share

the newsletter of the
Australian National Data Service

OCTOBER 2015
ISSUE 23

MANAGING CONNECTING DISCOVERING REUSING
RESEARCH DATA

Making the case for data reuse

Opening up research aids new discoveries - and boosts the economy

Emily Petroff

In May 2014, a ghost that haunted radio telescope data materialised for the first time. A fast, powerful burst of radio waves which had travelled billions of light years through space was picked up 'live' by the Parkes Radio Telescope—otherwise known as The Dish—in central New South Wales.

Lasting only milliseconds, astrophysicists think 'fast radio bursts' (FRBs) may signal extreme events in our universe such as the collapse of a neutron star to form a black hole.

"These emissions encode information about what they have travelled through to reach us—about all the matter between where they occurred and here," says Emily Petroff, the Swinburne University PhD student who witnessed the new burst.

FRBs were first revealed in 2007 by American astronomers combing archival data from Parkes for unrelated information, though none had ever been intercepted live before. Emily and her team were the first to do so. The phenomenon lasted only a millisecond – but in that time may have given off as much energy as the Sun does in a whole day.

Yet no-one would know about them without access to the huge amounts of radio data accumulated by The Dish being opened up to international researchers.

New discoveries about our universe are always exciting. What is not yet known, however, is the exact reason why the FRBs happen, and how they are caused. These are questions still being analysed by astrophysicists – though Emily Petroff thinks it is only a matter of time before further bursts are recorded to help answer them. "We've set the trap," she says. "Now we just have to wait for another burst to fall into it."

Continued on page 2

INSIDE ISSUE 23

ISSN 1836-9731

Executive Director's report

Ross Wilkinson, ANDS

Big data! The tsunami of data is upon us! We hear a great deal about the complexity of the new data environment and the infrastructure that supports it, but in this issue of *Share* we focus on the Champions of Research Data – those researchers who have embraced this new environment, using data from a wide variety of sources, to make new discoveries.

Some areas of research, such as the social sciences, have a rich history of using shared data resources to conduct some of their strongest research – the Household, Income and Labour Dynamics in Australia (HILDA) Survey and The Australian Survey of Social Attitudes (AuSSA) are both widely cited in the literature. In other areas, shared data is becoming an absolute necessity, such as with global climate models which run into petabytes and represent vast amounts of High Performance Computing.

We highlight areas where data sharing has not been routine, but where the practice of conducting open science has had many benefits. Some are obvious, but the ability to attract great students internationally has occurred in both chemistry and marine biology, not to mention tackling globally significant challenges such as malaria eradication, and understanding the ocean's response to climate change.

Dr Ross Wilkinson

The word 'globally' is key here. Australia is assembling terrific data assets, and great data infrastructure, but we need our Champions of Research Data to lead Australia in international collaboration over data – because neither data nor research challenges respect borders! So read through these pages, and see what's happening here and around the world in data with oceans of Argo floats, and data coming in from the Universe, and see how our Champions are tackling the challenges.

(continued from page 1)

Economic value of open research

The Parkes Radio Telescope is just another example of open or shared data being reused to come up with fascinating new research – sometimes completely unrelated to the aims of the original study.

A 2014 report commissioned by ANDS, *Open Research Data*, estimated that data curation and sharing is worth up to $5.5 billion annually to Australia. Of that figure, between $1.4 billion and $4.9 billion is yet to be realised, the authors calculated.

The report shows there is much to be gained for Australia by increasing access to and sharing data, through investing in infrastructure and framing the right policies to encourage such activity.

It also makes a strong case that a relatively small investment in data policy and infrastructure can lead to a significant increase in value to Australian innovation, research, and the broader economy.

"Research funders can realise both economic and scientific benefits from open research data, and there are a growing number of funders mandating open data," say the report authors. "Research institutions benefit from the enhanced visibility ... that open data brings and from a reduced risk of inadvertently playing host to scientific fraud."

In 2008 the Australian Bureau of Statistics (ABS) began allowing unfettered use of its data, rather than selling access. A cost-benefit analysis of the provision of this ABS data by Professor John Houghton (also one of the authors of the *Open Research Data* report), showed that, while the Bureau lost the revenue it had previously made from selling its data, there were substantial savings in no longer needing to maintain a 'shopfront' for sales and related inquires. And the benefits to the Australian economy arising from the ABS making data freely available were likely to have outweighed the costs more than five times over, Houghton suggested, because the nation put that information to wider use. If that analysis is correct, the Australian Government would have gained substantially more from the tax revenue earned on the additional economic output than it lost on fee income.

Research funders can realise both economic and scientific benefits from open research data, and there are a growing number of funders mandating open data

Making research data freely available for reuse is a win-win for policy makers, the research community and the taxpayers that fund them.

Based on an original story by Science in Public.

The Open Research Data report can be viewed in full via the 'Quick Links' section of the ANDS website.

Ben Derrick in action at Falls Creek

Reusing a mountain of data

National Park uses six decades of monitoring to support local business, safety and the environment

Victoria's Alpine National Park is the largest and one of the most magnificent in the state, covering a full 6,474 square kilometres of the Great Dividing Range. Stretching from the small Victorian town of Mansfield towards the New South Wales border, it is also home to a number of popular snowfields such as Mt Bogong and Falls Creek.

The high numbers of visitors to the park provide a significant boost to the region's economy, benefitting local businesses and their employees. But to continue wooing the tourists, this important economic and environmental resource must be managed carefully and sustainably throughout the whole year.

"We face a number of land-management challenges where we have to balance the needs of the business, safety and the environment," says Ben Derrick, Director of Economic Development and Land Management at Falls Creek Alpine Resort.

His job includes harmonising the demands of tourists with the need to protect the environment for future generations. Working in one of Australia's least forgiving yet most stunningly beautiful places, it's a role Ben is passionate about.

"My job is fantastic," says Ben, who was previously a semi-professional cross-country skier. "Not only is the work stimulating and rewarding, but I also get to work in an environment that I love."

