Submission to the Productivity Commission Inquiry into Data Availability and Use

Made by the National Computational Infrastructure

July 2016

National Computational Infrastructure
NCI Building, 143 Ward Road
The Australian National University
Acton ACT 2601

www.nci.org.au

OVERVIEW

The National Computational Infrastructure welcomes the opportunity to contribute to the Productivity Commission's inquiry into Data Availability and Use. NCI strongly supports initiatives to make the data gathered by, and held in, government and publicly-funded entities more accessible and more useable. Doing this requires investment in data management and services by custodians. However, if done properly, improving access to such data and making it more useable has the potential to deliver a significant innovation dividend for Australia – particularly through improved outcomes in the research and government sectors.

About NCI

As Australia's national research computing service, NCI is funded by the Australian Government to support advanced computational and data-intensive methods by providing high-performance infrastructure services to Australia's researchers, and in so doing raise the ambition, impact, and outcomes of Australian research.

NCI operates as a formal collaboration of the ANU, CSIRO, the Bureau of Meteorology (BoM), and Geoscience Australia (GA), strengthened by formal partnerships with science agencies, medical research institutes, and a number of research-intensive universities supported by the Australian Research Council. The infrastructure investments underpinning NCI have been provided by the Australian Government under its National Collaborative Research Infrastructure Strategy (NCRIS).

NCI has established a widely used, highly regarded High Performance Data (HPD) Node, which provides robust, high quality data infrastructure services, including managing and serving data collections of high-value. This is done through open standard protocols, underpinned by a trusted repository and overlaid with tools for analytics which are supported by necessary compute capacity. NCI's capability is well demonstrated by the development and deployment of the National Environmental Research Data Interoperability Platform (NERDIP), which makes over ten petabytes of major data collections available both in situ and as data services. The platform brings together the data and metadata, overlaying these with a comprehensive range of modern information services. More information on the NERDIP is in the example box, below; more detail on NCI's HPD capability is at **Attachment A**.

NCI is now recognised for its considerable experience managing and serving high-value public data sets in a way that maximises the benefit they provide to the broader research sector. Within the international scientific community, there is considerable interest in NCI's model. It is on the basis of this expertise and recognition that NCI is making this contribution to the inquiry. An overview of NCI's input is immediately below, with more detailed responses to some of the questions in the Issues Paper in the following sections.

Summary of Input

Significant benefits and innovation accrue from making data available in way which will enables re-use, including for research and through data analytics. To achieve these benefits, consideration must be given not only to which data sets to make available and the management of those data sets, but also to the way the data is made available. The latter includes how the data is linked, referenced and put in a form that can be readily analysed.

Wherever possible, data needs to be managed and provided in accordance with well-established community and international standards, noting that data management is an ongoing activity.

- Standards provide a reference point for application software and services to be developed with confidence.

- Implementation of methods and processes to ensure standards compliance increasingly need to be automated, and standardisation requires ongoing resources as relevant technologies continue to evolve.

Issues such as which data to make available in what order (including criteria for this), which standards to use and which tools and services to bring together with the data, are usually best addressed at the level of the community, or domain – above the level of the organisation, but sub-national.

Further, it is important to be aware that because big data collections are more difficult to copy and to synchronise, even a basic level of data availability now needs to include the following.

- The provision of machine-readable network data services.
- The provision of related information in a linked way, where such data needs to be trusted.
  - This includes data provenance and lineage, as well as data product information.
- Such access to data is provided without entering into service-level agreement (SLA) requirements.
  - More robust arrangements may be required by some parties, depending on the level of and type of demand from the community of interest.

Accordingly, high-value public sector data is most actively used and cited by—and therefore delivers the greatest benefit from—users, including from the researcher government sectors, when it is available widely through open standard protocols, underpinned by a trusted repository and overlaid with tools for analytics supported by necessary compute capacity – that is, a platform. Delivery of public sector data through a platform, such as NCI's NERDIP, significantly improves research outcomes by ensuring that the broader community of potential users are able to:

- Analyse definitive data sources;
- Combine or compare sources for analysis; and
- Create innovative applications and processes of benefit to otherwise unsupported sectors of the community.

The extent of the benefit derived by users such as researchers is thus partly a function of the extent of the investment in making the data accessible and useable. Making data available and then maintaining that availability requires the application of resources, including expertise as well as either or both of ongoing capital and operational costs, depending on the approach.

An approach which has achieved great efficiencies in making agencies' data accessible and useable is the approach enabled through NCI, under the auspices of the NCRIS program. The NCRIS funding arrangements acted to facilitate deep collaboration between government agencies, universities and other stakeholders. By adding resources to a government investment, a capability has been delivered which exceeds what could be provided by any one or two partners. This capability is an aggregation of multiple infrastructures and skills sets, providing robust, high quality data infrastructure services which meet the needs of multiple parties, to whom it is responsive as funders. This model, in which direct support combines with agency cash co-investment to underpin a facility specifically geared to facilitating broad uptake and access, has the greatest potential to drive necessary innovation in the development and operation of data services, and is likely to be most effective for agencies.

