

Data: adding value to the Australian economy and environment

Tim Thwaites, Science in Public



IMOS collects a wide range of data throughout Australia's coastal and open oceans. Image courtesy of Kim Brooks, AIMS

Early in 2013, Australian and Japanese researchers announced that a pod of seals had helped them find a fourth outflow of cold dense water from Antarctica into the world's oceans - a discovery of great significance for understanding the world's climate. The foraging seals were carrying sensors that transmitted critical data to Hobart via satellite for posting on the website of Australia's Integrated Marine Observing System (IMOS: imos.org.au).

Up above, much of the early evidence for the mysterious cosmic 'fast radio bursts' was discovered while combing through information in the public archive of pulsar data from the Parkes Radio Telescope set up by the CSIRO with the help of the Australian National Data Service (ANDS). A team led by a doctoral student at Swinburne

University has just detected one of these celestial events 'live' for the first time.

These are just two of the important scientific discoveries that would never have happened without Australia's increased capacity to capture, store, process, access and recycle data. The National Collaborative Research Infrastructure Strategy (NCRIS), which funds both IMOS and ANDS, has played a pivotal role in moving Australia to the forefront of the collection, organisation and use of data as a scientific tool.

"There's been a perceptual change in science," says Prof Aidan Byrne, Chief Executive Officer of the Australian Research Council.

Continued on page 8.

Inside Issue 21 - The value of research data to the nation

- » Adding value to Australia's research data
- » How past observations are building astronomy's future
- » CSIRO's marine vessel steers towards a new data future
- » Med.Data: where data intersects with national health
- » Data the key to regional development
- » Costs and benefits of data provision

Executive Director's report

Ross Wilkinson, ANDS

How can Australia address the challenge of living in bushfire prone city fringes? How can Australia most effectively farm and preserve our precious soil? How can Australia understand the Great Barrier Reef? No single discipline can answer these questions, but to address these challenges data is needed from a range of sources and disciplines.

Research data that is well organised and available allows research to make substantial contributions vital to Australia's future. For example, by drawing upon data that is able to be used by soil scientists, geneticists, plant scientists, climate analysts, and others, it is possible to conduct the multidisciplinary investigations necessary to tackle truly difficult and important challenges. The data might be provided by a Terrestrial Ecosystems Research Network OzFluz tower, insect observations recorded by a citizen through the Atlas of Living Australia, genetic sequencing of insects through a Bioplatforms Australia facility, weather observations by the Bureau of Meteorology, or historical data generated by CSIRO over many decades. Each will provide a part of the jigsaw, but the pieces must be able to be put together. This requires careful collection and organisation, which together deliver enormous value to the country.

However, nationally significant problems are often tackled by international cooperation, so Australia's data assets enable Australian researchers to work with the best in the world, solving problems of both national and international significance. Australia's data assets and research data infrastructure provide Australian researchers with an excellent platform for international collaboration.

Australia has world-leading research data infrastructure: our ability to store, compute, discover, explore, analyse and publish is the best in the world. The ability to capture data through a wide range of capabilities, from the Australian Synchrotron to Integrated Marine Observation System [IMOS: imos.org.au] ocean gliders, the combination of national storage and computation through RDSI, NCI and Pawsey initiatives, the ability to publish and discover data through ANDS, the ability to analyse and explore data through Nectar, and state and local eResearch capabilities, highlights just some of the capabilities that Australian researchers are able to access. Importantly, their international partners are able to work with them using many of these resources.

As well, Australian research organisations are assembling many resources to support their research. These include policies, procedures, practical infrastructure, and very importantly – people! The eResearch team and the data librarians are always keen to help.

This issue of Share highlights how the data resources of Australia are providing a very substantial national benefit, and how that benefit is being realised.



Dr Ross Wilkinson

Connecting the infrastructure pieces to add value to Australia's research data

Andrew Treloar, ANDS

It's always good to have an advantage over one's competitors. It's even better to be valued as a research partner when overseas researchers are looking for collaborators to tackle big problems. And it's best of all when one is based in a country that is only responsible for 3 per cent of the world's research output (despite producing just 1 per cent of the world's GDP with 0.3 per cent of the world's population!).

Which country? Australia. Australian researchers can draw on the legacy of a decade of farsighted investment in research

e-infrastructure, and seven years of investment specifically in research data infrastructure. This has delivered world-leading capabilities to allow researchers to store, compute, discover, explore, analyse and publish their data.