Long-term monitoring sites

Local managers like Ben are in luck: six decades of scientific monitoring effort has been invested in capturing how Victorian alpine ecosystems work and how they respond to different kinds of disturbance, such as fire, grazing, development, invasive species and climate change.

The network of long-term alpine ecosystem monitoring sites, some of which were established in the 1940s, and which are now supported through the Long Term Ecological Research Network (LTERN), is delivering data to help underpin understanding of the area and its needs.

The findings are having a major influence on decision-making for the ongoing sustainable management of Alpine National Park. Moreover, they are also greatly valued by local businesses such as Falls Creek Alpine Resort.

"Having robust science on which to base our weed or fire management programs means we have much greater confidence that the effort we put in will deliver the desired outcome in an efficient way," says Ben.

Detailed knowledge of the fundamentals of ecosystem structure and function via decades of persistent ecological monitoring means those overseeing the park can more efficiently and effectively manage the land for sustainable use.

It is a great story of data being reused to balance the needs of business, safety and the environment. It is also important to remember that 60 years is a relatively short time in the Alps, and there are still many things we don't fully understand.

Long-term monitoring is vital for increasing our knowledge, anticipating change and managing the alpine environment sustainably. Who knows what else the data will tell us over the next 60 years?

With thanks to the Terrestrial Ecosystem Research Network (TERN). Based on an original story published in the TERN newsletter.

Data sharing out of the blue

The marine economy is boosted by data from an army of robotic floats

In Greek mythology, the *Argo* was a ship sailed by Jason and his Argonauts in their quest for the legendary Golden Fleece. Today, marine industries and modern-day sailors on a different kind of quest also make use of an Argo. For almost 15 years, the Argo international ocean-monitoring project [http://www.argo.ucsd.edu/index.html] has collected and shared data for climate and oceanography research through sensor-equipped floats.

Australian researchers play a key role in this latter-day Argo, jointly led by the national science agency CSIRO and the University of California and involving 31 countries. Dr Peter Oke, a CSIRO Ocean Modelling and Data Assimilation Research Scientist for Australia's own ocean forecasting system BLUElink, says Argo has changed the way researchers do business, encouraging data reuse and spawning new systems like BLUElink.

The Argo community has really led the way in creating a data-sharing culture. By making data access free and open, it's breaking down the silos once set up to collect and protect observations

"The Argo community has really led the way in creating a data-sharing culture. By making data access free and open, it's breaking down the silos once set up to collect and protect observations," he says.

BLUElink is an ocean forecasting system built on Argo data and the Ocean Forecasting Australia Model. CSIRO, the Bureau of Meteorology and the Royal Australian Navy collectively run the project, established in 2001. Forecasts are based on data including temperature, salinity, sea levels and currents, all measured in real-time at different locations and depths by the autonomous Argo floats.

In 2014, Oke and his colleagues used BLUElink to provide intelligence in the search for Malaysian Airlines Flight MH370, which vanished with 239 passengers on board.

"We dropped all tools in our support for the search. We made educated guesses about the splash points and used BLUElink to look at where the debris was most likely to go," says Oke.

Then, in July this year, a wing part washed up on Réunion Island, east of Madagascar. "We did historical runs to backtrack with BLUElink to see if the debris could be from the plane. Argo data was a key component."

The current locations of over 3800 Argo floats appear on a map like confetti across the oceans. Each float's sensors collect temperature and salinity data profiles at depths of up to 2000m. Every 10 days the floats come to the surface to relay data to satellites. More than 10,000 profiles per month provide oceanographers, climate scientists and others with comprehensive subsurface ocean data, accessible via the IMOS (Integrated Marine Observing System) web portal [http://imos.org.au/]. Since 1998, Argo data has generated dynamic maps of ocean currents and resulted in over 2000 scientific research papers.

As for BLUElink, its users range from marine industries such as shipping companies to individuals like surfers and sailors. Each year, using forecasts from BLUElink, sailors in the famous Sydney to Hobart yacht race are briefed on the conditions they'll encounter on some of the world's roughest seas.

Argo data improve safety for oil and gas workers and help analyse risks of oil spills on sensitive coastlines. Data also inform decisions about fishing area boundaries and catch limits.

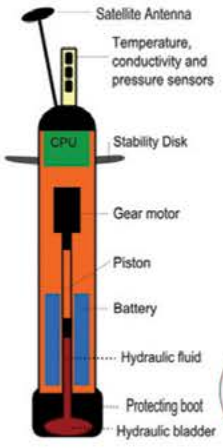
Oke is particularly excited that the field of operational oceanography, which aims to make ocean monitoring and prediction routine. He sees improved ocean forecasts resulting from the Global Ocean Data Assimilation Experiment (GODAE), [https://www.godae.org/], which is exploring how BLUElink can be used more efficiently. Data from sensors on marine gliders closer to shore will also be integrated with Argo data to create new coastal models.

"One day, thanks to Argo, we'll see ocean forecasts as reliable as the weather forecasts that we check-in on every day," Oke says.

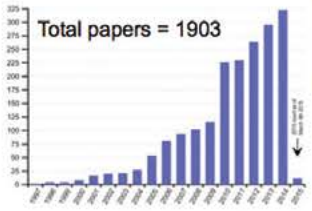
Story provided by Refraction Media.

Argo: global coverage in real time

Ocean Climate Dynamics Program
CSIRO Ocean & Atmosphere Flagship, Hobart, Australia



Argo array approaching 4000 floats!