We suggest that attempting to collate a list of datasets which various parties suggest may be of value if made accessible, will not be the most effective approach to identifying datasets which deliver the most value to the research, business and other sectors. Rather, the datasets with the highest value for a group or groups of users are those which represent a combination of the following:

a) Are aligned with defined priorities; and

b) Have certain technical characteristics.

With respect to a), the experience of NCI in establishing an HPD Node is instructive. The strategic goals of the partnership underpinning the HPD infrastructure and services led to the development of priorities for datasets to be potentially accessioned. The key elements of these prioritisations can be summarised as:

- Are directly associated with, or required access to the hard and soft infrastructure operated by the partnership to make more effective use of them.
- Are of ongoing national/international significance, and which were selected on this basis by the supporting stakeholders;
- Are requested by research communities to be served by the facility.
- Will complement and augment collections in the categories above when combined with them, delivering benefit to cognate research communities;
- Combining datasets held by different research communities into coherent collections; and
- Will benefit from being managed within the rich environment of high-end computational and data-intensive services operated by the facility.

With respect to point b), above, while there is no simple set of characteristics that define high-value datasets, in NCI's experience there are broad technical features of high-value datasets which can be categorised as follows.

- They are able to be used in a variety of ways; and/or
- They are widely used information and services, forming part of processes, or chains of processes that deliver outcomes of high value – such outcomes including new datasets, services, etc.; and/or
- They are an input to research, or other activities which are high-impact; and/or
- Their provenance can be traced.

More specific technical characteristics of high-value datasets, and data services include the following:

- Enabled for programmatic, machine readable access, both in-situ and over the internet;
- Available via capable data services that allow subsetting, dynamic aggregation and other forms of data manipulation;
- Complies with scientific (or alternative relevant) data standards, and provides provenance information to expose the way in which the data was generated, curated, manipulated and disseminated;
- Version controlled;
- Both the data and the data services used are citable through Persistent Identifiers;
- The data uses open, flexible, high performance data formats; and
- Used by data services and that can be readily accessed by a broad set of tools that can be developed and extended by communities for their purpose.

Skilled labour in data science and data management remains very difficult to recruit and retain, and is a critical issue for entities delivering data management services in, or for, publicly-funded entities. The fact of these difficulties significantly increases the value of facilities which have built up an expert data science and data management capability, and it will be important not to spread the pool of skills too thinly, by supporting fewer, larger entities which build on existing capability.

RESPONSES TO QUESTIONS ON HIGH VALUE PUBLIC SECTOR DATA

Q. What public sector datasets should be considered high-value data to the: business sector; research sector; academics; or the broader community?

NCI has considerable experience managing and serving high-value public datasets to maximise the benefit they provide to the broader research sector. Based on this experience, we suggest that attempting to collate a list of datasets which various parties suggest may be of value, will not be the most effective approach to identifying datasets which deliver the most value to the research, business and other sectors.

We suggest that the datasets with the highest value for a group or groups of users are those which represent a combination of the following:

c) Are aligned with defined priorities; and
d) Have certain technical characteristics.

With respect to point a) and the matter of employing defined priorities to guide decisions on value, the experience of NCI in establishing a high-performance data (HPD) node for the research community under the auspices of the NCRIS program, is instructive (*see* **Attachment A** *for details of the development of the HPD Node*).

In establishing the node it was necessary to prioritise datasets for accessioning, which were of highest value to NCI's partners and their collaborators in Australia and internationally. There were two elements to the development of priorities. At a high level, the partnership's strategic goals necessitated prioritisation of data collections that:

- Required access to NCI's unique hard and soft infrastructure, to make more effective use of both existing and emerging technologies and techniques – in this case, high-performance computational (HPC) or data-intensive analysis through HPD services;
- Were associated directly with the NCI infrastructure classes above, and which should be co-located with them;
- Are of ongoing national/international significance, and which were selected, in the national interest, by the supporting stakeholders of NCI; and
- Are requested by research communities to be served by NCI.

More particular priorities were also identified. These were then used to prioritise datasets by establishing which would have a higher potential value to users when incorporated into NCI's HPD environment. In NCI's case, the priorities outlined are:

- Making accessible major data collections currently held by NCI's partner national science agencies - Australian Bureau of Meteorology, CSIRO, Geoscience Australia;
- Complementing and augmenting these collections with other nationally and internationally significant collections that are priorities for cognate research communities;
- Combining datasets held by different research communities into coherent collections; and
- Integrating the collections from both the research and government sector, and then establishing these collections in a rich environment of high-end computational and data-intensive services.

