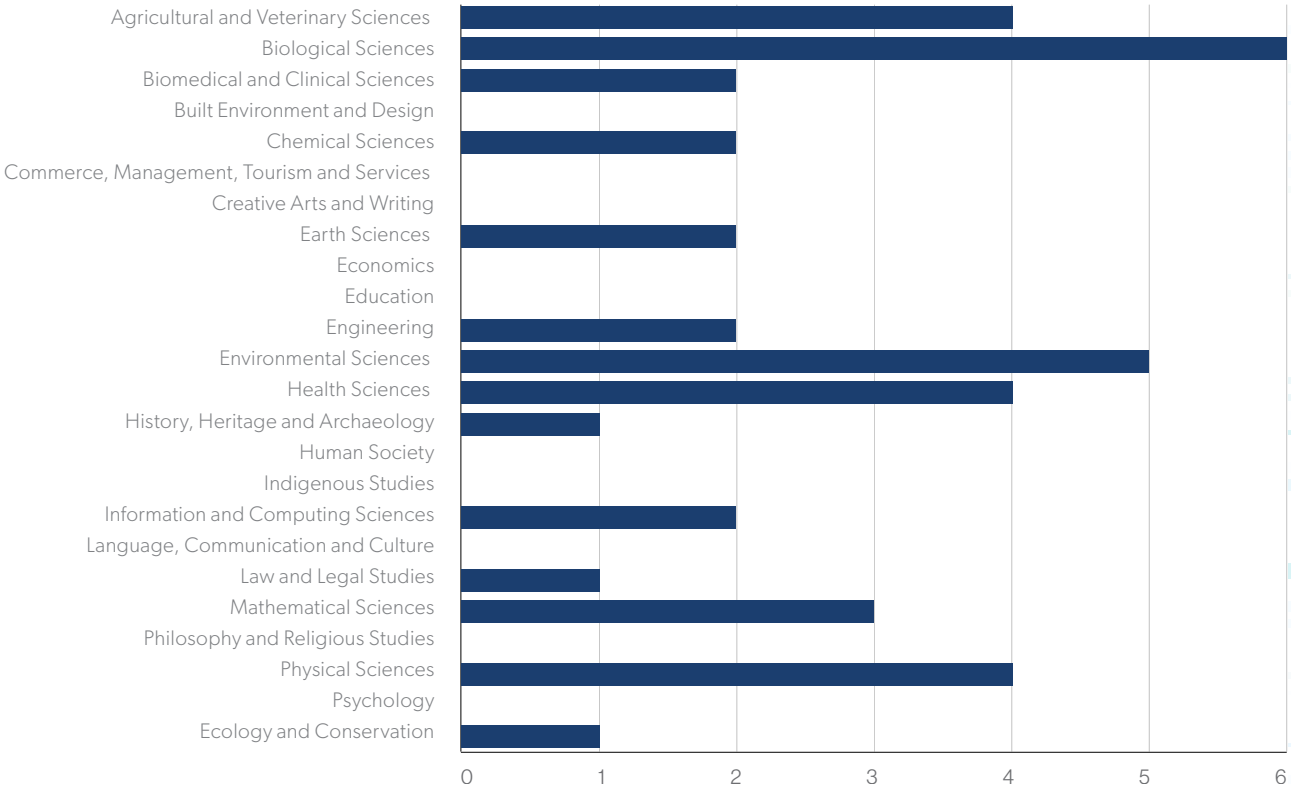
In the scenario narratives, workshop discussion participants were encouraged to think broadly about data issues in their field and what could be accomplished in the future with next-generation research infrastructure and skilled people and data policy to support research.

Consultation workshops

The approach to the workshop discussions was adapted from the Sutherland methods for 'collaboratively identifying research priorities and emerging issues in science and policy'.[41] The extensive community feedback and use of breakout rooms in workshops were not used in this project, but the general structure involving pre-event surveys and workshop discussions was retained.

Experts were identified by the Australian Academy of Science secretariat and project Steering Committee, and recommended by Academy Fellows, based on their expertise and engagement with research data issues impacting their field. Participants were selected to bring together three multidisciplinary groups representing a range of different disciplines in the natural and physical sciences. An indicative breakdown of the disciplines represented in the workshop sessions, by Field of Research, are shown in Figure 1. Three workshop sessions were held (one each on discovery, understanding and prediction), each with 10-14 attendees, with 32 participants in total across the sessions.

FIGURE 1 Fields of research of the participants who responded to the post-event survey following the workshop session.



People who were invited to participate completed a registration form to confirm their involvement, where they were provided with the scenario narratives and asked to indicate which of the areas (discovery, understanding or prediction) best aligned with their expertise and interest. Participants were allocated to a workshop session based on these preferences.