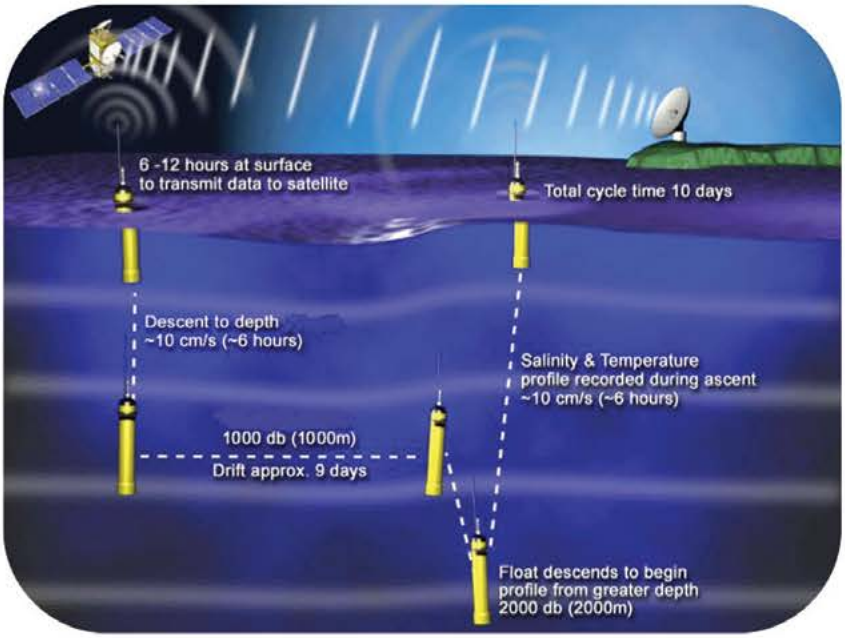


3843 Floats

ARGENTINA (4) · AUSTRALIA (371) · BRAZIL (3) · BULGARIA (3) · CANADA (51) · CHINA (204) · ECUADOR (3) · EUROPEAN UNION (5) · FINLAND (5) · FRANCE (287) · GABON (1) · GERMANY (129) · GREECE (5) · INDIA (114) · IRELAND (6) · ITALY (45) · JAPAN (193) · KENYA (1) · SOUTH KOREA (77) · LEBANON (0) · MAURITIUS (7) · MEXICO (3) · NETHERLANDS (16) · NEW ZEALAND (7) · NORWAY (9) · SOUTH AFRICA (1) · SPAIN (20) · SRI LANKA (0) · TURKEY (4) · UNITED KINGDOM (153) · UNITED STATES (2 116)

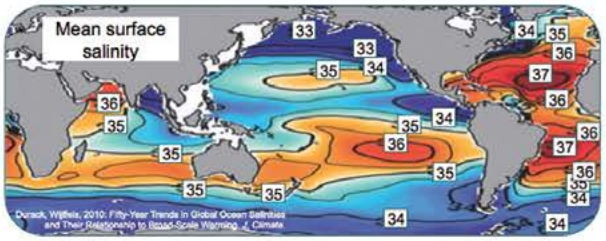
January 2015

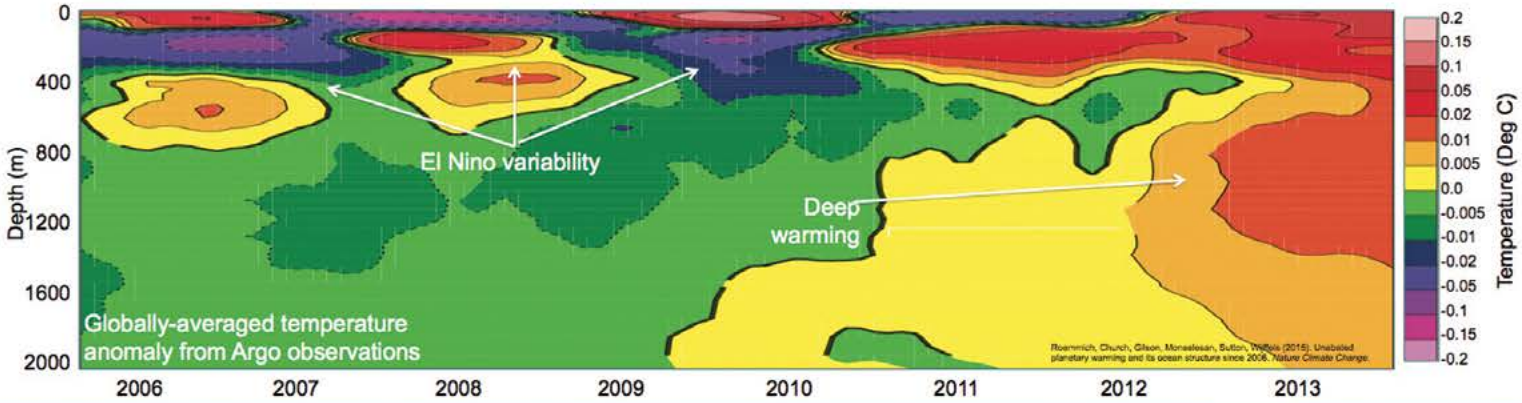
jcommops



What is Argo data useful for?

Ocean currents
Operational oceanography
Planetary energy balance
Mesoscale eddies
Seasonal prediction
Sea-level change
Air-sea interactions
El Nino
Change detection
Ocean oxygen
Ocean mixing
Ocean heat content
Interannual variability
Freshwater budget
Ocean forecasting
Model evaluation
Weather forecasting
Water mass changes





Credit: Argo Project

One small step for open data...

NASA makes its research accessible to the world

NASA has a plan. Not one, in this case, about spaceships and astronauts, but something far more 'down to earth': open data.

The organisation's *Plan for Increasing Access to the Results of Scientific Research* was first published in late 2014, laying out NASA's commitment to open up its datasets for international reuse. Full implementation of the plan is set to be in place from October 2015.

The plan aims, in NASA's words, to "ensure public access to publications and digital data sets arising from NASA research, development, and technology programs".

Done properly, opening up complex data sets for public analysis and reuse can lead to new and exciting discoveries, sometimes by those with nothing more than a keen amateur interest (or perhaps obsession) with the topic.

NASA is fully aware of this potential. It says it wants to support researchers to make new findings based on its data, not just in the US but around the globe. As if to prove the point, NASA's Data Stories website [https://open.nasa.gov/data-stories] highlights a number of case studies of people reusing its datasets in original applications, such as a 'Solar System Simulator' created by Canadian website developer Martin Vezina.

NASA also knows it needs to show commitment to scientific integrity and the accuracy of its research data and wants to encourage others to do the same. So by publishing its own datasets, NASA's team are setting a benchmark for researchers hoping to grab a slice of the organisation's annual research investment – a whopping US$3 billion. A condition of funding those research contracts, outlined in the 2014 document, is that researchers must develop their own data management plans describing how they will provide access to their scientific data in digital format. One small step for open data, one giant leap for new scientific discovery?