Implicit in these characterisations was the need to:

- Capitalise on the rich computational environment at NCI; and
- Engender an outcome of mutual value to NCI stakeholders, the national research community and the national innovation agenda that would make available national collections in such an environment

Based on this prioritisation, NCI has accessioned datasets which have made it a critical hub for research in a range of disciplines, for users from well beyond the circle of NCI's major partners. The significance of the examples lies in that fact that it is necessary to reconcile the value placed on datasets by their custodians, with the value placed on datasets by those who seek to re-use them. Establishing priorities for dataset availability within broad domains or communities of use, which encompass both data custodians and those who will re-use the data, represents the most effective approach to establishing which datasets have the highest value when made widely accessible and useable.

With respect to point b), above, NCI input on the characteristics of high-value datasets is provided below, at the question 'What characteristics define high-value datasets?'.

Q. What characteristics define high-value datasets?

There is no simple set of characteristics that define high-value datasets. NCI has dealt with this issue by identifying a dataset's value based on its alignment with defined priorities, and *technical* characteristics. Datasets held at NCI have been prioritised on the basis that they constitute reference datasets which are national in scope, and have been organised through trusted organisations and groups (see above). Of equal importance has been establishing the technical features of high-value datasets. In NCI's experience, these are:

- They are able to be used in a variety of ways; and/or
- They are widely used information and services, forming part of processes, or chains of processes that deliver outcomes of high value – such outcomes including new datasets, services, etc.; and/or
- They are an input to research, or other activities which are high-impact; and/or
- Their provenance can be traced.

In the era of Big Data, datasets can no longer be considered just for the value of the information that they contain. They must also be considered in the context of the way the data is made available, which includes how the data is linked, referenced and put in a form that can be readily analysed. These concepts necessitate consideration of fundamental technological advances that have taken place around data services and informatics. For example, datasets are more highly valued if they can be put into a form that is discoverable over the web, can be processed without the need for a user to manually control the process, and can be reliable to be used for third-party capabilities such as data analytics and machine learning functions.

As a consequence of NCI's experience and considering the issues outline, we suggest that the technical characteristics of high-value datasets, and data services includes the following:

- Enabled for programmatic, machine readable access including in-situ and over the internet;
- Data is available via capable data services that allow subsetting, dynamic aggregation and other forms of data manipulation;
- Complies with scientific (or alternative relevant) data standards, including known references for encoding the data, trusted processes by which the data was generated and that provides provenance information to expose the way in which the data was generated;
- Version controlled;
- Both the data and the data services used are citable through Persistent Identifiers;
- The data uses open, flexible, high performance data formats; and
- Used by data services and that can be readily accessed by a broad set of tools that can be developed and extended by communities for their purpose.

Working Example: Building on the Technical Characteristics of Data to Increase its Use, and Benefits

The development of the NCI National Environmental Research Data Interoperability Platform

The correlation between the technical characteristics of data and the value it has when made available is well demonstrated by the platform NCI has created for making datasets available both in situ and as data services. NCI's National Environmental Research Data Interoperability Platform (NERDIP[1]) provides users with a flexible and extensive digital environment for accessing all the major data collections at NCI. The platform has carefully brought together the data and metadata, and overlaid these with a comprehensive range of modern information services about the data – which is needed for modern machine accessible, programmatic methods. The platform provides both standards-based protocols and state-of-the-art technology to provide a full range of ways for research and innovation to take place on the data collections.

This approach has already delivered significant value by ensuring that the research community has a full range of software and methods, accessible through commonly accessible interfaces, to analyse the major national data assets – including conducting analyses across different big datasets. In the era of Big Data, such a platform is now an essential national infrastructure for working interdisciplinary way. The capacity of the platform to deliver this value, however, arises out of the initial investment in ensuring the data being used has the necessary technical characteristics.

It should also be noted that ensuring the NERDIP platform (or any similar platform) can continue to improve research outcomes is an ongoing task. For NCI this means continuing to manage quality assurance of the data, and improves the data organisation and data format to make best use of all the components of the system.

A further significant feature of this approach to building an interoperable data services platform over datasets with the necessary technical characteristics, is that additional community environments can then be built to interact with the platform. In the case of the NERDIP, examples of data services that have built directly on the NCI data management layers, and include NCI collaboration or partnerships (both nationally and internationally), are:

- NCI’s Raijin, VDI and research cloud environments for building powerful and flexible environments next to the data;
- The Earth Systems Grid Federation (ESGF) data service federation which builds on large international data synchronisation and provides portal access to climate and other Earth system data;
- EarthServer - a European Commission funded Horizon2020 project that provides another exemplar international data service federation that provides access to satellite data using both novel array services and open data standards;
- Copernicus Data Hub for Sentinel data replicated to NCI using underlying high performance data transfer methods;
- NICTA/Data61 National Map which accesses the NCI data services;
- CSIRO’s eReefs portal - which adds web access and brokerage to NCI’s hosted model and observation data for the Great Barrier Reef;
- NCRIS NeCTAR Virtual labs including Climate and Weather Science lab (CWSLab), AuScope Virtual Geophysics Laboratory (VGL), Astronomy Australia Limited (AAL) All-sky Virtual observatory (ASVO);