The Research Data Storage Initiative (RDSI: rdsi.edu.au) will have made available more than 50 petabytes of data storage at state-based nodes by the end of 2014.

The National Computational Infrastructure (NCI: nci.org.au) and Pawsey Centre (ivec.org/facilities/pawsey) already provide systems

ranked at 52 and 41 respectively in the Top 500 high-performance computer centres (as of November 2014) to process these large data volumes.

The National eResearch Collaboration Tools and Resources (NeCTAR: nectar.org.au) project is augmenting this peak performance with highly-scalable compute cores available to any Australian researcher, as well as a range of e-research tools and virtual laboratories to provide specialised functionality for researchers in particular disciplines.

These national investments are complemented by e-research support services provided by state-based nodes and by some universities.

ANDS is working with these state-based nodes to ensure these infrastructure pieces mesh seamlessly together to support researchers as they seek to answer big questions.

One example is a project ANDS is working on with the Genomics Virtual Laboratory (GVL: genome.edu.au/wiki/GVL) to ensure better connections between data stored on RDSI-funded storage, tools running on NeCTAR-funded compute cores, the existing GVL, and ANDS' data publishing services.

The result will be a more seamless data publishing lifecycle, and an enhanced GVL that is moving from its initial workflow-centric

version to something that aspires to be genuinely data-centric.

Another example (in which ANDS has contributed much of the funding) is a cluster of data services and systems in the area of climate adaptation research.

The TerraNova system (terranova.org.au) is improving the accessibility and availability of climate change adaptation data and information to researchers and decision-makers in Australia and around the world.

The CliMDDIR (Climate Model Downscaling Data for Impacts Research: climddir.org) service is making the outputs from Regional Climate Model (RCM) simulations of the recent and predicted future climate of Australian regions available to climate change impacts researchers in data formats they can process more readily.

The Ecosystem Production in Space and Time (ePiSaT: episat-software.blogspot.com.au) software models Gross Primary Production (GPP), a measure of the changing productivity of the continent, across Australia from OzFlux flux tower data, gridded climate and satellite data.

As Prof Andy Pitman says (see text box below), the result of these investments, and many others like them, is a system that when connected provides real competitive advantage to Australian researchers.

It's always good to have an advantage over one's competitors. It's even better to be valued as a research partner when overseas researchers are looking for collaborators to tackle big problems.

Prof Andy Pitman, Director of the ARC Centre of Excellence for Climate System Science at the University of New South Wales

"It seems to me... that no country has brought a system together quite as well as Australia to manage this complexity.

In terms of climate science, NCI provides our computer resources and RDSI our storage needs. ANDS helps us understand how to manage our data and publish it. NeCTAR provides tools to help process data, to manage codes and to develop an efficient workflow system (a virtual laboratory if you like).

Australia has put in place the pieces of the jigsaw in a way that really is world-leading and unique in covering all the key components...

The definite 'good news' story for Australia is that ANDS ... is clearly recognised as a leader in finding the solutions by similar programs overseas. ANDS, in combination with newer program NeCTAR, supported by hardware infrastructure through NCI and RDSI, genuinely seem ahead of their equivalents overseas."



Prof Andy Pitman

How past observations are building astronomy's future

Tim Thwaites, Science in Public

China is building the world's largest single radio telescope. Known as FAST (the Five hundred meter Aperture Spherical Telescope), it will be three times as sensitive as the current largest dish at Arecibo in Puerto Rico, and will be used to study pulsating neutron stars, or pulsars. China has limited pulsar capability, so many astronomers who will be using the FAST telescope are learning their trade by analysing the world's only public database of pulsar data, CSIRO's Pulsar Data Archive (atnf.csiro.au/research/pulsar/), established with funding from the Australian National Data Service (ANDS).

Not only has this drawn attention in China to Australia's expertise in this field - CSIRO's Parkes Radio Telescope has been involved in the discovery of more pulsars than any other facility in the world - it has also sparked a collaboration likely to lead to access to the latest technology and data for Australian astronomers. Already, Dr George Hobbs, a pulsar expert from CSIRO Astronomy and Space Science (CASS), is off to spend two months working on the new telescope.

"Astronomy is an extremely collaborative and international subject," says Dr Jessica Chapman who heads up a data management team at CASS. "If you look at our research publications, around 85 per cent of them have teams that include both Australian and overseas people. In radio astronomy, the deal has been that we provide our facilities for free to those who are allocated observing time on the telescopes, but in return we can apply for telescope time on most

other international facilities." The pulsar data archive extends that philosophy nicely.