A pre-workshop survey was distributed to participants to gather open responses about gaps in Australia's research data ecosystem and opportunities for development to enable scientific research. Participants were asked to consider the research narrative and the development of their discipline over the next decade (to 2030) when responding to the following questions:

1. Please indicate your field of research based on the Fields of Research (FoR) Divisions below. You may select more than one field, or use the 'other' text entry box. (Participants were asked to choose from a list)
2. If nothing were to change, where will the gaps, weaknesses or risks in Australia's research data ecosystem be?
3. What data-related developments would you like to see to advance research and research impact in Australia?
4. How does the data capacity of your research discipline or community need to develop to address emerging challenges or research themes?
5. What will the research data ecosystem (e.g. research infrastructure, skills, data policies, data capabilities) need to look like to be able to do the research that you or your field is beginning to contemplate?

Responses to the survey were summarised into key ideas and re-distributed to the participants in a second survey. The second survey asked participants to select and rank their top five responses from the list in order of priority. The results of the surveys were distributed to participants in advance of the workshops. The five highest ranked responses for each question were highlighted as a starting point for discussion in the workshop, but open discussion was not limited to these topics.

The workshops were held virtually via Zoom and participants could contribute to the discussion verbally and via the comment tool and live polls using the online polling tool Slido. The workshop sessions were broken into the following three topics which aligned with the pre-workshop survey questions:

- Current and emerging research data needs to enable research in Australia
- Gaps, weaknesses and risks in Australia's research data ecosystem
- Data-related developments and opportunities to enhance research and research impact in Australia.

Participants were asked three live poll questions during the workshops:

- What are the data-related barriers to your current research or projects you might pursue in the future?
- What opportunities are arising from data-enabled research in your field?
- What are examples of international data trends or initiatives that Australia could adopt? Or Australian exemplar initiatives?

A summary of key points raised by participants in the workshop sessions is provided in Appendix 2. Appendix 3 provides the barriers and opportunities identified by participants in response to the live poll during the workshop.

A post-event survey was distributed to participants, to capture their feedback on the topics discussed. The post-event survey included the following questions:

1. Please indicate your field of research based on the Fields of Research (FoR) Divisions below. You may select more than one field, or use the 'other' text entry box. (Participants were asked to choose from a list)
2. Which points/ideas did you particularly agree with?
3. Which points/ideas did you particularly disagree with?
4. Of the ideas discussed in the workshop, which do you think is most critical?
5. Was there any topic that wasn't discussed today, that should have been?
6. Do you have any feedback for the organisers?

The responses to the surveys, live polls and the outcomes of the workshop discussions was collated and synthesised into this report.

Workshop chairs

Professor Martina Stenzel FAA, UNSW Sydney

Professor Kerrie Mengersen FAA FASSA, Queensland University of Technology

Professor Jane Elith FAA, The University of Melbourne

Workshop participants who contributed to consultations

The following people contributed through workshops, interviews and document reviews.

Professor Adrian Barnett, Queensland University of Technology

Dr Jonathan Beesley, QIMR Berghofer Medical Research Institute

Dr Phill Cassey, The University of Adelaide

Simon Costello, Australian Climate Services

Distinguished Professor Noel Cressie FAA, University of Wollongong

Professor Drew Evans, The University of South Australia

Associate Professor Daniel Falster, UNSW Sydney

Dr Rebecca Farrington, The University of Melbourne and AuScope

Professor Nick Golding

Mr Donald Hobern, Australian Plant Phenomics Facility and Atlas of Living Australia

Professor Rob Hyndman FAA FASSA, Monash University

Professor Michael Kidd AM FAHMS, Australian National University

Associate Professor Sarah Kummerfeld, Garvan Institute of Medical Research

Dr Bryan Lessard, Australian Biological Resources Study, Department of Agriculture, Water and the Environment

Professor John Mattick AO FAA FTSE FAHMS, UNSW Sydney

Professor Jodie McVernon FAHMS, Melbourne Medial School and The Peter Doherty Institute for Infection & Immunity

Dr Vanessa Moss, CSIRO

Dr Tim Rawling, AuScope

Professor Andrew Robinson, The University of Melbourne

Dr Susie Robinson, Australian Plant Phenomics Facility, University of Adelaide

Professor Julie Anne Simpson, The University of Melbourne

Dr Manodeep Sinha, Swinburne University of Technology, ARC Centre of Excellence for All-Sky Astrophysics in 3 Dimensions (ASTRO 3D)

Dr Mohammad Taha, The University of Melbourne

John Curtin Distinguished Professor Steven Tingay, Curtin University

Professor Ian Wright FAA, Western Sydney University

Dr Andre Zerger, Atlas of Living Australia

8. Appendix 1 Scenario narratives

Discovery in research

The following statement is intended to encourage exploration of data issues and data-related needs to meet current and emerging challenges in the context of 'discovery' in research. This includes the data infrastructure, tools and services, skills, data policy, data management plans and data capabilities required to support research and research impact in Australia within and across research disciplines.