[1] http://nci.org.au/systems-services/national-facility/nerdip/

- The Australian Geoscience Data Cube - a collaboration between Geoscience Australia, CSIRO and NCI to provide new tools to analyse time series data managed at NCI;
- Research Data Australia (RDA) discovery portal developed by the NCRIS Australian National Data Service (ANDS);
- The find.gov.au / data.gov.au portals[2];
- The NCRIS Terrestrial Ecosystems Research Network (TERN) - providing access and additional land products within NCI's MODIS satellite collection
- The NCRIS Integrated Marine Observing System (IMOS) Satellite Ocean data products that are hosted within the NCI's MODIS satellite collection
- NCRIS AuScope Geophysics research data products and portal access via both AusGIN and AuScope portal.

Q. What benefits would the community derive from increasing the availability and use of public sector data?

Delivery of public sector data through a platform such as NERDIP is an investment, but one which significantly improves research outcomes by ensuring that the broader community of potential users are able to:

- Analyse definitive data sources;
- Combine or compare sources for analysis; and
- Create innovative applications and processes of benefit to otherwise unsupported sectors of the community.

High-value public sector data is most actively used and cited by—and therefore delivers the greatest benefit from—the research community (both academic and government science agency), when it is available widely through open standard protocols, underpinned by a trusted repository and overlaid with tools for analytics supported by necessary compute capacity – that is, a platform.

It is important to understand that that the extent of the benefit derived by users such as researchers is thus partly a function of the extent of the investment in making the data accessible and useable. Because Big Data collections are more difficult to copy and to synchronise, even a basic level of data availability now needs to include:

- The provision of machine-readable network data services.
- The provision of related information in a linked way, where such data needs to be trusted.
  - This includes data provenance and lineage, as well as data product information.
- Such access to data is provided without entering into service-level agreement (SLA) requirements.
  - More robust arrangements may be required by some parties, depending on the level of and type of demand from the community of interest.

Therefore, significant benefits and innovation through research accrue from making data available in way which will enables 'Big Data' analytics. However, these benefits are dependent on the application of appropriate resourcing and expertise to the process of making the data accessible, if it is to deliver such outcomes.

[2] http://find.ga.gov.au/ / http://data.gov.au/

RESPONSES TO QUESTIONS ON COLLECTION AND RELEASE OF PUBLIC SECTOR DATA

Q. What are the main factors currently stopping government agencies from making their data available?

From NCI's dealings with partners in government who have sought to improve the benefit they and others can gain from their data through making it more available, it is clear there are a number of key considerations for both initially making data available, and *sustaining that availability* into the future.

These are:

- The level of direct benefit to the government organisation in terms of support for its work, and the work it undertakes with others, from investing in moving to a more openly accessible environment;
- Access to IT infrastructure that is designed to make the datasets appropriately available, noting that this will include costs for ongoing storage upgrades, network infrastructure *and* soft infrastructure, such data service development and ongoing operations;
- Cost of technical and management skills to curate, manage a quality service that is useful for external user communities;
- Staff resources to move data from older infrastructure to modern available systems;
- Ongoing support and maintenance costs;
- Prioritising and valuing the data for third-party use;
- Security concerns; and
- Reputational damage associated with misinterpretation of data, and/or use of data beyond its known limitations.

Q. How could governments use their own data collections more efficiently and effectively?

If datasets are to deliver value when made widely accessible across government, they must be able to be used in ways other than that for which they were originally collected and organised. However, many government data collections have not been exposed to modern data management (in particular for Big Data and international collaborations), cutting-edge infrastructure, improved standards, or properly curated in way that will allow the data in the collections to be used beyond the initial purpose for which it was collected.

It is still common for government datasets to require heavy management by human processes to locate them, evaluate the reason for accessing them, select and then transfer the data. Once received by another party, the data typically is stored again, and reformatted to suit individual software tools – with this process often occurring multiple times over. Further, if data is managed by a narrowly purposed community, then the data availability is narrowed (captured) by the needs of that particular community. And lastly, even if such data is provided online, the data is commonly provided either through human interaction via portals which are designed for a specific set of use-cases, or an old-style "data download" and end-user data wrangling.

To efficiently enable usage by multiple other parties, including within government, data must be made available in a machine-readable way over the internet, as well as be capable of being used in-situ by those other parties. That is, other parties can access a dataset or datasets remotely and analyse them in the place they are stored. Data made available in this way can service multiple user communities simultaneously while also lowering the cost of data management and storage, and allow for far more effective data, software and workflow sharing environments through aggregated resources and expertise.

The NCI case is a demonstration of this. By providing a shared platform bridging the government science agencies and research communities, and in an international context, NCI has also fostered knowledge sharing and innovation, and enabled transformative research to take place. Since NCI has commenced providing the capacity to both manage and analyse the data, it has been possible to more effectively understand and improve the quality of data for the analysis that is required, rather than just to make the data available for download to be used elsewhere on other infrastructures.