"Many astronomers have little experience with looking after data," says Hobbs. "Historically, each observer took their own data from the telescope. Then their copies often disappeared somewhere. Now we have the professionals doing it for us." Initially the archive team had to chase up historical data from important projects, but now the system is automated.

Every six months data from all the projects undertaken at Parkes are described and archived. The astronomers who took the observations have exclusive use of their data for 18 months. Then they become available to all, retrievable through CSIRO's Data Access Portal (DAP: data.csiro.au) developed by CSIRO Information Management and Technology and CASS with ANDS support, or Research Data Australia where descriptions (metadata) of what is in the archive are automatically stored.

The archive contains 100 terabytes (100,000 gigabytes) of pulsar data and is rapidly increasing each year. The use has been phenomenal - more than 40,000 visits and 90,000 downloads by 600 users last year from all over the world including the US, Italy, Canada, Russia, the UK, Germany, South Africa and India.

The archive data can be searched for missed pulsars using powerful modern computers, techniques and algorithms. It also provides researchers with material in which they can look for celestial phenomena that were unknown when data was originally collected, such as the fast radio bursts originally thought to come from well outside our Galaxy, and the mysterious 'Perytons' which appear to be fast bursts located from inside our Galaxy. These radio bursts were first discovered in pulsar archive data and have only just been 'seen' live for the first time.

The expertise gained in building the pulsar data archive is being put to good use, with similar archives created for the massive Square Kilometre Array (SKA) and its precursor the Australian SKA Pathfinder (ASKAP). Data collected from ASKAP's 36 linked small dishes will be stored and processed at the Pawsey Supercomputing Centre in Perth. As an international project, it needs to be widely accessible.



An almost full moon shines brightly over the CSIRO Parkes radio telescope. Copyright: Shaun Amy

CSIRO's marine vessel steers towards a new data future

John Morrissey, CSIRO

CSIRO's new marine research vessel, *Investigator*, has been constructed to collect significantly more data than any of its predecessors.

Investigator, currently undergoing early science and engineering sea trials, carries a larger scientific crew, has a significantly expanded suite of scientific instruments and a greatly extended range. It can operate to the south as far as the ice edge.

An active marine research vessel such as *Investigator* is essential in order to provide the capacity to accurately and rapidly detect and predict changes in the ocean environment, monitor coastal ecosystems and living resources, and investigate and assess the unique marine geology surrounding Australia.

Operated by CSIRO, access to the Marine National Facility (MNF: mnf.csiro.au/), and its research vessels has always been open to Australian researchers and their international collaborators, and is allocated on the basis of scientific merit.

The MNF's operations are funded by the Australian Government to enable oceanographic, geo-scientific, fishery and ecosystem research.

Some of *Investigator*'s new capabilities include atmospheric instruments to measure wave heights, precipitation, sunlight, aerosols, and other components of the marine atmosphere.

Capability for geosciences includes a coring facility, side scan sonar and sub-bottom profiler.

These will often be augmented by additional sensor platforms brought onboard for particular voyages by researchers from external organisations.

Investigator is also well equipped for collecting biological samples from a wide range of marine environments.

Combining *Investigator*'s diverse datasets with live data feeds and voyage logs presents a challenging environment in which MNF staff have to manage the data for both physical and digital assets delivered ashore at the end of each voyage.

The MNF collects a significant amount of data managed by CSIRO on behalf of the nation. This management involves archiving raw and processed data, creating and publishing suitable data descriptions and publishing these data in ways suitable for later reuse.

So how does the MNF and CSIRO make all these data visible and available to researchers across the globe within a reasonable timeframe?

A major project underway at CSIRO is the creation of a new long-term digital archive for all end-of-voyage data, known internally as the Marine National Facility Science Data Archive (MNFSDA).

This archive will utilise CSIRO's existing Data Access Portal (DAP):



The 93.9 metre *Investigator* when it was under construction. Image courtesy of CSIRO (CC BY 3.0)

data.csiro.au) to ensure all these data are managed and available as a long-term national asset.

As MNF Director Ron Plaschke says, "The new data archive is an important step to providing a secure and enduring archive for increasing data flows from *Investigator* as well as improving accessibility for voyage scientists towards making MNF data visible and publicly available from a range of sources."

The raw data from *Investigator* aren't always in a format that is optimal for access by researchers or industry. They usually require some form of post-processing, which creates a number of data products that then need to also be managed and made accessible to others.

