share
the newsletter of the Australian National Data Service
october 2010 : issue 06

Champions of Research Data

Murtho Floodplain, Murtho, South Australia. Photo courtesy of Ms Tanya Doody, CSIRO

Fluid Resources Build a Foundation for CSIRO

By Pollyanna Sutton (Pollyanna is a freelance journalist)

Water resources management is one of the crucial issues for Australia now and into the future, so making data available from one of the country's significant water assessment projects was a good test for CSIRO's newly formed Data Management Service. The Murray-Darling Basin Sustainable Yields (MDBSY) project proved to be a strong testing ground because of the volume and diversity of the data and the prospect of multiple end users. CSIRO's Water for a Healthy Country Flagship led the MDBSY project to assess the current and future water availability in the Murray-Darling Basin. It was the first water resource assessment at this scale in the world taking into account climate change and surface and groundwater interactions.

CSIRO has a long history of research in hydrology. The challenge has been to reuse data outside individual projects in a universal way. For example, one dataset may include 100 years of climate scenarios with estimates of rainfall, and origins in thousands of files that give rainfall at individual locations.

Dr Edward King, a research scientist in marine and atmospheric science at CSIRO worked with the Data Management Service to help build an infrastructure and tools that would allow better management of data. "A lot of the datasets were large and unwieldy because they were organised in a way that would fit the analysis of that particular project, this didn't really allow the structure of the data in space and time," he said. The MDBSY project used a broad array of data from climate control historical data from 1895 to 2006 (rainfall run-off models and observed climate data) to future global warming scenarios. These include future daily climate scenarios across the Murray-Darling Basin, global climate models and global warming scenarios of high, medium and low.

Project Manager for CSIRO's Data Management Service, Dan Miller, said, "The volume of raw data was in terabytes; that meant literally millions of comma-separated value files which were difficult to provide to anyone after the research was conducted."

...continued on page 2

Inside Issue 06

ISSN 1836-9731

The team identified a format that was more useable and built tools to convert the raw data to NetCDF files (Network Common Data Form). Once the data was aggregated with an associated metadata set and ingested into the system, some of the key processes around managing the datasets were developed, including tools to administer, and tools for people to access and discover the information.

CSIRO has built a data access portal and will collaborate with ANDS to publish the datasets to make them more discoverable. "The tools and processes we have developed are generic enough to handle the nuances of this dataset and be reapplied for future datasets," Miller said. The MDBSY project data has raised some salient issues about data sharing for the organisation. Miller explained that because the data originated from various external sources, CSIRO is not the owner but the custodian, so it requires a process of correct approvals for use. "That is one of the reasons we tackled this area first, because it brings out some of the tough questions we have to answer in CSIRO for data sharing," he said.

Oceanographers are in the Vanguard of Data Sharing

By Pollyanna Sutton

Professor Nathan Bindoff. Photo courtesy of Professor Nathan Bindoff

Professor Nathan Bindoff has spent a lot of his life looking for climate change in the global oceans combining oceanographic and atmospheric research, which in his words, has led to a deep involvement in data. As Professor of Physical Oceanography at the University of Tasmania and CSIRO Marine Research Laboratories, Director of the Tasmanian Partnership for Advanced Computing (TPAC) and a Project Leader in the Antarctic Climate and Ecosystems Cooperative Research Centre (ACE CRC), he has been proactive in developing systems that allow sophisticated modelling and comparisons, to predict how climate changes will affect our future.

Oceanographers have been in the vanguard of data sharing, with projects such as the World Ocean Circulation Experiment (WOCE) beginning in the late 1980s. WOCE was a billion dollar experiment funded by countries around the globe that saw 12 different geographic data centres assemble hundreds of datasets from individual researchers. These were combined to form a single coherent searchable global resource that conformed to standards. These data collections were distributed initially on CD and eventually through the internet as it developed in the early '90s.

Over time, the research showed that the state of the ocean was evolving. Oceanographers found that the salty surface of the ocean in the subtropics was evaporating faster; there were changes in ocean temperature and ocean salinity caused by the oceans taking up heat and the changing global patterns of rainfall. His current project with colleagues is a synthesis of simulations of the Future Tasmania Climate exploring how the climate of Tasmania is likely to change from 1961 to 2100. The three-year project has required massive amounts of complex modelling output of both observational data (more than 70 Terabytes), to project how the climate over the island will evolve. Results show that Tasmania will increase in temperature by 2.9 degrees, which is slightly cooler than the projections of the global average temperature change of 3.4; warm spells, and temperature extremes are 2-3 times more common. The east coast of the island will be wetter and the 1 in 100 year rainfall event will become up to 70% more intense with fewer drizzle days; this in turn will affect river flow and flooding. An indication of the impact on agriculture is in the wine industry. In 2000, vineyards were growing Pinot Noir grapes. By 2030 they will be able to grow Shiraz.