The concentration such a large volume of data on a shared platform has also allowed for the building of critical mass of skills to work more effectively on deeper data management issues around a broader set of co-located, cross-disciplinary datasets.

Q. Should the collection, sharing and release of public sector data be standardised? What would the benefits and costs of standardising? What would standards that are 'fit for purpose' look like?

Wherever possible, data needs to be managed and provided in accordance with well-established community and international standards, in order that the utility for community and stakeholders is maximised. Such standards provide a reference point that allows application software and services to be developed with confidence. It is important to recognise that standards are most effectively employed for data at the level of the community, or domain – above the level of the organisation, but sub-national.

The process of standardisation includes implementing digital methods for checking data against the standards, and processes for improving compliance. Such methods and processes increasingly need to be automated. It should be recognised, too, that there are ongoing costs for standardisation. In an increasingly linked world, multiple data standards apply and the technologies for delivering these need to be managed to accommodate evolving standards.

Well used international data standards must be able to accommodate ongoing innovation in data use and access. Thus, tdata standards must themselves be underpinned by appropriate mechanisms for both accommodating user community requirements, and allowing the standards to evolve the standards for innovation. Ensuring this generally requires representation on the applicable standards body.

It should also be noted that compliance with community standards is distinguished from ensuring data is "fit for purpose". The latter is a more extensive activity that provides additional levels of quality assurance processes to be formalised. This may include representative use-cases, benchmark measures, collaboration, and quality assurance in an innovation environment.

Q. What criteria and decision-making tools do government agencies use to decide which public sector data to make publicly available and how much processing to undertake before it is released?

Clearly, there will be different criteria applied for different data collections. As noted in the response to the first question, addressing issues such as which data to make available in what order (including criteria for this), which standards to use and which tools and services to bring together with the data, is often best decided in consultation with key parties from a broad domain or community of usage, within which the data will have the most relevance and will deliver the most value through re-use. In the case of NCI's NERDIP, for example, this is a broadly defined environmental research domain.

Community/domain level approaches to this question can be useful in resolving issues which sometimes arise in relation to decisions over which data to release, and how to do so. For example, in some cases there is concern from potential users that relevant data is not made accessible, and that approaches to processing could be improved by the research community. While this is often the case, the overall value of this would need to be assessed in each situation.

Q. What specific government initiatives (whether Australian Government, state, territory or local government, or overseas jurisdictions) have been particularly effective in improving data access and use?

The ongoing support under the National Collaborative Research Infrastructure Strategy (NCRIS) support for the NCI has been the essential enabler of the NCI activities in data management within an integrated high performance computing environment.

This has been through both NCRIS support for the NCI computational capability (I.e., the NCRIS NCI project), and through support for research data initiatives which have been leveraged by NCI (the NCRIS Research Data Storage Initiative [RDSI] and its successor, the Research Data Services [RDS] project). The latter supported the NCI High Performance Data (HPD) Node, which was itself made possible by the leading national capability which has evolved at NCI. Additionally, the NCRIS NeCTAR program funded both the virtual laboratory infrastructure developed at NCI, and the high performance cloud infrastructure managed at the NCI.

Delivering an integrated outcome from all these initiatives was achieved through the consolidated NCI strategic plan, supported with co-funding and contributed resources from the NCI partners, the Australian Research Council's *Linkage Equipment Infrastructure and Facilities* scheme, and in-kind contributions from research communities such as the ARC Centre of Excellence in Climate Systems Science, ARC Research Hub for Basin Geodynamics and Evolution of Sedimentary Systems (Basin Genesis Hub) and other NCRIS community infrastructure activities such as the Australian National Data Services (ANDS), AuScope, IMOS, and TERN.

RESPONSES TO QUESTIONS ON DATA LINKAGE

Q. Which datasets, if linked or coordinated across public sector agencies, would be of high value to the community, and how would they be used?

In line with an earlier response, we suggest that an alternative approach which will deliver significant benefit is to identify datasets from different public sector agencies which are of a similar *type*, which would be more valuable if better linked.

Examples include are linking federal and state/territory LIDAR data, as well as federal and state/territory geological survey geochemistry data.

Q. Which rules, regulations or policies create unnecessary or excessive barriers to linking datasets?

There are two elements to this matter, which operate at the inter-organisational and intra-organisational level. Regarding the former, NCI is aware that not all the state agencies are aligned with the federal rules, regulations and policies. In these circumstances, the regulatory barriers to linking datasets are often a matter of regulatory harmonisation, rather than identifying regulations to be removed.

At the intra-organisational level, agencies with a specific purpose and mandate (E.g., science agencies) can often encounter issues when complying with whole-of-government regulations, resulting in some datasets being not linked or inaccessible.