'Discovery' is concerned with targeted projects where data may be collected or collated then analysed to describe a particular phenomenon. Data may also be explored to uncover new research questions and form new hypotheses. Research projects may involve working with quantitative or qualitative data. While prediction may be involved in discovery, the focus here is on research activities or challenges aimed at discovery.

The below examples illustrate situations and data-related tasks to consider in the context of 'discovery'.

1. A researcher aims to collect historical ocean observation data and blend it with rainfall data over Australia and the Pacific Islands, to identify patterns in climate drivers that influence drought in Australia. They plan to access observations from diverse and non-standardised sources including NOAA, the WMO, the Bureau of Meteorology, IMOS, NCI and Digital Earth Australia. They are also considering using text and data mining of news sources from Trove. They intend to publish data, code and analyses so the full analysis is reproducible.
2. A researcher plans to gather and analyse available data on national and international marine pest and invasive species to identify biological characteristics that determine invasiveness of pests in Australian marine environments. The data will include databased records plus expert knowledge from a broad range of peoples including Indigenous groups, professional trawlers, and recreational divers. The data collected will be a combination of quantitative and qualitative. They intend to publish the data and analyses.
3. A researcher is involved in a national consortium studying Australia's soil biodiversity. They aim to collect soil samples and sequence the eDNA present to produce DNA sequence information of organisms present at the site. They will compare the sequence information from samples against available reference genomes. Their data will form part of a project to create a reference map of Australian soil biodiversity. They aim to make their data publicly available in a university or national repository and need to integrate with international reference standards.
4. A number of cases of a similar, unknown disease have been reported at a regional hospital. The disease is suspected to be caused by a new virus. Samples from the patients are collected and sequenced to establish the virus genome. The genome sequence is shared via an open online discussion forum for virus evolution. There is an immediate necessity to obtain as much genetic information from as many virus strains as possible. International teams of researchers compare the sequence to known viruses to identify close relatives and characterise the new virus. Researchers share new data rapidly via online repositories and pre-print publications, enabling rapid development of tests and vaccines.

Understanding in research

The following statement is intended to encourage exploration of data issues and data-related needs to meet current and emerging challenges in the context of 'understanding' in research. This includes the data infrastructure, tools and services, skills, data policy, data management plans and data capabilities required to support research and research impact in Australia within and across research disciplines.

Here we use 'understanding' to refer to research and analyses that deepens understanding of or explains systems or processes. The examples below focus on research targeting understanding but excluding prediction. The focus is on larger scales, greater complexity and broader data analyses than may be considered for research aimed at 'discovery'. Research projects may involve working with quantitative or qualitative data.

When reflecting on 'understanding' in research as it applies to your field, include in your thinking new, ambitious research undertaken at a scale that is world leading and the requirements to enable it.

The following examples illustrate situations and data-related tasks to consider.

1. A researcher plans to conduct a project to enhance understanding of air quality drivers in Australian cities, including all cities of more than 100,000 residents. They will be integrating time series of observed meteorological data, meteorological reanalyses and local air quality data across the cities and supplement this with simulations with an air pollution model to understand the causes of hazardous air pollution.
2. A suspected exotic invasive plant pathogen is detected in an inbound shipment of nursery plants. The following research is required: (a) the species needs to be identified; (b) a research program needs to be planned and initiated to understand the life cycle of the pathogen in Australia, and to identify likely hosts; (c) a system for collating all existing and new data on the species needs to be made operational, so information can be shared across states and agencies while maintaining the security of sensitive biosecurity information.
3. Researchers plan to conduct fieldwork to understand the impact of severe bushfires on a threatened native species. This will require collection of a range of data including from farmland, National Park and Indigenous protected areas. Some of these sites will be very remote. Data need to be quickly databased and analysed to inform recovery from the fires.
4. Advances in next generation sequencing techniques are enabling analysis of the function of genetic variations associated with disease. Connecting variants to their functions in particular genes, cell types and pathways that lead to disease can guide therapeutic discovery and support the goal of precision medicine. A research group aims to understand the mechanism of potential pathogenic genetic variants shared among individuals with a particular disease. To identify variants of interest, they access annotated sequence data stored in a large public database that archives data on relationships between human genetic variations and phenotypes. They plan to use a cellular model system and use high-throughput functional assays to analyse the functional impact of the variants. They plan to share their data in an international data repository as required by the journal they publish their study in.
5. Managing the COVID-19 pandemic requires researchers to understand how the virus is changing as new variants emerge, and understand how these changes affect the spread of the virus and severity of the symptoms. Samples are taken from many thousands of individuals and require rapid genetic sequencing. Sequences need to be made readily accessible to all interested parties, and analysed using robust, reproducible and open source methods. International collaboration drives rapid turnaround of this information.

Prediction in research

The following statement is intended to encourage exploration of data issues and data-related needs to meet current and emerging challenges in the context of prediction in research. This includes the data infrastructure, tools and services, skills, data policy, data management plans and data capabilities required to support research and research impact in Australia within and across research disciplines.