Increased access to data is being made possible through the Marine and Climate Data Discovery and Access Project (MACDDAP) co-funded by ANDS and ARCS. The project brings together the research of Integrated Marine Observing System (IMOS), TPAC, CSIRO, the Bureau of Meteorology and Architecta. MACDDAP has already delivered a TPAC digital library with a greater geo-spatial search capability, and a data aggregator service to extract data with various attributes, and combined them into a single user product, and capabilities to deliver standards used by the geospatial community.

Because organisations are agreeing to use common formats in common ways, there is a potential for greater integration. In the past a search across datasets for something as simple as ocean temperature may have returned an ambiguous result because the "synonyms" in the data weren't clear. Indeed, in some of these datasets there can be hundreds of different terms for the same variable, for example sea-surface temperature.

Professor Bindoff predicts that with the right support, in less than a decade, there will be standardisation of functions that will be applied to legacy data and the new datasets that are coming on line from organisations like IMOS, Terrestrial Ecosystem Research Network (TERN), and the Atlas of Living Australia. "Once we get the data to a standard, the datasets become interoperable, and then we will see the new scientific discoveries because we can cross disciplines transparently."

He said that the tasks confronting those working in the Environmental Sciences are getting higher, and there is a demand to deliver results at a faster rate. "Data is expensive to obtain, multiple access means it is used in ways we wouldn't have anticipated, and this drives the real value of data, instead of it being used once or a few times," he said. Because projects like MACDDAP and TPAC are helping scientists to publish their datasets and make them more discoverable, scientists can spend more time on the analysis and synthesising with models to bring them together.

New Data in the Commons

ANDS aims to encourage researchers to make publicly-funded research data part of the Australian Research Data Commons so that it is discoverable and reusable by other researchers. Pioneering the way is Queensland University of Technology (QUT) and Griffith University, which in late 2009, undertook ANDS funded projects to identify, capture and expose research data.

"Involvement in the ANDS project has, in conjunction with the *Australian Code for the Responsible Conduct of Research*, been a stimulus for developing the policy on the management of research data and guidelines that follow that policy," said Martin Borchert, Associate Director, Library Services (Information Resources and Research Support) at QUT.

Research Data Librarians, Craig Milne and Ellen Thompson were tasked with identifying and describing the datasets associated with ARC or NHMRC research activities for the period 2000 to 2009. Interviews were conducted with researchers in relation to 424 research activities and the metadata was described and recorded using the RIF-CS schema.

Martin Borchert believes making data available is a good way to attract additional attention to research publications and that there is evidence that this increases the citation count of published research. It is seen as a useful way to increase QUT's research output and its standing in the research community. "We recognise that one of the greatest costs in conducting research is actually creating the dataset," he said. "When curation and annotation are done retrospectively a great deal of work is required to make datasets shareable and reusable and this can be a major impediment to sharing," said Dr Joseph Young, Manager of High Performance Computing at QUT. Building data sharing into research agreements increases the likelihood that data can be shared because it provides researchers with the opportunity to liaise with the Office of Research, ethics committees and the people involved with contracts and commercial services from the very beginning of a project. A QUT interface has been developed by the High Performance Computing and Research staff in collaboration with the data librarians to allow the data librarians to enter the metadata about research datasets into a database.

At Griffith University the IT and Library group are integrated and services are designed around the research cycle from discovery right through to publications. Malcolm Wolski, Acting Director, Scholarly Information and Research Services says that, "Griffith University, like most universities, has a publications repository but did not have an integrated or enterprise solution to deal with the data situation and that is what we are aiming to build... Participating in this project has made what we need to do internally in terms of policy, architecture, technology and resource budgeting much clearer." Preparation is underway on a policy issues paper covering ethics policy, publishing, open access, IP and Creative Commons. "To facilitate data sharing, institutions need to implement a process that preferably captures the data and metadata automatically from the very beginning of a project," Wolski said.