We also aware that some agencies are obliged by regulations and their statutes to operate on a cost-recovery basis. These agencies may therefore be limited in their capacity to make data more easily accessible, including by linking their data.

Q. How can Australia's government agencies improve their sharing and linking of public sector data? What lessons or examples from overseas should be considered?

The cited example of moving data from different agencies to NCI, and building it into a multi-disciplinary, integrated HPC/HPD platform for Earth systems information has provided a significant platform to improve access, collaboration and innovation within the research and government science agencies. The platform is built on the infrastructure and expertise aggregated at NCI, and represents an investment of resources and skills which any one partner would have been unable to make. By providing an avenue for multiple parties to participate in the development and use of such a platform, NCI has delivered researchers from all partners—along with their Australian and international collaborators—the capacity to undertake science they would not otherwise have been able to do. Within the international scientific community, there is considerable interest in this model.

Conversely, we are aware that there is considerable concern in some US government agencies about the model of using public companies to 'freely' host data, and then limiting others ability to effectively access that data, including by putting it behind paywalls. This trend is a double-edged sword. It is disruptive to agencies who then have less performant access which inhibits their potential for innovative activity. It also results in important capabilities being outsourced to a privately held entity that has captured the data, and which acts in the commercial interests of the company rather than the broader interests of community, government, research or even other businesses. These potential users may only deliver maximum socio-economic benefits in response to mostly unimpeded access to useable government datasets. This approach to data management needs to be carefully understood since there are both benefits and deep concerns around this trend. The NCI model avoids these issues by operating as partnership of agencies who have 'skin in the game' by virtue of their support for operations, ensuring NCI acts in their interests.

RESPONSES TO QUESTIONS ON RESOURCE COSTS OF ACCESS

Key Points:

- Resources are required for:
  - Ensuring data is well specified, consistently defined, accurate and available in a usable format.
  - Maintaining consistency with evolving standards, being aware that information systems and data collections standards do not apply uniformly.
  - Upgrading digital storage media and digital security in line with changes in technology.
- Preparing data for release is a resource intensive task.
- Skilled specialist staff are required for all stages, including making the data accessible and then maintaining it.

Q. How should the costs associated with making more public sector data widely available be funded?

Accommodating government agencies' operational and business models when moving to ensure they make their data more available—thereby requiring them to cover the cost of supplying a public good—needs to be considered when planning for data availability. In particular, the issue of sustainability for smaller and medium sized agencies of maintaining high-quality data accessibility needs to be considered. This reflects that fact that making data available and then maintaining that availability can be costly, encompassing either or both of ongoing capital and operational costs, depending on the approach.

This is particularly the case when it is remembered that once an agency makes its data available, there is a high chance that dependencies will develop as other agencies (including front-line agencies such as emergency services) incorporate that data into their own operational processes. For this reason, data availability must be reliable – with reliability being ensured by highly robust processes and infrastructure. Needless to say, robustness not only comes at a price, but requires considerable expertise.

An approach which has achieved great efficiencies in making agencies' data accessible and useable is exemplified by the approach enabled through NCI, under the auspices of the NCRIS program. The funding arrangements established under NCRIS acted to facilitate deep collaboration by government agencies with each other, and with other stakeholders in the research community. This was done in a way which, as a consequence of the NCI collaboration having to meet both academic and agency research needs, delivered data infrastructure services which have driven not just research outcomes, but innovation in research processes.

For reasons of both efficiency and effectiveness in achieving outcomes (I.e., the outcome of improved use of the data), this is approach is far more preferable to the siloing of data activities within individual agencies, or funding these only through vertical activities run solely by government (departments and agencies). By adding resources to a government investment, a capability is delivered which exceeds what could be provided by any one or two agencies. The capability represents an aggregation of multiple infrastructures and skills sets, and can provide robust, high quality data infrastructure services which meet the needs of multiple parties, to whom it is responsive as funders.

Importantly, this approach has delivered not just reliable, high quality services, but also *innovation* in the delivery of data services for government and research data. Innovation is, and will continue to be, essential in the area of data services, as *data is not static*: collections grow and change, standards and tools continually develop, and user needs constantly evolve. The services provided by NCI therefore differ from the commercial hosting of a defined data service, because the NCI

Collaboration exists to provide for its partners *ongoing*, *evolving* research data needs. NCI is therefore geared to respond to its partners changing needs, by innovating in its development and provision of data services and management techniques.

This model, in which direct support combines with agency co-investment to underpin a facility which is specifically geared to facilitating broad uptake and access, has the greatest potential to drive necessary innovation in the development and operation of data services, and is likely to be most effective for agencies.

NCI therefore suggests that any funding associated with making data widely available (accessible and re-useable) should build upon extant large national investments, such as have been made under NCRIS. The marginal costs of building on extant specialist (research) data infrastructures to make them sufficiently robust and tailored for broader agency use, renders such an approach more efficient than building multiple new, separate infrastructures for this purpose.