In the MNFSDA's case, some of these raw data will be post-processed by other organisations and research groups such as Geoscience Australia and the Bureau of Meteorology. This is where the DAP's ability to make data selectively available to collaborative partners is being used to great advantage.

Working with such a complex platform forces one to take a wider view of data management with the realisation that digital repositories are just one small component in what can be considered an 'ecosystem' of services and tools required to support national research facilities.

One key outcome of the project will be the creation of a series of 'use cases' that will be used to further develop a series of follow-on projects looking at a number of areas, including workflow automation for speeding up the delivery of MNF data products to researchers, and provenance management systems that include workflows involving external partners.

To learn more about *Investigator*, read the MNF's blog (csirofrvblog.com), which is updated on a regular basis.

John Morrissey is an eResearch Planner at CSIRO.

Med.Data: where data intersects with national health

Leonie Hellmers, Intersect

A new national initiative, named Med.Data, will create a data collection of profound significance for Australia.

The facility will enable, for the first time, the sharing of health and medical datasets generated, curated and managed by many of the dozens of universities, medical research institutes and state government health departments who contribute to health and medical research in Australia. It will also make available medical research data which is currently not accessible.

Med.Data will provide the underlying infrastructure to support the safe retention and sharing of health and medical research data so as to enable collaboration and reuse with the necessary ethical and technical access protocols.

Intersect, an eResearch agency, leads Med.Data which is funded by the Research Data Storage Infrastructure (RDSI) project.

Med.Data will be hosted and operated by Intersect, Queensland Cyber-infrastructure Foundation (QCIF), Vicnode and eResearch SA, each being a 'node' of the RDSI project and each based in a different state of Australia.

To date, 47 health and medical research institutions have nominated data collections to be hosted by Med.Data.

To make the richness of these data collections valuable, they need to be well-described and readily discoverable; this will assist collaboration for researchers across the spectrum of the many contributing institutions.

Better research data management leads to better research impact.

To this end Intersect and ANDS have agreed to collaborate in creating a rich store of medical data which offers additional value

Continued on page 9.

Who is contributing data?



The data resources being assembled through this initiative vary widely and will be drawn from the agencies above.

Data the key to regional development

Tim Thwaites, Science in Public, and Shannon Lindsay, ANDS

The Australian Government demonstrated its keenness last year to develop the nearly 40 per cent of the country north of the Tropic of Capricorn by releasing a green paper to stimulate discussion.

This year, it plans to deliver a white paper to guide development policy for the region. Access to data is pivotal to all of the policy directions discussed in the green paper: delivering economic infrastructure; improving land use and access; improving water access and management; promoting trade and investment, and strengthening the business environment; fostering education, research and innovation; and enhancing governance.

Information is also central to the success of the proposed AgNorth Cooperative Research Centre, a collaboration between government, business and academia to support the development of industries to help satisfy the significant new markets for food and fibre opening up in the countries to the north.

A team at James Cook University's eResearch Centre, supported by ANDS, has created the Tropical Data Hub (TDH: tropicaldatahub.org) to bring together information on anything to do with the tropics in Australia.

With campuses in Townsville and Cairns, JCU is one of only two tropical universities in Australia and one of the few tropical universities in developed nations.

Yet the tropics is where two-thirds of the world's population lives, facing distinctive issues of health, agriculture, the environment and managing resources.

The TDH is aiming to become a repository of tropical information of national and global importance. For developing Australia's north, the TDH houses a host of useful datasets, such as the potential for food and wine tourism in the Cairns region, algal biochar's impact on plant growth, and a time-space distribution of iron oxide copper-gold deposits to lead to more focused mineral exploration and research.

The TDH also features many datasets relevant to the Great Barrier Reef, from current surveys to threats such as herbicides and river drainage.

Protecting this World Heritage area is a national and global issue. Large parts of the tropical north rely on the reef as a drawcard for tourism.

Protecting natural resources partly in the name of tourism is also an issue for the Gold Coast.

Griffith University's Centre for Coastal Management (GCCM) works in partnership with the City of the Gold Coast to improve and develop new best practices in beach sustainability. GCCM has



James Cook University's Tropical Data Hub features many datasets relevant to the Great Barrier Reef. Image courtesy of Jon Milnes/shutterstock.com

produced a 15-year strategy: *Gold Coast Shoreline Management Plan* (GCSMP). It contains a set of 77 recommendations for the economic, social and environmental sustainability of the Gold Coast's beaches that will assist the Gold Coast City Council with best practice coastal management.