NASA data is being opened up to researchers around the globe
(Credit: NASA Creative Commons)

How public data is being reused: The Australian Survey of Social Attitudes

The Australian Survey of Social Attitudes (AuSSA) is the main source of data for the scientific study of the social attitudes, beliefs and opinions of the nation.

It measures how those attitudes change over time as well as how they compare with other societies, which helps researchers better understand how Australians think and feel about their lives. Similar surveys are run in other countries, meaning data from AuSSA also allows us to compare Australia with countries all over the world.

Access to the AuSSA data has allowed independent researchers to explore changes in social attitudes in Australia over time. For example, Andrew Norton (now at the Grattan Institute in Melbourne) has analysed AuSSA to examine changes in attitudes towards same sex relationships between 1984 and 2009, noting the major shifts in favour of same sex relationships during that period.

AuSSA is often used as a reference point for public policy debate. A number of media articles have been based on its findings, discussing topics as diverse as climate change, the welfare state and the kindness of Australians.

Similarly Australian Policy Online includes 18 different papers making use of AuSSA, including papers on perceptions of democracy, population growth, cultural identity and tax policy.

AuSSA datasets can be accessed via its website at http://aussa.anu.edu.au/data.php.

With thanks to Steve McEachern, Director of the Australian Data Archive at Australian National University.

This plan will "ensure public access to publications and digital data sets arising from NASA research, development, and technology programs"

Coral study leads to new partnerships

Jackie Wolstenholme, a research services librarian at James Cook University, wrote her PhD thesis on species boundaries in corals.

After making her work open and accessible, one small table of data led to interest from two professors at the University of Sydney: Emeritus Professor Liam Burke, a physiologist, and Professor Maria Byrne, a marine biologist. Together with Japanese researcher Dr Yoko Nozawa from Academia Sinica, Taiwan, they have teamed up to study new perspectives on the timing of coral spawning. Jackie told *Share* more...

Data on coral habitats led to new research partnerships (Credit: Dr Yoko Nozawa)

What was your original research about, and what was Yoko Nozawa's original research about?

JW: Spawning data has traditionally been collected to study the biology and reproduction of corals. I investigated the fertilization potential of five similar species for my PhD. Yoko has separately examined the influence of temperature on coral reproduction.

So when exactly does mass spawning occur?

JW: That is something that has eluded researchers. We know on which month the corals will spawn for many locations. We also have a good idea about the time of night that corals will spawn, even down to the detail of variation between species. What we can't yet predict is exactly which nights the spawning will occur. We know that mass spawning occurs after the full moon, around the time of the third-quarter moon. Although this knowledge narrows it down to one of several nights in the year, it still means researchers may need to wait night-after-night to make sure they 'catch' the coral spawning.

And why is it useful to know when corals will spawn?

JW: Beyond research, there are also commercial implications. A major drawcard for dive tourism is to swim amongst mass spawning corals. I have also worked for a film crew to capture spectacular footage of mass spawning corals.

Along with Yoko Nozawa, you have now teamed up with Professor Liam Burke and Professor Maria Byrne for new research. Tell us more about the project.

JW: The idea for our current project is Liam and Maria's. Liam has published on the relationship between human vision in eyes affected by macular oedema and the lunar cycle. He found that visual acuity (sharpness) is better at the time of the full moon in the eye with macular oedema. On what seemed to be an annual cycle, Liam also found that a decrease in atmospheric pressure correlates with an increase in visual acuity, both in the oedemetous eye and normal eye. Liam was keen to test his findings in another biological system, and Maria suggested mass spawning corals. To test it, Liam needed data on the timing of spawning by individual coral colonies. Fortuitously, colony data is the level of detail recorded by Yoko and me.

And your previous research is a key part of the new project?

JW: Neither Yoko nor I could have imagined the project that we are now part of when we were collecting our coral spawning data. Yoko's data systematically documented the timing of spawning of four species corals, and was the focus of his paper. In my case, timing of spawning was one of many characters I used to assess the potential for corals to breed. My data is therefore from just one small table in my PhD, and the resulting journal article.

Are any other open datasets being used for the project?

JW: We are using climate data to develop our understanding of the timing of coral spawning. I kept a close eye on the weather while doing my coral spawning field work to determine where we could dive each day but only recorded anecdotal notes. Fortunately, the freely available historic climate data from the Australian Institute of Marine Science (AIMS) has provided the information we need. Without this, our current project would not be possible. Similarly, Yoko has obtained freely available data for our project from the Japan Meteorological Agency.

It sounds like a good advert for open research data.

JW: Yes. Importantly, the coral spawning data collected by myself and Yoko had been managed in a way that enabled it be retrieved. In my case, this was 15 years after the data was collected. Improving the discoverability of coral spawning data could facilitate further projects that, at this point, cannot be imagined by the researchers collecting the data. The climatic data we used are also great examples of how carefully curated data can enable projects that were not anticipated by the data collectors or data custodians.

Report: Open Research Data Showcase

19 June 2015, Hyatt Hotel Canberra

Hon. Senator Scott Ryan: "Australia's research data is a significant national resource"

Katrina Trewin and Kim Heckenberg with Western Sydney University's Open Research Data Showcase display

Shortly after the previous edition of *Share* went to press, ANDS hosted the Open Research Data Showcase in Canberra to highlight the work of the ANDS Major Open Data Collections (MODC) and Open Data Collections (ODC) programs.

Attended by 100 invited guests, the event welcomed representatives from 33 research institutions, as well as the Department of Education and Training, the Australian Research Council (ARC) and other NCRIS programs.

The event was opened by Hon. Senator Scott Ryan, then Parliamentary Secretary to the Minister for Education, who used his speech to flag the importance of the NCRIS program and the Australian research community. "This event shows how opening data collections for wider use has a multiplying effect, as these collections support researchers to increase and improve research, producing better quality data in higher volumes and often faster," Senator Ryan told the audience. "Australia's research data is a significant national resource."