Predictions are made from various types of models. Predictions may be made once or continually updated, they might be spatial or temporal, they may have one or millions of stakeholders and users. Predictions might be discarded, or kept and managed so they are insulated against changes in technologies or software.

The following examples illustrate situations and data-related tasks to consider.

1. The incidence and impacts of extreme weather events such as heatwaves, extreme rainfall and destructive storms will increase with global warming. Different regions across Australia will experience different challenges e.g. severe cyclones in northern Australia, increased heavy rainfall and flooding in Central Australia, and increases in droughts in southern Australia. Using climate models to project the changes in the risk of these events, or using operational prediction systems to predict specific occurrences of events, requires access to multiple streams of data including satellite and in situ data, in combination with 4-dimensional simulations from models. Analysing results from simulations require management and analysis of petascale data, simultaneously with local-scale information, to provide the detailed local level probabilistic climate projection and information on risks associated with extreme weather events and compound climate extremes required by local decision makers and other users. Increasing granularity is required to identify where climate risks will most severely affect Australians to inform longer-term strategies, risk assessments and adaptation measures.
2. An exotic invasive plant pathogen is found in Australia for the first time, suspected to have entered with a shipment of nursery plants. Overseas it has been known to affect species with many close relatives in Australia. Biosecurity professionals will need to draw on a range of biodiversity, supply chain and transport data to assess risk, predict potential dispersal over the next 5-10 years, identify species the pathogen could infect and determine areas for surveillance in the environment.
3. A research team aims to develop a genetic risk prediction model for a complex disease using a machine learning method. The team plans to use single nucleotide polymorphism (SNP) data from genome-wide association studies, gathered from international databases, as well as de-identified genomic data and clinical information from individuals, including data from Indigenous patients, to develop the model.
4. Multi-institutional teams run epidemiological models to predict the spread of the COVID-19 virus under different public health measures, informing decisions about how to respond to outbreaks. Predictions of virus spread are updated weekly. Researchers use a range of data which may include effective reproduction rate, demographic data, transport links, healthcare system capacity, size of social networks and people's activities. Models may be required at different levels of resolution, and need updating as new information emerges. Data are sourced across state, federal and international jurisdictions. Researchers share information on the virus and its spread across nations, to enable fast and accurate learning.

9. Appendix 2 Workshop discussions summary

The tables below provide a summary of key ideas raised in the workshop discussions and surveys, arranged by themes.

THEME: COORDINATION AND INTEGRATION		
Workshop	**Challenges, gaps, and needs**	**Opportunities and solutions**
Discovery	• Greater leadership required from funding agencies to support data management and sharing • High-level leadership from relevant agencies to create national data assets • Integration is critical to allow greater exploration and discovery • In Australia there are less drivers to aggregate data to advance research • Missed opportunities and lost added value of bringing datasets together to see the big picture • No incentive for 'data poor' disciplines to consider the value in aggregating data into central repositories	• Discipline national data centres to bring datasets together and provide centralised data support • Funding agencies adopt a framework for using and storing data • Leadership to set national strategic data priorities • A dedicated national central service for hosting research data in a systematic way could help a lot of fields
Understanding	• Lack of agreements between jurisdictions makes data sharing difficult, leading to missed opportunities for impact and impeding collaboration (repeated in 'Data sharing') • State-based data are not brought together at a larger scale or integrated, and cannot be linked to other datasets • Risk in duplicating effort and wasting resources • Disciplines have varying ability to speak with a single voice to advocate for data needs • Ongoing challenge to integrate commercial data, social and biophysical data, data on different spatial and temporal scales • Culture of competition rather than collaboration • Focus seems to be on universities making their data available for industry, rather than viewing data as something that society interacts with as a whole, this will limit large-scale, integrative, transdisciplinary research data	• Implement coordinated, standardised data sharing agreements between jurisdictions (repeated in 'Data sharing') • Map existing databases and infrastructure that exist nationally and determine steps to link them together • Communication brokerage between smaller discipline interest groups to transfer knowledge and ideas to strategy and investment conversations • National coordination for a standardised approach to find, access and retrieve data with metadata, definition and dictionaries • Large, centralised, multi-institutional structure to implement big studies quickly (rather than competition for small grants) • Grand challenges can be an effective way to work together • Engage early and regularly with industry areas that are interested in data to ensure there is a meaningful translation/connection between publicly funded research and industry and feed this into designing and integrating data into meaningful assets

THEME: COORDINATION AND INTEGRATION		
Prediction	• Human linkages and shared understanding of how data that is collected is going to be used and the purpose is important and will be critical to enable multidisciplinary approaches • Challenge to bring datasets together from different projects (commercial and publicly funded) from disparate places • Important to consider non-traditional data streams to think innovatively and make the case for greater linkage, augmentation and integration • Large datasets become difficult to manipulate at an institutional level • Difficult to have a one size fits all model • Difficult to create central discipline specific repositories as the IP sits with different institutions, organisations etc.	• Create use cases to illustrate power of access and integration of data to provide motivation and rationale • A national and state commitment to harmonisation • Federated structure (locationally devolved infrastructure that are linked)