These recommendations were based on research datasets gathered through a range of major projects.

There are a number of other examples of how data initiatives and datasets contribute to regional development in direct or indirect ways.

The Australian Urban Research Infrastructure Network (AURIN: aurin.org.au) is helping create more liveable Australian suburbs and cities through integrating multiple data sources (e.g. state, local and federal government and utilities data) and producing such software tools as employment modelling and walkability simulations.

The Terrestrial Ecosystem Research Network (TERN: tern.org.au) and the Atlas of Living Australia (ala.org.au) have combined ecological and genomics data of Australian soils, animals and plants to produce a software tool – Soils2Satellites (soils2satellites.org.au) – to help plan for and stem economic and environmental losses due to plant and animal pests.

When it comes to improving the Australian economy and environment – it's all about the data.

Data: adding value to the Australian economy and environment

(continued from page 1)

"We used to see physical hardware - equipment - as the principal thing. Now we're in an environment where data itself has an intrinsic value."*

For instance, while the Square Kilometre Array (SKA), the world's biggest radio telescope, part of which will be built in Western Australia, will produce fascinating astronomy," he says, "that's not the only reason why people are interested in it. They are also interested because it will become the template for how to deal with high volumes of data."

NCRIS is an Australian Government fund that currently provides more than \$100 million a year to support major national research infrastructure, beyond the reach of a single institution, but not on the large scale of facilities such as the Australian Synchrotron or the Opal Research Reactor.

NCRIS facilities include those which generate, accumulate and provide access to data, such as IMOS; the Terrestrial Ecosystem Research Network (TERN); AuScope for earth science data describing Australia; the Australian Urban Research Infrastructure Network (AURIN); Astronomy Australia; the Population Health Research Network (PHRN); Bioplatfroms Australia (BPA) for genomic data; and the National Imaging Facility (NIF) and the Australian Microscopy and Microanalysis Facility (AMMRF) for large and complex image datasets.

But probably more innovative in international terms are the facilities

that add value to data. They include ANDS which assists institutions big and small to find solutions to data management problems; the National Computational Infrastructure (NCI) which provides access to supercomputers and other research computing; National eResearch Collaboration Tools and Resources (NeCTAR) which provides software tools, virtual laboratories and a cloud environment for research; and Research Data Storage (RDS).

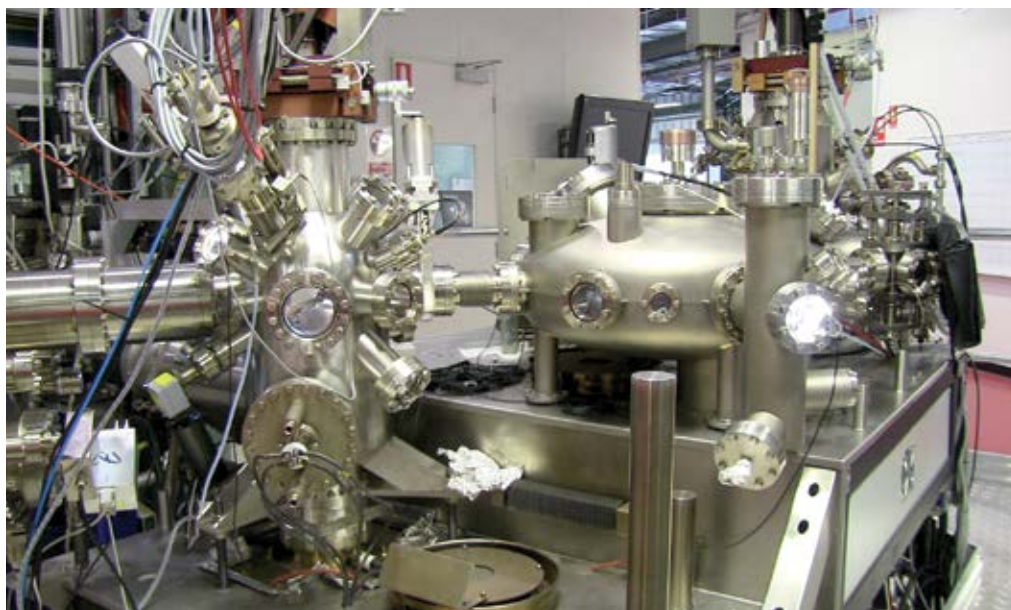
Monash University's Vice-Provost (Research and Research Infrastructure), Prof Ian Smith has been the driving force behind the development of an array of about 20 underpinning technology facilities for research, which are among the best in the world at any university. "All of our capabilities, whether they be for imaging, 'omics data' or materials analysis, are generating increasingly sophisticated and complex datasets. The NCRIS facilities are helping us in terms of storing the data, curating the data, processing the data, and distributing the data."