~

NCI notes that a deficiency currently evident is that continued funding—in general, not only for NCRIS—for core research capabilities which sit outside agencies and departments to provide specialist services for these parties, is not assured. What funding there is, has uncertain timelines.

The key issue arising from this uncertainty is for agencies themselves, who cannot undertake strategic planning based on the funded external capability, and so cannot commence to manage their own risks in the most efficient way possible (I.e., by collaborating around data and pooling resources for specialist activity). A follow-on effect of this is multiple agencies making similar but independent requests for funding for similar activities.

Q. What pricing principles should be applied to different datasets? What role should price signals play in the provision of public sector data?

When Government considers this issue, it will be important to be clear on the purpose of any policy to improve accessibility to data. There will be likely be a significant issue if the purpose is to drive innovation in all socio-economic sectors (rather than just commercially), because pricing datasets comes with significant risks to uptake, and therefore to any wider innovation dividend. This was touched on earlier, when we noted that stipulating agencies must operate on a cost-recovery basis has the potential to negatively impact how much they invest in making data available, and therefore its accessibility and useability (I.e., potential value) to a wide range of stakeholders.

Of more significance, however, is the high likelihood of an inverse relationship between the price of public data, and its uptake. Indeed, in the context of driving improved outcomes in government and the publicly-funded research sector (the two most like users) by making public sector data more available, there is something of a contradiction in requiring agencies to invest in making their own data available, and then further requiring that they pay for access to data from other agencies.

It will be necessary to understand the potential impact of data in order to evaluate the benefits or drawbacks of pricing it. This will include understanding who will make most use of public sector data when it is made available, and thus ascertaining price sensitivity of the most likely users (there will be little point going to the effort of making data available if it comes at a price that most users cannot, or will not pay). More importantly, however, it will require understanding the use to which data will be put and the nature of the benefit which will be derived from it. Many of the uses to which public sector data will likely be put, will not deliver clear commercial benefit. Rather they are likely to contribute to outcomes such as improved evidence for government policy, or increased accuracy of research findings.

NCI suggests that developing and implementing impact measures for data is a more established approach to understanding the value of those data, and associated data services. Impact measures would enable a clearer understanding of the return on the investment in making data available – noting that public sector data has already been gathered and used for its initially intended purpose. More importantly, implementing impact measures would enable appropriate prioritisation of resources around datasets, including through understanding the characteristic performance requirements, and costs to support access to them.

Q. Is availability of skilled labour an issue in areas such as data science or other data-specific occupations? Is there a role for government in improving the skills base in this area?

Skilled labour in data science and management is essential to support data being made accessible and useable, but remains very difficult to recruit and retain. This is a critical issue for entities who rely government funding, or on support from publicly-funded organisations, to deliver data services. It is also a particularly pressing issue for entities who provide services for high-end users, such as researchers.

There is an increasing awareness of the need for skilled labour, and the difficulties in recruiting and retaining it. Translating this awareness into increased supply of skilled labour will take time, however. In the meantime, difficulties in recruiting and retaining suitable expertise significantly increases the value of entities which have built up an expert data science and data management capability, such as those enabled built by the NCRIS initiatives. Accordingly, it will be important not to spread the pool of skills too thinly, by supporting fewer, larger entities which build on existing capability.

The NCI Data Capability

What is NCI?

The National Computational Infrastructure, as Australia's national research computing service, is funded by the Australian Government to provide high-end services to Australia's researchers, and in so doing raise the ambition, impact, and outcomes of Australian research through access to advanced computational and data-intensive methods, support, and high-performance infrastructure.

NCI offers comprehensive and integrated high-performance services that span the gamut of computationally based research. This encompasses computationally-intensive and data-intensive workloads, with a focus in environment fields, and in climate and earth system science, in particular.

The infrastructure investments at NCI, which total approximately $80M since 2007, have been provided by the Australian Government under its National Collaborative Research Infrastructure Strategy (NCRIS), including support under the Super Science Initiative.

NCI operates as a formal collaboration of the ANU, CSIRO, the Bureau of Meteorology (BoM), and Geoscience Australia (GA), which is strengthened by formal partnerships with science agencies, medical research institutes, and a number of research-intensive universities supported by the Australian Research Council. These organisations collectively provide considerable co-investment that, complemented by support from NCRIS, provides for NCI's recurrent operations.

NCI's High-Performance Data Capability

NCI has established a widely used, highly regarded High Performance Data (HPD) Node for the research community. This has been done using funding from under the Research Data Storage Initiative (RDSI), made available as part of NCRIS, an Australian Government program. The HPD Node manages and serves data collections of high-value to the research community. NCI's strategic goals for the HPD Node have led to it prioritise the accessioning of data collections that:

a) Require either high-performance computational (HPC) or data-intensive analysis through (HPD) services in order to make more effective use of the data, with either existing or emerging technologies and techniques;
b) Are associated directly with the HPC/HPD class above and which should be co-located with them;
c) Are of national/international significance, and which is selected in the national interest, by the supporting stakeholders of NCI; and
d) Are requested by research communities to be served by NCI.