THEME: COORDINATION AND INTEGRATION		
Workshop	**Challenges, gaps, and needs**	**Opportunities and solutions**
Discovery	• Lack of knowledge and ability to share data in a form that is useful for others • Sharing is constrained by only using local repositories • No mandate for government and agencies to store and contribute data • Need to consider how industry and other stakeholders can interact with and contribute to data being shared • Challenge to share data from different fields in a useful way	• Sharing incentives including recognition of data and software as research outputs

THEME: COORDINATION AND INTEGRATION		
Understanding	• There are valuable datasets that are unavailable as they are only collected privately • Incentives are required to encourage data sharing • Data volumes and sharing data in a form that people in another field could use, risk of errors in use and interpretation • Need to make data accessible to diverse people (e.g. people with a disability) to interact with and use • Systems and procedures not built to de-identify data or export data in the volume or format required • Need a large, centralised, multi-institutional structure for sharing (rather than institutes fighting for small grants) in public health and clinical science • Data for publication is not always suitable for sharing • Fields with greatest freedom to share data (e.g. biodiversity) have the least money to do it • Better resourced disciplines (e.g. medicine and agriculture) have a lot to gain from sharing data but there are more commercial interests in keeping data exclusive/profitable • Models and data are not necessarily easily discoverable for use in other contexts • Lack of understanding of legal agreements, IP, and privacy are barriers to sharing data (repeated in 'Policies and governance') • Researchers need to be confident in their rights when sharing data (repeated in 'Policies and governance') • Lack of agreements between jurisdictions makes data sharing difficult, leading to missed opportunities for impact and impeding collaboration (repeated in 'Coordination and integration')	• Engage with industry to ensure translation connection and design and integrate data into meaningful assets • Implement coordinated, standardised data sharing agreements between jurisdictions (repeated in 'Coordination and integration') • There could be a tightly coupled view with a plan for the infrastructure and architecture or we could map what exists and how they are used to understand who has the data and who to contact to access it
Prediction	• Cultural shift required in some disciplines to encourage data sharing • Government organisations make data available as a data product, but this limits analyses that can be done • Need to consider end-user needs and the way data is delivered, its discoverability and usability to cater for the range of stakeholders who will use it – there are a lot of great data sources available but the capacity to exploit them is low • Sharing constraints between state and federal governments create challenges • National data platforms for sharing data (e.g. in health) haven't been sustained in the past, so people can be risk averse in contributing their data to these solutions	• Recognise data and software as a research output and reward making data open and sharing open-source software • Incentivise data sharing e.g. through grant and publication requirements • Learn lessons from why previous national data sharing platforms have not worked

THEME: COORDINATION AND INTEGRATION		
Workshop	**Challenges, gaps, and needs**	**Opportunities and solutions**
Discovery	• Advantage to making data available in one location and moving code to data • Movement of large-scale data is a challenge • Resources are required to allow people to remotely work on data	
Understanding	• Databases and infrastructure need to be future ready and adaptable to new forms and sources of data	
Prediction	• Enable moving compute to data, or data storage co-located with compute in some cases • Systems need to be flexible and future-proofed • Standardised systems with integrated compute resources may inadvertently stifle innovation • Disciplines which collect large amounts of data need to address where to store it and data management plans are required to preserve data • Cultural shift required to store data and make data available in raw forms as FAIR data • In some cases, pre-processing is required to be able to store data, in other disciplines it is critical to store raw data • Matters less where the data repository is located, what is important is that it is well managed and discoverable • Data need to be machine discoverable • Downloadable data not suitable for all situations, in these situations computing is done in the cloud • Need to be aware of download costs associated with data • Reliance on commercial services for data storage and sharing is a risk as things could be removed at any point • There are different needs for different types of data, some can probably go in a general data archive while for others there is potential value in a bespoke platform that captures metadata in a convenient way and facilitates meta-analyses	• Make decisions about storage and compute at the same time so you can compute where data is stored

THEME: COORDINATION AND INTEGRATION		
Workshop	**Challenges, gaps, and needs**	**Opportunities and solutions**
Discovery	• Need to understand the veracity of data	• Standards to ensure data quality in repositories
Understanding	• Data in databases is unverified and methods are not reproducible due to lack of transparency • Sufficient metadata is required to trace provenance back to original measurements/observations in a transparent and reusable way • A standardised approach for how to find, access and retrieve data with metadata, definitions, dictionaries is required	
Prediction	• There are different needs for different types of data, some can probably go in a general data archive while for others there is potential value in a bespoke platform that captures metadata in a convenient way and facilitates meta-analyses	• Data is federated but with central metadata