"They've given researchers more confidence to take on and handle large datasets and, more importantly, to extract the information they need, to mine and interpret and interrogate their data."

"NCRIS is a national program which, at Monash, has also given support to MASSIVE, the Multi-modal Australian Sciences Imaging and Visualisation Environment. MASSIVE provides supercomputing nationally to process complex imaging datasets, and RDS which provides storage for nationally important archives. These initiatives not only underpin Australian research, they knit together other NCRIS-funded capabilities, all of which produce large amounts of data."

"The Characterisation Virtual Laboratory (CVL) funded through NeCTAR, for example, provides unique desktop tools while allowing other NCRIS-supported facilities such as NIF and AMMRF real-time cloud access to MASSIVE. Such a capability allows the seamless integration nationally of imaging instrumentation, including at the Australian Synchrotron."

*See p 11 for an overview of the new Houghton report which highlights the value of public research data.



Inside the Australian Synchrotron. Image courtesy of Dave Jones (CC BY 2.0)

UWA's Prof Sydney Hall receives CODATA Prize 2014

The CODATA Prize 2014 was awarded to University of Western Australia Emeritus Prof Sydney Hall for his leadership in the world of scientific and technical data.

CODATA, an interdisciplinary scientific committee of the International Council for Science, said Hall's efforts reached beyond his field of crystallographic data, into many other science disciplines.

He devised a universal self-defining text archive and retrieval (STAR) file format that evolved into the Crystallographic Information Framework (CIF), a momentous contribution in the area of data characterisation, and well known to structural chemists and biologists in particular as both a data and publications standard.

It enables data validation for articles published by International Union of Crystallography journals, an approach pioneered by Hall in his role during the 1990s as Editor of Acta Crystallographica Section C: Crystal Structure Communications.

Continued on page 10.



Emeritus Prof Sydney Hall receives his award from President of CODATA, Sarah Graves, Professor of Computer Science at The University of Alabama.

Med.Data: where data intersects with national health

(continued from page 6)

to the researchers and the institutions providing the data.

Med.Data can become a huge asset for Australian medical research. By providing extra services; for example, connecting to other collections, describing the data, and providing richer insight into the data, the data will provide extra visibility and enable stronger international collaboration.

Offering these extra services on top of storage creates extra value to the storage service that Intersect provides, and makes a stronger case for providing custody of these important collections in the long-term.

Med.Data will harmonise the distinct and dispersed data from the many contributors who all have a role to play in Australian health and medical research.

Med.Data can then fulfill its aims to:

- » Deliver richly described collections that support discovery and citation
- » Provide cost-effective and sustainable data storage

- » House and support data underpinning medical research of identified national and international significance
- » Increase availability of medical research data for reuse and clinical analysis
- » Supply a foundation architecture that facilitates controlled access to, use and sharing of medical research data in accordance with ethical and privacy constraints
- » Offer access, storage, high performance compute and analysis optimised to machine and laboratory workflows
- » Permit mirroring of, access to, and linkage to other international datasets of similar interest to facilitate comparative studies
- » Support interface and sharing capabilities between academic and clinical researchers.

Med.Data acknowledges funding from the Research Data Storage Infrastructure (RDSI) Project, an initiative of the Federal Government, which is funded from the Education Investment Fund under the Super Science (Future Industries) initiative.

Chair's report

Ron Sandland, ANDS' Steering Committee Chair

One of the disadvantages of having a relatively low population is that the size of our market is quite small, especially compared to massive and affluent countries like the United States.

Entrepreneurs in the US have a ready market for their new businesses and products, not to mention a strong pipeline of discoveries and inventions locked in a Darwinian struggle for survival and ultimate success.

Companies like Amazon, Google, Apple, Microsoft, Facebook, Twitter and many others like them have dominated the ecology of the internet and social media revolutions.

And many of these companies have been built around the ruthlessly innovative collection, analysis, interpretation and display of data. They have discovered the secrets of unlocking the value of data. One comparative advantage that Australia has is our relatively small size means we can find our way around the whole system, embracing our research institutions and Government agencies.

This has enabled the Australian Government to invest in research infrastructure that embraces the national research enterprise and underpins a recognisable competitive advantage. Of course I am talking about NCRIS – the National Collaborative Research Infrastructure Strategy.