More particular priorities were then identified. These were used to prioritise datasets by establishing which would have a higher potential value to users when incorporated into NCI's HPD environment. In NCI's case, the distinctive characteristics that led to data collections being prioritised as high-value, and to be managed at NCI were:

- Making accessible major data collections currently held by NCI's partner national science agencies - Australian Bureau of Meteorology, CSIRO, Geoscience Australia;
- Complementing and augmenting these collections with other nationally and internationally significant collections that are priorities for cognate research communities;
- Combining datasets held by different research communities into coherent collections; and
- Integrating the collections from both the research and government sector, and then establishing these collections in a rich environment of high-end computational and data-intensive services.

Implicit in these priorities was the intent to:

- Capitalise on the rich computational environment at NCI; and
- Engender an outcome of mutual value to NCI stakeholders, the national research community and the national innovation agenda that would make available national collections in such an environment

To determine how to allocate resources—including storage—to datasets of highly valued data collections, NCI undertook a multi-stage peer review process. The ability to do this was underpinned by a funding model supported by the NCI NCRIS business plan with further support and co-funding support from NCI's partner organisations. In particular, the NCRIS RDSI initiative provided approximately 10 Petabytes of data storage on NCI's high performance filesystems, which included augmented the data centre capability to house this data alongside the other high performance capabilities at NCI.

Proposals for datasets were assessed by a Data Allocation Committee to ensure that the data collections aligned with the NCI prioritisation and high-level objectives. The Committee was supported by a review panel that consisted of skilled and respected scientific and technical staff to ensure that the proposed data collections were thoroughly reviewed and met both the science and technical merits of being allocated resources at the NCI.

The data collections available at NCI can be discovered through its catalogue discovery portal: http://geonetwork.nci.org.au/

Datasets were ranked particularly highly by the Data Allocation Committee where making the data available was beyond the capabilities of any individual, research group, or even any single organisation. In the era of Big Data, such massive data needs to be co-located with high performance computing infrastructure so that the data can be processed and analysed in-situ without requiring the additional overheads of downloading data, and installing additional software environments.

Truly massive data collections are beginning to emerge that are beyond the capacity of any one nation to manage. One of the first nationally significant examples of this was the Coupled Model Intercomparison Project (CMIP) Phase 5[3]. This data collection is internationally approximately 30 Petabytes and is managed across several continents using the Earth System Grid Federation[4], and NCI is the node for the Australasian region. This data underpinned the IPCC Annual Report 4[5] which is used for global climate analysis.

As a consequence of the capabilities NCI demonstrated in the process outlined above, existing and new stakeholders of NCI have identified priority datasets to be stored in the HPD environment, so as to capitalise on the capabilities now in place. Some of these partners have contributed additional funding for further data storage as well as contributions to operations, in order to store data at NCI. Stakeholders' reasons for this decisions include aligning their data with the datasets already organised at the NCI, and taking advantage of other capabilities of NCI such as its data management and sharing practice. Two key examples are:

- The Australasian Copernicus Hub which hosts the European Sentinel Earth observation data has now been selected to be managed at the NCI[6]. The hub is projected to provide access to over 12 Petabytes of satellite Earth observation data by 2025, and is expected to go beyond simply providing users with the ability to download Copernicus data. The NCI data access hub

[3] http://cmip-pcmdi.llnl.gov/cmip5/

[4] http://esgf.llnl.gov/

[5] https://www.ipcc.ch/report/ar5/

[6] http://www.ga.gov.au/news-events/news/latest-news/Establishment-of-a-satellite-data-hub-to-benefit-Australia-and-international-partners

is the largest facility of its kind in the southern hemisphere, taking advantage of the Australian government's investments in science and research infrastructure to support the region. This data will then be managed in a consistent way to the other long-term satellite data - notably the Australian Landsat archive, and the NASA MODIS archive.

- The Garvan Institute of Medical Research has chosen to host their genomics data at NCI to ensure that the large volume of reference data can be managed for initial processing and subsequent gene sequencing analysis at the NCI.

NCI, in partnership with its stakeholders also assists in turning private sector data into public sector data. One example is GA's geophysical data held at the NCI. This is data collected by the private sector such as for oil and gas exploration, which is then submitted via State and Territory Geological Surveys or directly to GA. In their submitted formats, these datasets by themselves are difficult to analyse. However, by creating higher level products using tools hosted within the NCI HPD environment, these data can be made into data products which form national reference datasets.

Further, protected government data which is not directly available to the broader community, is used in a compute intensive process calibration process with other very large datasets that is used to generate open data products that are able to be made publicly available. One such example is from satellite Earth Observation where over 30 years of individual data scenes have been calibrated and combined into a dataset that can be analysed spatially at either a national scale or at the scale of a local farmer's paddock and at time intervals from 30 years to one month.