THEME: COORDINATION AND INTEGRATION		
Workshop	**Challenges, gaps, and needs**	**Opportunities and solutions**
Discovery	• Legal and ethical frameworks for responsible collection and use of data are not yet sufficiently developed	
Understanding	• Data isn't viewed as a public good • Lack of understanding of legal agreements, IP, and privacy are barriers to sharing data (repeated in 'Data sharing') • Researchers need to be confident in their rights when sharing data (repeated in 'Data sharing') • Different stakeholders have difference perceptions of risk in relation to data • Areas differ regarding data confidentiality which creates a barrier and constraints by state legislation	• National review and data policy reset across universities and the research sector
Prediction	• Length of time and administrative burden of processes to obtain access to health data is a significant barrier • Need to balance privacy concerns with making data available and viewing it as a public good	• Focus on meaningful ethics processes, e.g. community engagement rather than 'box-ticking' • There are opportunities to change policies to support the ongoing use of data (e.g. the 5-year destruction of data requirement in health research limits what is possible)

THEME: COORDINATION AND INTEGRATION		
Workshop	**Challenges and gaps**	**Opportunities and solutions**
Discovery	• Small projects do not have the resources to manage data and make data interoperable • Inadequate funding and people limit use of existing infrastructure to their full capacity • Funding required for dedicated people to curate valuable historical datasets • Lack of resources to compile, preserve, maintain, and digitise data • No funding available for critical long-term monitoring beyond the life of a grant or project • There is a requirement that infrastructure, management, and support is commensurate with any national agreements	• Fund postdoc level positions to curate datasets and bridge gaps between datasets and entities
Understanding	• Critical foundational collections and datasets are underfunded • Culture of competition rather than collaboration • Funding is not allocated in a way that encourages movement out of discipline silos • No capacity to refresh data, making data collected and stored obsolete • Lots of resources are needed to curate databases, make sure they are reliable and link them to other national databases • Unable to fund a full-time data professional under one grant	
Prediction	• Some valuable big data projects don't fit within normal funding cycles as they require significant investment and cannot be done as a single project • Human resources are required to bring datasets together • Cannot hire data managers from grant funding • Need an open view regarding location of data and compute (in some cases it may be cheaper to use commercial cloud services than institutional storage and compute) • Long-term funding and continuity for data storage is critical • Lack of resources for ongoing management and curation	

THEME: COORDINATION AND INTEGRATION		
Workshop	**Challenges, gaps, and needs**	**Opportunities and solutions**
Discovery	• Loss of data scientists and data experts from academia • Baseline data literacy required to work with data specialists • Need to train people who are not data scientists to use databases and contribute their own data • Need a sustainable, central source of data expertise with broad applicability to support research	• Recognition and rewards (e.g. funding, promotion) for data curation, analysis and software development • Data literacy included in undergraduate curriculum • Potential funding for data activities could be directed at addressing EMCR job insecurity and capacity building
Understanding	• A cohort of experts is required to support large-scale, data-intensive questions • Need clear career paths for data experts in research • Risk losing skilled people due to fixed term positions, with some leaving for opportunities overseas • Investment in training and skills to be able to utilise data for large scale studies • Need cohort of experts (e.g. software engineers, code optimisation, data curation and management experts) who are at the interface between discipline experts and compute facilities • Knowledge sharing and learning to implement sophisticated workflows • Risk of underusing existing infrastructure, in part due to lack of training • Need to complement development of infrastructure and efforts to integrate systems with training • Need investment in the training and skills and right people to be able to look at and use the data is the bottleneck with genomic data, big scale studies are happening but need to people to be able to look at it locally	• Training and skills development to supplement existing infrastructure, and sharing example workflows using data infrastructure • Need to build teams with complementary expertise • We have a pool of expertise who are utilised at a project or program level to build databases or access information, but this pool of expertise can be tapped to tackle problems as a whole research data access problem
Prediction	• Urgent need to retain data experts in research/academia • Need pathways to bring skilled people into Australia who don't fit into a clear box • Need to raise statistics and data management skills among researchers to improve understanding when taking on research projects and collecting data • Disciplines working with big data need a level of computational literacy to be able to examine their data • Competencies required in different disciplines (e.g. by the end of a PhD) need to be examined • Need people to understand each other and have complementary ways of working as there is a trend towards working in large teams with different expertise	• Recognise and reward data professionals in research • Create clear career pathways and career progression

10. Appendix 3 Barriers and opportunities identified in the workshop live polls

WORKSHOP	DATA-RELATED BARRIERS TO RESEARCH	OPPORTUNITIES ARISING FROM DATA-ENABLED RESEARCH
Discovery	• Responses to this question were not collected in the discovery workshop session.	• New datasets that don't fit the mould require new statistical methods • Intersecting remote-sensed data with other data types • Combining data from different scales to ask new types of research questions • Broadening the collaborations • Global partnerships and access to international data • Larger available data sets • Improved data • Stored data for generations after us • Better utilisation of data for more than one purpose – e.g. legacy data to answer questions not originally thought of when the data was collected in the first place • No unnecessary repetition of experiments • Comparisons • Initiation of discussions • New ideas • Identifying new research directions where data is lacking • Continental scale biosecurity risk assessment/modelling • Understanding multi-scale processes (basic science) • Research translation • More meaningful interaction and crossover with industry partners, especially in data science • Faster identification of materials to translate into products • The transformation of medical research and healthcare