At the end of September 2014, I attended the Department of Education's NCRIS Showcase in the Great Hall of Parliament House in Canberra. The Department's description of the NCRIS investment ran thus: "Established in 2004, NCRIS provides Australian researchers with the major research facilities and networks necessary to deliver world-class research. Over \$2.5 billion invested by successive governments on NCRIS has successfully attracted

over \$1 billion in co-investment from universities, research facilities, state governments and industry. NCRIS currently employs over 1,500 staff in 222 institutions and supports an estimated 30,000 researchers across Australia and internationally."

When I attended this Showcase, I was singularly impressed by the breadth of the investment and its maturation.

In NCRIS, Australia has created something that enables its researchers to participate in an increasingly competitive international research system.

And it also enables them to tackle the large-scale problems that are vital to Australia's future, whether they be in health, the environment, urbanisation or industry.

The collection, transformation, analysis, and interpretation of data is a thread that runs through the great majority of NCRIS initiatives.

Its success depends to a significant extent on our ability to see the data resources that have been created not as a static, disposable thing but a living entity whose value is being continually enhanced through its discoverability, aggregation, use and interpretation in addressing major research challenges.

The continually evolving role played by ANDS is central to enhancing the value of data and ensuring its availability.

Part of that role has been in enhancing the capacity of institutions, themselves the largest investors in research infrastructure, to maximise the value of their own collections and to participate effectively in some of our biggest national challenges.

That is why continuing investment in research data infrastructure is of the utmost consequence to the nation.

UWA's Prof Sydney Hall receives CODATA Prize 2014

(continued from page 9)

The CIF standard has been adopted by journals publishing crystal structure results. It also forms the essential core for data in prominent databases such as the Cambridge Structural Database (more than half a million entries, growing by more than 40,000 each year), and the Protein Data Bank (more than 100,000 entries, growing by about 10,000 per year).

These databases are widely used by researchers from many science disciplines.

In addition, the STAR File approach is being applied to such diverse applications as botanical taxonomy, quantum chemistry, chemical informatics and biographical databases.

Hall received the CODATA Prize at SciDataCon 2014, the International Conference on Data Sharing and Integration for Global Sustainability, in New Delhi last November.

The CODATA Prize is a major biennial award that acknowledges outstanding achievement in the world of scientific and technical data.

Costs and benefits of data provision

Greg Laughlin (ANDS) and John Houghton (VISES)

What is the value of data being used in public research? The new Houghton-Gruen Report suggests it could be up to \$6 billion a year; making this data available for sharing and reuse, via data infrastructure and appropriate policy settings, would result in enormous benefits for Australia. But if it is not available for use, this value is unrealised.

In 2011, ANDS commissioned the Victoria Institute of Strategic Economic Studies to examine the costs and benefits of public sector organisations making their Public Sector Information (PSI) data freely available. The study was undertaken by Prof John Houghton, a prominent economist and researcher in the 'open access' field. It involved a number of public sector agencies, each suggesting different, but positive, benefit:cost ratios. Overall, the study demonstrated that the benefits of making PSI data freely available far outweigh any costs, by some considerable margins.

In late 2014, ANDS commissioned a follow-up report to estimate the total public spend on research, the value of the data created and analysed during the research process, and evaluate the benefits of curating and openly sharing public research data.

The new report, entitled *Open Research Data*, is now available: ands.org.au//resource/open-research-data.html.