Australia's Chief Scientist Professor Ian Chubb spoke about the importance of science and ICT to the pursuit of research. He went on to pose a number of open questions about the future of research data. (See box on opposite page)

The final speaker was Professor Geoffrey Boulton, Regius Professor of Geology Emeritus at the University of Edinburgh, and principal author of the Royal Society's influential report *Science as an Open Enterprise*. An engaging presentation provided Professor Boulton's perspective on the importance of open data as a glaciologist and the benefits of this approach to the wider scientific community.

An afternoon panel session was facilitated by ANDS Executive Director Dr Ross Wilkinson. Attendees were encouraged to share their views on the value of open data, how it can be further supported, and the hurdles still to overcome. Despite some (welcome) variance of opinion, there was broad agreement in many areas, such as the importance of collaboration and the different relationships required to facilitate open and shared data, particularly internationally.

This event shows how opening data collections for wider use has a multiplying effect, as these collections support researchers to increase and improve research, producing better quality data in higher volumes and often faster

How are we going to build smart cities without the data to understand how traffic flows?

How are we going to get mining projects off the ground without the data to maximise their efficiency?

How are we going to manage our environment – the sustainable use of freshwater or protection of our marine resources – without the data to understand and predict how these systems work, and to monitor how our actions are having an impact, or whether they are having an impact?

How are we going to increase farm yields without a wealth of knowledge from bio-sensors, gene sequencing and climate models?

How are we actually going to open up the North, in the way that is now proposed, without some understanding of the differences between growing legumes and grains down here versus up there?

Data underpins our aspirations for research – and so it underpins our aspirations for Australia.

Australia's Chief Scientist Professor Ian Chubb, speaking at the Open Research Data Showcase

Australia's Chief Scientist Professor Ian Chubb shares a light moment with Senator Ryan

Dr Craig Johnson, Head of the Ecology & Biodiversity Centre at the Institute for Marine & Antarctic Studies (IMAS) [left] speaks to Professor Peter Rathjen, Vice-Chancellor of the University of Tasmania

Open Research Data posters

The display of 35 posters included examples drawn from projects right across Australia, reinforcing the diversity of research being undertaken and the collaborative efforts across NCRIS programs.

As Senator Ryan stated in his opening speech: "What these examples, together with today's event, show clearly, is the importance of supporting research data at the institutional level as well as at the national level."

All Posters displayed at the showcase and the interviews are available on the ANDS website at http://ands.org.au/partner/open2015poster.

Six video interviews were also commissioned in cooperation with individual projects, which were shown during the day. They are now available on the ANDS YouTube channel at www.youtube.com/andsdata.

Guest speaker Professor Geoffrey Boulton, who made the trip over from the UK, with ANDS' Partnerships Programs Manager Angeletta Leggio

Secure data is a SURE thing

Teams reuse survey data via remote research lab

A resource providing data on the health, social and environmental factors of more than 60,000 NSW residents is proving rich information for researchers, who are accessing the data through the Sax Institute's remote, secure research laboratory to embark on wide-ranging investigations.

Three research teams are drawing on the Social, Economic and Environmental (SEEF) project, a sub-study of the Sax Institute's 45 and Up Study, Australia's largest long-running population cohort study into healthy ageing.

"Through providing a centralised, secure environment through which researchers can collaborate, SURE facilitates greater secondary use of data in research, as demonstrated by the SEEF project"

SURE Program Manager Ms Jo Khoo

The data is accessible through the Sax Institute's Secure Unified Research Environment (SURE), a purpose-built virtual research laboratory that enables researchers to access sensitive data without having to store information in their own computing environments.

SURE was established with funding from NCRIS as part of the Population Health Research Network. It brings researchers together from across Australia and the world to collaborate on large-scale projects tackling major health and social issues such as population ageing, diabetes and mental health.

Associate Professor Geoff Morgan, from the University of Sydney, is a member of the SEEF Environment/Spatial Team, which is developing a Geographic Information System framework for the 45 and Up Study and SEEF project participants for use in environment and health research.

The SEEF Environment/Spatial Team is drawing on the SEEF database to investigate physical activity and its relationship to the 'walkability' of people's neighbourhoods, as well as to conduct a study into the association between exposure to air pollution and health.

Associate Professor Morgan said the SURE facility enabled data from the SEEF project – in which subjects completed a 78-item questionnaire covering factors ranging from their attendance at medical practices to home ownership and their sexuality – to be linked to other information such as mortality or hospitalisation data.

Associate Professor Geoff Morgan uses Sax Institute's SURE laboratory to access detailed health data

"The SURE environment has the ability to host very cost effective cohort study information for people who are followed up over many years by linking the 45 and Up Study data with routinely collected health and environment data," he said.

The research team had also been able to obtain data at a much higher spatial resolution than previously, meaning it was possible to narrow down the impact of environmental factors like air pollution on health to much more precise areas – such as residential addresses – while ensuring subjects' confidentiality.

"There are processes in place by which we can access that data and yet ensure it is secure and confidential," said Associate Professor Morgan. "The residential data is not provided in a way in which it can be linked directly to an individual participant or their health data."

The resulting findings would be more precise, improving the validity of the research, he added.

Another study, led by Professor Mark Harris from the Centre for Primary Health Care and Equity at the University of New South Wales, is also accessing the SEEF dataset via the SURE facility. The project will explore the social, economic and environmental factors associated with preventive care, including how such factors impact on people's access to preventive healthcare and health checks, multidisciplinary care for chronic diseases and allied health services.

A third study by a team from the Australian National University is using SEEF data and linked health data to try to untangle the relationship between mental health, antidepressant use and weight gain.

Story provided by Megan Howe, Sax Institute

Data dungeons and dragons

Aerial surveys monitor kelp cover off Tasmanian coast

Photo courtesy of Ivan Hinojosa, IMAS. All rights reserved.

Unearthed from the dungeon-like depths of a government building in Hobart, photographic data from aerial surveys is helping to protect the underwater forest home of Tasmania's dragons. This dragons' tale is a data reuse success story set against the threat of a warming world.