Of particular significance was the work Griffith University did with the National Library of Australia's Party Infrastructure Project team to pilot the creation of researcher profile records for Griffith researchers for the online database, Trove. The records will also be ingested to Research Data Australia. The NLA project assigns unique identifiers that will allow researchers to build their profiles over multiple projects and across institutions and locations throughout their research life span. "Most systems hold bits of information about projects, publications, data or staff but there are very few systems that collect the relationships between all of those," commented Wolski. Data preservation is the next issue to be addressed at Griffith. Skills need to be developed and policies established as it will not be possible to keep all research data.

Associate Professor Michael Blumenstein, Dean (Research) of the Science, Environment, Engineering and Technology Group said, "Griffith University holds some significant repositories of data in entities such as the Eskitis Institute for Cell and Molecular Therapies and the Queensland Compound Library. It will be hugely beneficial for researchers to share this data nationally and internationally and to collaborate with others." Griffith University is now exploring ways to de-identify some of their sensitive datasets so that they can also be made shareable.

"...one of the greatest costs in conducting research is actually creating the dataset,"

Chair's report – Ron Sandland

I can't let this opportunity pass without paying tribute to Clare McLaughlin for her tireless work in DIISR on behalf of the national eResearch capabilities. ANDS is greatly in her debt for her assistance and strategic advice, often on matters of some delicacy and complexity. Clare has moved to a new policy role in DIISR where I'm sure she will continue to acquit herself with distinction and verve. Added to which she's a great Lyle Lovett fan. Thanks Clare!

Clare has passed the eResearch baton to Cheryl Kut whom I know well from working with her on the AURIN project. Cheryl is extremely capable and I'm sure she will make a very positive contribution to the eResearch portfolio and ANDS in particular.

The links between AURIN and ANDS were not initially as obvious as they have now become. So what is AURIN? AURIN (Australian Urban Research Infrastructure Network) is a $20 million EIF funded initiative to provide research infrastructure for urban and built environment researchers. At present researchers do not have access to critical data which is often held in very diffuse repositories in all three tiers of government, government corporations and industry (as well as in research laboratories).

The aim of AURIN is to provide these researchers with access to appropriate data and models to enable them to address problems which are critical to the sustainability of our urban environments. These data might relate to water and energy usage, transport, climate change for example. They may arise in very different forms, from well-managed collections in the ABS to social networking data and vary from hard to soft. Addressing these problems absolutely requires an ability to integrate hard data and qualitative data from the social sciences.

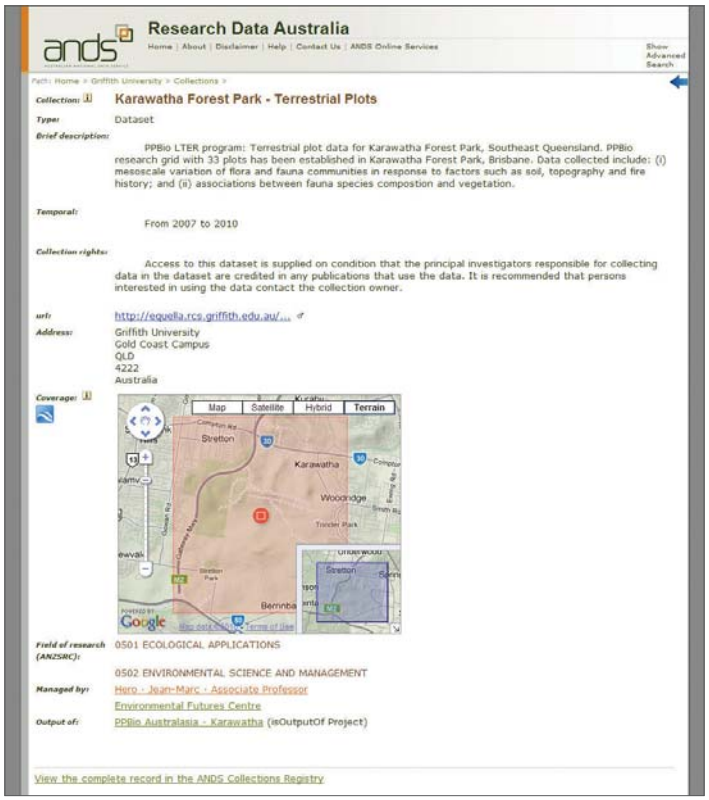
The challenges in AURIN are enormous, not least of which is negotiating access to the data through a maze of privacy, confidentiality and IP considerations. Which data should be prioritised for collection? Where do they currently reside? Can access be via *in situ* data repositories or is a central store required? Ultimately researchers will access AURIN through a portal that points to the collections, provides rich descriptions of the data to aid the search and probably interpretative tools and models.