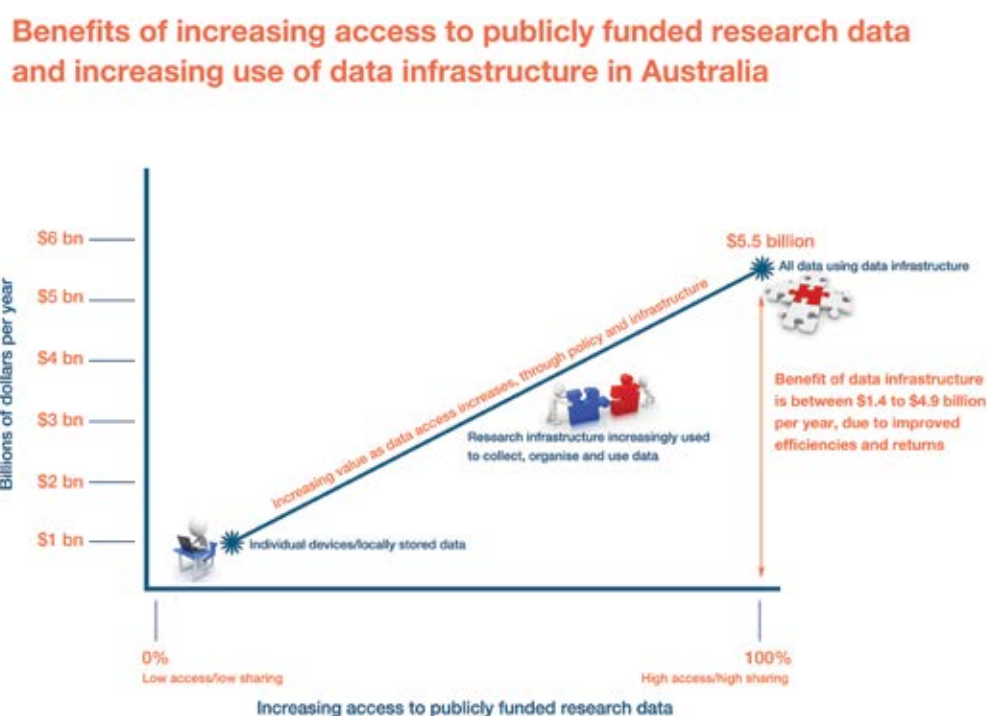
The main findings are really significant: the total spend on public research is in the order of \$13 billion annually, some \$9 billion

of which comes from the Commonwealth Government. Using **conservative** methods, the authors estimate that the value of data used in public research is between \$1.9 billion and \$6 billion a year (the range reflects different methods: labour cost share of Commonwealth funding versus total public research spending). Not all of these data are Australian; the numbers are an estimate of the total value of data used in Australian public research. It is believed that only 10 per cent or 20 per cent of the data created in Australia's public research is currently made publicly-available.

The report also shows the value-adding power of data infrastructure in Australia, based on extrapolations from UK repository studies. The report suggests that data infrastructure can add \$1.4 billion to \$4.9 billion a year. This value incorporates time saved for data centre users, reinvestment of that time in more research, and returns from the reuse of the data by researchers who could neither create nor obtain the data themselves.

Overall the report shows that a relatively small investment in a combination of data policy and infrastructure provides a significant increase in value to Australian innovation, research, and the broader economy.

The figure below illustrates the major implications of the new report.



The vertical axis represents the billions of dollars gained and the horizontal axis represents how much research data is curated and publicly available. Infrastructure and policy, well used, can unlock up to \$4.9 billion per year as data becomes more usable and used.

Forthcoming events

Seminar

Take Control of Your Data Assets - a Practical

Introduction to Licensing Data for Research

When: Thursday, 26 March, 9.30am-12.30pm

Where: University of Adelaide

Roundtable

ORCID Implementers Roundtable with

Dr Laure Haak (ORCID)

When: Tuesday, 14 April, 9.30am-4.30pm

Where: University House, ANU Canberra

Webinar

Unveiling the New Research Data Australia

When: Monday, 20 April, 12.30-1.30pm AEST

Seminar

Tools & Approaches for Managing Environmental Data

Throughout the Data Life Cycle

When: Thursday, 23 April, 9.00am-2.00pm

Where: JG Crawford Building#132, ANU Canberra

Webinar

Changing Paradigms for Data Management,

Publication and Sharing

Presenter: Prof William Michener, DataOne

When: Thursday, 23 April, 2.30-3.30pm AEST

ALIA Presentation

Ready, Set Go! for Research Data at Your Library

When: Tuesday, 28 April, 5.30-6.30pm

Where: State Library of WA

Workshop

Back to Basics: a "Research Data Starter" for WA

When: Wednesday, 29 April, 9.30am-12.30pm

Where: Technology Park Function Centre, Bentley, Perth

Webinar

Cost and Benefits of Data Provision

Presenter: Prof John Houghton, Victorian Institute of Strategic Economic Studies

When: Wednesday, 29 April, 12.30-1.30pm AEST

Workshop

Hot Topics - What's Trending in RDM in WA?

When: Wednesday, 29 April, 1.30-4.30pm

Where: Technology Park Function Centre, Bentley, Perth

Seminar

Take Control of Your Data Assets: a Practical

Introduction to Licensing Data for Research

When: Thursday, 21 May, 9.30am-12.30pm

Where: Technology Park Function Centre, Bentley, Perth

See the ANDS Events webpage (ands.org.au/events)

for further information and to register for these and other events.