Giant kelp *(Mycocystis pyrifera)* grows in magnificent undersea forests, where it thrives in nutrient-rich, cool waters. Kelp forests can be 30m high and reach the sea surface to sunbake as a floating canopy 40m across, clearly visible from the air. Giant kelp is the foundation species for its ecological community, providing habitat to weedy sea dragons, big-bellied seahorses, abalone, sponges, corals and myriad other marine species.

These richly biodiverse habitats were thought to be in decline off Tasmania's east coast but actual data was lacking. Professor Craig Johnson and Dr Piers Dunstan, researchers from the University of Tasmania's Institute for Marine and Antarctic Studies (IMAS), found the raw data they needed in archives of aerial survey photos going back to the 1940s. It was then that state governments began using ex-WW2 aircraft and photographic reconnaissance equipment to map coastlines, mostly for planning. Fellow IMAS researcher Dr Neville Barrett used the same imagery in the late 1990s to undertake the first seabed habitat mapping of Tasmanian waters for marine protected area planning.

"It's classic data reuse. Those aerial surveys were done to map Tasmanian land-based features but we found we could also use the survey photos to study marine habitat," Barrett says. "Craig and Piers dug up all the aerial photos of the east coast they could find and estimated the percentage of kelp cover in every bay, year by year. They identified a 90% decline in kelp cover between 1945 and 2000."

Warmer oceans

Why the decline? Climate change is thought to be the main culprit. Oceanographic shifts first seen off Tasmania's east coast in the 1970s coincided with IMAS' observations of reduced kelp forests. During the period studied, noticeable strengthening of the East Australian Current created warmer ocean temperatures and reduced nutrient load, conditions which are unfavourable to kelp growth.

Kelp forests are among the most dynamic and productive ecosystems on Earth, yet are unknown to many Australians. Most people could name iconic terrestrial forests like the Daintree in Queensland but would struggle to name a comparable marine equivalent. In 2012, thanks to the work of the IMAS researchers and others, the giant kelp marine forests of southeast Australia became the first marine ecological community to be protected with an endangered listing under Australia's national environmental laws.

New types of data, including image data from robotic submersibles (autonomous underwater vehicles or AUVs) and satellite data from advanced sensors, are aiding in conserving and managing marine environments. For example, a recent University of Tasmania project measured kelp beds from hyperspectral satellite data exploiting the species' unique spectral signature.

Studies on restoring kelp communities are underway and, in a twist to the dragons' tale, the original archive of irreplaceable aerial photos that helped protect their home has now been digitised. The data is now reusable by researchers anywhere.

"Wonderful outcomes like this become more and more possible every year as ANDS and other organisations enable access to more data and share it more readily," says Dr Barrett.

But, in the end, if we're to preserve these iconic forests, we need to tackle the root cause, which is the changing ocean conditions. How the forests will ultimately fare in this warming world is a tale whose next chapter is yet to be written.

Story provided by Refraction Media.

Chair's report

Ron Sandland, ANDS' Steering Committee Chair

It is pleasing to report that the overwhelming arguments in favour of the continuation of NCRIS funding have borne fruit with the Government's decision to continue funding while awaiting the findings of the Clark Review into research infrastructure. ANDS can now plan its activities over the next two years with a great deal more confidence.

A significant event that has taken place since I last wrote was the Open Research Data Showcase. It was a spectacular coming together of users of research data, showing just what can be achieved when researchers provide open access to facilitate its reuse and grow important collaborations. We saw some great examples (some of which are highlighted elsewhere in this edition of *Share*) which support the thesis of the Houghton–Gruen report *Open Research Data*, namely that the potential benefits from appropriate curation of open research data run into the billions.

Indeed ANDS is providing the infrastructure to facilitate this occurring. However, while this contribution was acknowledged by those attending the Showcase, what was even more gratifying was the level of buzz that characterised the discussions that took place. One of the unwritten laws of effective open data sharing and reuse is that there needs to be a fundamental change in research culture to underpin the open data revolution. And it was evident in the Showcase participants. It is no longer good enough for researchers to say 'this is my data and I'm not sharing it with anyone in case they find something that I want to find'. In fact opening one's research data can lead to new collaborations and greater research impact (whichever way that is estimated).

A feature of the Showcase was a substantial number of posters and high-quality videos illustrating the value of open research data. A particularly good video was from the Tropical Data Hub at James Cook University where a number of examples were displayed. One such example was VecNet, bringing together malaria data and modelling tools, as well as guidance tools, which could potentially multiply many-fold the value of the research collections. And the impact of openness is real in terms of delivering new insights into the prevention and treatment of malaria.

Indeed, Shakespeare could have been talking about open data in *The Merchant Venice* when he wrote of mercy: "It is twice blest; it blesseth him that gives, and him that takes".

The potential impact of unlocking data, as has been done in the Tropical Data Hub, is staggering.

Open source approach puts malaria on notice

Collaborative University of Sydney project tackles global killer

According to the *World Malaria Report 2014*, there were an estimated 198 million cases of malaria worldwide in 2013, and an estimated 584,000 deaths. Tragically 437,000 of these were African children who did not reach their fifth birthday.

In response, the World Health Organisation (WHO), which authored the report, put a call out to the global health community for further investment into programs to combat malaria, particularly to target gaps in prevention and treatment. Current treatments exist but they can be expensive, often do not reach those in need, and may not work. For many treatments, parasitic resistance is also building. New drug discovery is therefore critical, compelling researchers to adopt innovative approaches.

Last July, *Share* featured a story, *Opening new ways to treat Malaria*, about an open source drug discovery project called Open Source Malaria (OSM). Through open collaboration, the project extends research by providing mechanisms for data sharing and reuse, all done in real time. Now a year on, OSM's team lead Professor Matthew Todd says the project is making progress.

"We're now working on molecules that are very promising – the current set are able to cure a mouse of a form of malaria, so the molecules can wage their war on the parasite inside an animal, which is a major step forward."