WORKSHOP	DATA-RELATED BARRIERS TO RESEARCH	OPPORTUNITIES ARISING FROM DATA-ENABLED RESEARCH
Understanding	• Lack of funding (2 votes) • Metadata (2 votes) • Dirty-data • Training • Integration • Transdisciplinarity (2 votes) • Interoperability • Database experts • Data security • Harmonisation • IPD Meta-analyses • Loss of staff knowledge • Lack of staff resources • Siloed databases • Federated data • Supercomputing • Reusable models • Data/capability hording • Repeatability • Transparency • Retrievability	• Forming integrated views on crop performance across our continent in light of a changing climate • Better water regulation and allocation based on an innovative but data intensive data-model fusion approach • Ability and opportunities to tackle systems interoperability problems. • To build a very large cohort of data scientists that will diffuse into other disciplines and into industry • Reliable extrapolation on data and proposals • Better quality data leading to robust and reproducible findings • Understanding at larger spatial scales • True cross-disciplinary integrated research • Automation of data capture and integration into data infrastructure and delivery • Linking primary and secondary care data • Linking of spatial and temporal elements • Positioning plant development data in the context of whole-landscape crop and environment data, time-series and trends • Integration of social and biophysical factors • Global scale comparative studies, through space and time • Predictive capacity
Prediction	• Incentives, lack of incentives • Poor data hygiene • Quality of metadata • Ownership • Federated system • Ethical approvals • Discoverability • Persuading govt to fund (in relation to big data projects that don't fit within normal grant funding cycles) • No systematic organisation • Cost of storage • No long term storage • Data repository funding • Integration of datasets	• We are producing a large cohort of data scientists • Healthy 'scepticism' about existing systems • Prediction based on uncertainty and sparse data • Much better ability to predict future pandemic waves • Better genomic/microbiome-based predictions supporting sustainability and profitability of crop and livestock systems • Science with archival data (astronomy) • Automated writing of research papers • Automated checking of research papers • Subpopulation identifiability • Managing shipping container biosecurity risk in a more fine-grained risk-aware way • Combined inversion

12. References

1 Australian Academy of Science, 2021. Advancing data-intensive research in Australia. science.org.au/files/userfiles/support/documents/advancing-data-intensive-research-in-australia-11-11-21.pdf

2 Australian Government, 2022. Australia's Digital Economy. Australian Government digitaleconomy.pmc.gov.au/ [Accessed June 5, 2022].

3 Department of the Prime Minister and Cabinet, 2022. Australian Data Strategy: The Australian Government's whole-of-economy vision for data. ausdatastrategy.pmc.gov.au/ [Accessed June 1, 2022].

4 Australia's Chief Scientist, 2021. Unlocking the academic library: Open Access. Australia's Chief Scientist chiefscientist.gov.au/news-and-media/unlocking-academic-library-open-access [Accessed June 8, 2022]

5 UNESCO. UNESCO Recommendation on Open Science, 2021. en.unesco.org/science-sustainable-future/open-science/recommendation [Accessed June 9, 2022].

6 Department of Education Skills and Employment, 2021. 2021 National Research Infrastructure Roadmap. dese.gov.au/national-research-infrastructure/resources/2021-national-research-infrastructure-roadmap.

7 Department of Industry Science Energy and Resources, 2022. 2022 Critical Minerals Strategy. industry.gov.au/data-and-publications/2022-critical-minerals-strategy

8 Australian Academy of Science, 2018. Decadal plan for Australian Geoscience: Our Planet, Australia's Future. science.org.au/supporting-science/science-policy-and-sector-analysis/decadal-plans-science/australian-geoscience

9 Bradshaw, C. J. A. et al, 2021. Detailed assessment of the reported economic costs of invasive species in Australia. NeoBiota 67, 511–550.

10 Pest plants and animals cost Australia around $25 billion a year – and it will get worse. theconversation.com/pest-plants-and-animals-cost-australia-around-25-billion-a-year-and-it-will-get-worse-164969 [Accessed May 1, 2022].

11 CSIRO. Australia's Biosecurity Future, 2020. csiro.au/en/work-with-us/services/consultancy-strategic-advice-services/csiro-futures/futures-reports/agriculture-and-food/biosecurity-futures

12 Hahnel, M., 2022. Guest Post: A Decade of Open Data in Research — Real Change or Slow Moving Compliance? The Scholarly Kitchen scholarlykitchen.sspnet.org/2022/03/30/guest-post-a-decade-of-open-data-in-research-real-change-or-slow-moving-compliance/?informz=1 [Accessed May 29, 2022]

13 Confederation of Open Access Repositories, 2022. Open Consultation – COAR Community Framework for Best Practices in Repositories. coar-repositories.org/news-updates/open-consultation-coar-community-framework-for-best-practices-in-repositories/ [Accessed June 1, 2022].

14 Australian Bureau of Statistics. Multi-Agency Data Integration Project (MADIP). abs.gov.au/about/data-services/data-integration/integrated-data/multi-agency-data-integration-project-madip [Accessed June 1, 2022].

15 Data Availability and Transparency Act 2022. Commonwealth of Australia, 2011.

16 GO FAIR. FAIR Principles. go-fair.org/fair-principles/ [Accessed June 8, 2022].

17 Global Indigenous Data Alliance, 2019. CARE Principles of Indigenous Data Governance. gida-global.org/care [Accessed June 8, 2022].

18 Australian Research Data Commons, 2022. ARDC Institutional Underpinnings Framework. zenodo.org/record/6392341 doi:10.5281/ZENODO.6392341.

19 Department of Education and Training, 2014. The Australian research data infrastructure strategy. apo.org.au/node/42792

20 Tenopir, C. et al, 2011. Data Sharing by Scientists: Practices and Perceptions. PLOS ONE 6, e21101.

21 Tenopir, C. et al, 2015. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. PLOS ONE 10, e0134826.

22 Popkin, G., 2019. Data sharing and how it can benefit your scientific career. Nature 569, 445–447.

23 Campbell, E. & Bendavid, E., 2003. Data-sharing and data-withholding in genetics and the life sciences: Results of a national survey of technology transfer officers. J of Health Care Law Policy 6.

24 Savage, C. & Vickers, A., 2009. Empirical study of data sharing by authors publishing in PLoS journals. PLoS ONE 4.

25 Stuart, D. et al., 2018. Whitepaper: Practical challenges for researchers in data sharing. doi:10.6084/M9.FIGSHARE.5975011.V1.

26 Digital Science et al, 2021. The State of Open Data 2021. Digital Science doi:10.6084/M9.FIGSHARE.17061347.V1.

27 Gabelica, M., Bojčić, R. & Puljak, L. Many researchers were not compliant with their published data sharing statement: mixed-methods study. Journal of Clinical Epidemiology doi:10.1016/J.JCLINEPI.2022.05.019.

28 Tenopir, C. et al., 2020. Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. PLOS ONE 15, e0229003.

29 Sayogo, D. S. & Pardo, T. A., 2013. Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. Government Information Quarterly 30, S19–S31.

30 Mason, C. M., Box, P. J. & Burns, S. M., 2020. Research data sharing in the Australian national science agency: Understanding the relative importance of organisational, disciplinary and domain-specific influences. PLoS ONE 15.

31 National Committee for Data in Science, 2021. National Committee for Data in Science Submission to the Department of Education, Skills and Employment - 2021 National Research Infrastructure Roadmap Exposure Draft. science.org.au/files/userfiles/support/submissions/2021/aas-response---draft-nri-roadmap.pdf

32 Fisheries Research and Development Corporation, 2020. Australian Agrifood Data Exchange (OzAg Data Exchange): Deliver an interconnected data highway for Australia's AgriFood value chain. frdc.com.au/project/2020-126 [Accessed June 1, 2022].

33 Beagrie, N., 2019. What to keep: a Jisc research data study. repository.jisc.ac.uk/7262/

34 Digital Curation Centre, 2014. Five steps to decide what data to keep: a checklist for appraising research data v.1. dcc.ac.uk/guidance/how-guides/five-steps-decide-what-data-keep [Accessed June 1, 2022].

35 Bello, M. & Galindo-Rueda, F. Charting the digital transformation of science. (2020) doi:10.1787/1b06c47c-en.

36 Buchhorn, M., 2019. Surveying the scale of the research-IT support workforce - a survey and report comissioned by the Australian Research Data Commons (ARDC).

37 OECD Global Science Forum, 2020. Building digital workforce capacity and skills for data-intensive science | en | OECD.

38 Melbourne Data Analytics Platform (MDAP). mdap.unimelb.edu.au/ [Accessed June 1, 2022].

39 CSIRO. Next Generation Graduates Programs, 2022. csiro.au/en/work-with-us/funding-programs/programs/next-generation-graduates-programs [Accessed June 1, 2022].

40 Christian, K., Johnstone, C., Larkins, J. A., Wright, W. & Doran, M. R., 2021. A survey of early-career researchers in Australia. Elife 10, 1–19.

41 Sutherland, W. J., Fleishman, E., Mascia, M. B., Pretty, J. & Rudd, M. A., 2011. Methods for collaboratively identifying research priorities and emerging issues in science and policy. Methods in Ecology and Evolution 2, 238–247.