Obviously, defeating malaria is not a trivial venture. Many challenges exist. Significant amongst these is gaining physical samples of molecules for evaluation. Here the open source approach comes into its own. Through innovative thinking and sound course design, an exciting collaboration has emerged with a group of high-achieving chemistry undergraduates at the University of Sydney.

Credit: Open Source Malaria

Alice Williamson, Postdoctoral Research Associate, developed a new component of the 'Special Studies Program' in Chemistry. Along with Professors Adam Bridgeman and Peter Rutledge, Williamson designed a set of experiments in line with OSM's focus on the set of "very promising" molecules to which Todd refers.

Several new molecules were synthesised by the students with the aim of creating new antimalarial compounds. Their project work and experiences were captured in open lab notebooks and videos, and what they discovered was incorporated back into the OSM data pool. The outcomes in terms of drug discovery are still unknown since the compounds are undergoing biological evaluation. If the results are good, these molecules will progress to further biological assessment and hopefully from there to testing in humans.

The OSM consortium's history with non-traditional crowd-sourced collaborations is characteristic of the project. Students from Sydney Grammar School and a group of 40 undergraduate Lawrence University (USA) students, working from their own labs, have also made new molecules and evaluated them for activity against the malaria parasite, *Plasmodium falciparum*.

As Williamson notes, this could scale. "One can imagine lab courses based around current needs in any open source drug discovery and development project, meaning we could, with proper mentorship, bring to bear very large levels of skilled human resource to tackle actual project needs, with global coordination between cohorts."

"It's a win-win situation. Students learn research skills and an appreciation for organic chemistry, whilst tackling a real problem."

Professor Todd is equally optimistic. "It remains a particular goal for the future to engage scientists in the consortium from malaria endemic areas," he says.

"Traditionally this has always been a problem of logistics and money, but the open source approach has the best shot of involving researchers from endemic areas as equal collaborators. After all, because we share everything, there are no project outsiders".

More information means better predictions

CMIP5 modelling climate change and extreme weather events

CMIP5 models are helping researchers better understand cyclones
(Credit: NOAA/NASA Creative Commons)

In the era of 'big data', researchers are reaping the rewards from working with increasingly vast amounts of information about our planet. And datasets don't get much larger than those used for modelling climatic events and simulating the impacts of global warming on the Earth's surface.

The primary tools for modelling the climate are Atmosphere–Ocean General Circulation Models (AOGCMs). To improve the credibility of AOGCMs, the World Climate Research Programme [http://www.wcrp-climate.org] established the Coupled Model Intercomparison Project (CMIP). This facilitates comparison of different models to identify common deficiencies and stimulate investigation into their possible causes.

Being the latest generation, the CMIP5 models are the most valuable resource we have in the field

CMIP5 [http://cmip-pcmdi.llnl.gov/cmip5] is the fifth phase of CMIP and a multi-model framework of unprecedented scale. It incorporates many more simulations than earlier versions, including those based on historical concentrations, experiments for investigating climate sensitivity, and four emission scenarios reflecting differing potential pathways to 2100.

Use of datasets produced by CMIP5 is widespread: several thousand researchers access the CMIP5 datasets via the Earth System Grid website [https://www.earthsystemgrid.org/home.htm], and 28 modelling groups worldwide work on models that input to CMIP5 activity. Over 1000 peer-reviewed papers using the datasets have been published in a range of respected climate journals, for example: *Journal of Climate* (184 papers), *Geophysical Research Letters* (129 papers), and *Climate Dynamics* (122 papers).

In Australia, researchers at The Centre for Australian Weather and Climate Research, a partnership between the Bureau of Meteorology and CSIRO, have employed output from CMIP5 models to further our understanding of the current and future climate in the Pacific region.

The research, undertaken as part of the Pacific-Australia Climate Change Science Adaptation Planning (PACCSAP) program, provides insights into the current and future impacts of climate change on the Pacific and the implications for communities in the region.

The work further reinforces the strong credentials of climate research in Australia, which also boasts centres such as the ARC Centre of Excellence for Climate System Science at the University of New South Wales (see Share issue 21).

One of the research streams of PACCSAP has projected the impact of extreme weather events, such as tropical cyclones, onto the region's future climate. The output from CMIP5 models was key to simulating the conditions for the genesis and behaviour of tropical cyclones.

"Being the latest generation, the CMIP5 models are the most valuable resource we have in the field," says Dr Sally Lavender, Research Scientist at CSIRO's Oceans and Atmosphere division. "The real advantage with CMIP5 is there are more models than the previous generation with a broader set of experiments, and all the models are much better in terms of sophistication. They also tend to be higher resolution and more have sub-daily time fields which, for modelling tropical cyclones, is very important."

Dr Lavender is currently working to extend previous research using CMIP5 models to observe why and where cyclones form, and what determines their tracks. "We're analysing the CMIP5 models to see how well they represent those processes in the real world to produce a selection of models that are good at representing tropical cyclones over the Australian region. We can then use these models to generate more informed projections of tropical cyclones under future climate scenarios."

Research to date shows there is likely to be a reduction in the overall frequency of tropical cyclones in the Australian region; however, the proportion of high intensity cyclones is likely to increase. That needs to be taken into account in future building standards and disaster readiness planning.

Story provided by Refraction Media.

Opening the doorway to cancer data

Genome data portals support scientists in hunt for cancer cure

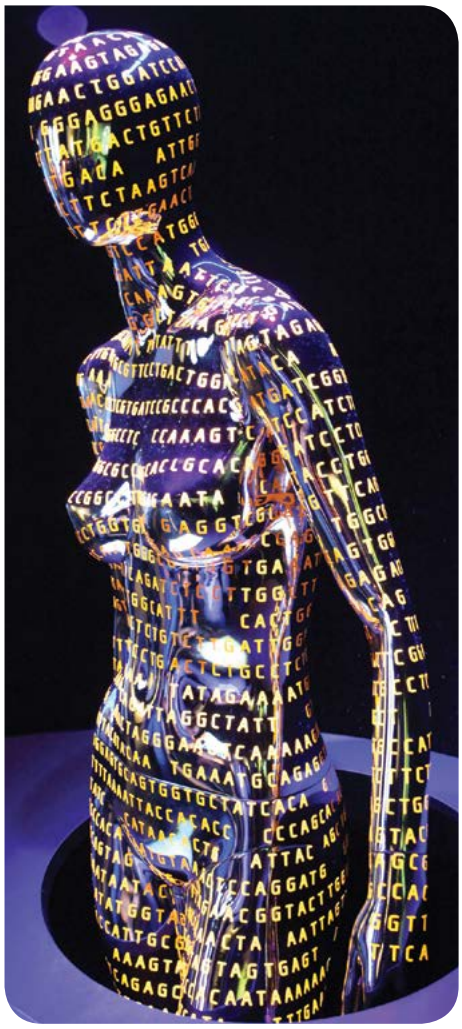
Precision medicine is offering hope to cancer patients. The power of genomics and the masses of data it creates is transforming cancer research and allowing personalised treatments with more proven effects.

Like hundreds of other cancer researchers, Mark Ragan and his team at The University of Queensland's Institute for Molecular Bioscience (IMB) need to design experiments based on data from human and cancer genetics. Using data chips and next generation sequencing they must assemble their genetic data, interpret it to understand what genes their data refer to by comparison with other samples, and then classify patients' cancer into subtypes. If they can't match to an existing subtype, they identify a new one. Ragan says this intensive work requires access to as much genetic data as possible.

"It would literally be impossible without the data reuse that TCGA and other genome research programs offer"

Luckily, there are portals with this type of data. One of the first to start collecting cancer genome data was the The Cancer Genome Atlas (TCGA) [http://cancergenome.nih.gov/]. The initials TCGA also make up the four-letter code of nucleotide bases thymine, cytosine, guanine and adenine that DNA uses to 'write' genetic information.

TCGA was started by the US National Institutes of Health (specifically the National Cancer Institute and the National Human Genome Research Institute) in 2006. Ragan says its initial goal was to generate data from researchers across research institutions on two cancer types. Early success expanded the initial goal to collect and profile more than 10,000 samples from over 20 tumour types. While the sample collection phase



Credit: Greyloch (Flickr / Creative Commons)

ended in 2013, data reuse ensures the data generated from those samples are still being analysed. Over 2700 papers have been published by TCGA data so far, including Australian researchers.

The data portal for the TCGA is "amazing" says Ragan. "It's a really powerful portal that lets you ask questions and interrogate gigantic amounts of cancer genome data, including sequences, survival rates and subtype classifications."

"Just about everything in it is open access, and the raw data, which isn't open access, is made available by applying through research institutions' ethics committees."

A newer initiative inspired by the success of TCGA, the International Cancer Genome Consortium (ICGC), is an international project in which Ragan's colleagues play a part. ICGC is built on the TCGA project, which provides about 60% of the patient data in ICGC's Data Coordination Center. ICGC aims to cover 50 tumour types and currently funds 78 international cancer genome projects like the Australian project at IMB.

"Our research into breast cancer subtypes and survival would literally be impossible without the data reuse that TCGA and other genome research programs offer. We can tell if we've discovered a new cancer subtype or not, or even whether the existing data need reinterpreting," says Ragan.

New treatments

Knowing a patient's cancer subtype allows more tailored, evidence-based treatment, potentially increasing survival rates and quality of life by allowing clinicians to more confidently focus on prescribing the drugs most likely to succeed for a particular patient.

One of the exciting things Ragan and other researchers are finding from the data is that some quite different cancer types have a similar genetic basis. This means drugs to treat one type of cancer, such as breast cancer, could be used for another, such as ovarian cancer.

"Instead of waiting 10 years for a new drug to be developed, patients may be able to be treated straight away with a drug that's already available for another cancer," says Ragan.

That's good news for patients, and it also makes drug development, which can cost hundreds of millions of dollars per drug, more cost-effective. This potentially creates a larger market for a given drug, and makes some drugs financially viable that otherwise wouldn't get to market.

Story provided by Refraction Media.

share

the newsletter of the
Australian National
Data Service

ANDS News

ANDS launches Research Vocabularies Australia

ANDS has launched a new controlled vocabulary service of agreed research terms used by Australian researchers.

Research Vocabularies Australia (RVA) comprises tools and support for creating, managing, storing, publishing and discovering vocabularies for research. The service aims to describe vocabularies commonly used by or relevant to Australian researchers.

Research Vocabularies Australia is now available at https://vocabs.ands.org.au.

For information about how to use the system, see the documentation on the ANDS website or contact services@ands.org.au. A recording of the webinar introducing the service is also available to view online via the ANDS YouTube channel [youtube.com/andsdata].

A controlled vocabulary is a set of agreed terms used to label data concepts, meaning the descriptions used in research are consistent across the research community. This supports the discovery, linking, understanding and reuse of research data.

Events

ANDS runs a full calendar of events for the data research community, including numerous online workshops and webinars. For a full list of the latest events go to:

ands.org.au/events

Subscribe to andsUP

andsUP is a fortnightly e-newsletter including all the latest news and events relevant to Australian research data. Sign up at:

ands.org.au/newsletters

Twitter

Follow ANDS on Twitter for updates, news and events:

twitter.com/andsdata

What we're reading

Journal of Librarianship and Scholarly Communication (Volume 3 – Issue 2)

Special issue of the Journal with a great range of commentary, articles and best practice for researchers and librarians. Read it online at:

http://jlsc-pub.org/10/volume/3/issue/2/

ANDS
AUSTRALIAN NATIONAL DATA SERVICE

ANDS Office
Building F, Level 6, 900 Dandenong Road, Caulfield East, VIC 3145
Telephone: +61 3 9902 0585 Email: contact@ands.org.au

NCRIS National Research Infrastructure for Australia
An Australian Government Initiative

ANDS is supported by the Australian Government through the National Collaborative Research Infrastructure Strategy Program. Monash University leads the partnership with the Australian National University and CSIRO.

ANDS Project Partners

MONASH University
Australian National University
CSIRO

THIS NEWSLETTER IS DESIGNED BY GRIT BRAND + ONLINE gritcomms.com.au