The overlaps with ANDS are clear. The problems are challenging and of real national significance. It's a tremendous opportunity for the ANDS capabilities to make a difference to the Australian research community.

Executive Director's report – Ross Wilkinson

In the last issue of *Share*, Ron Sandland referred to the hard task of developing research data momentum. A lot of effort has taken place in the last year and a half to develop that momentum since ANDS officially started and it is nice to start seeing the effects. Two of our early engagements were with Griffith University and Queensland University of Technology. They worked together to enhance their internal data management, and also to make some of their researchers' data visible through Research Data Australia (RDA), the ANDS portal onto the Australian Research Data Commons. Just recently, I was at a meeting in Queensland where I was listening to a talk about an international initiative to understand climate impact and adaptation in the tropics. I typed in the name of the speaker in RDA, found him, clicked on his data collections, and was taken to the collection page, which took me to the Griffith data repository and to the PPBio (Program for Planned Biodiversity and Ecosystem Research) data. His name – Associate Professor Jean-Marc Hero (Deputy Director of the Environmental Futures Centre). The capture of this metadata and data is also supported by the Terrestrial Ecosystem Research Network initiative.

It's great to see the hard work of Griffith University and Queensland University of Technology, together with the ANDS team, laying down paths that enable us to start to more effectively explore Australia's research data.



NeAT Project update on the Human Variome Project

The Australian node of the Human Variome Project is a NeAT project, jointly funded by ANDS and ARCS. The aim of this project is to establish a system to collect and make available information on genetic variations associated with human disease and to put this system into operation. The long-term intention is to make it possible to link together genetic variation, changes in the phenotype (the body encoded by that particular genome) and diagnostic results from particular tests associated with actual or possible disease. And of course, all this has to be done in a way that preserves patient confidentiality. The project is making good progress with a focus on establishing formal agreements. By its end in September 2011, it will have deployed the Australian Node database which will receive data and host the default search portal, as well as setting up three or more pilot sites for data collection.

In brief

DMP Online - An Innovative Data Management Planning Tool

The Digital Curation Centre (http://dcc.ac.uk/) has developed an innovative data management planning tool that adapts to the requirements of different research funding bodies. The DCC analysed funders' requirements and developed guidelines for two different versions of a data management plan; the first ('preliminary' version) for use at the grant application stage, and a second ('full' version) which is developed at the early-project stage (if the grant is successful). The user simply needs to go to http://dmponline.hatii.arts.gla.ac.uk/, login (first creating an account if needed), select the relevant funding body and work through a series of screens answering questions written in researcher-friendly language. The resulting plan can then be flexibly exported as PDF or HTML. While the funding bodies built into the plan are all from the UK at present, ANDS has been talking to the DCC about adapting and trialling the tool in Australia.

ESF advocates making data accessible in its new Code of Conduct

Original research data should be stored securely and should remain accessible to colleagues for a substantial period, according to the code of conduct for researchers that was launched recently by the European Science Foundation (ESF) (http://www.esf.org/). The Code cites examples of good practice and poor conduct in science and offers a basis for fostering confidence and integrity when international researchers work together. It is intended to serve as a point of reference for all researchers and to complement national and European regulations in this area. The ESF says that its Code is not intended to replace national codes, but instead to demonstrate what the 30 participating countries have in common. The foundation also believes that it could serve as the basis for a worldwide code of conduct. As well as data storage, the Code also concerns all other aspects of scientific research, such as the respectful observance of research procedures, taking responsibility for the health and safety of participants, transparency in publication and editorial responsibility. The full text of the Code is incorporated in an ESF report entitled *Fostering Research Integrity in Europe*, which is available at: http://www.esf.org/publications

"The Digital Curation Centre has developed an innovative data management planning tool that adapts to the requirements of different research funding bodies."

Korea Institute of Science and Technology visit

L to R: Dr Tae-Jung Kim (Researcher, KISTI), Mr Sun-Tae Kim (Senior Researcher, KISTI), David Groenewegen (ANDS) and Dr Andrew Treloar (ANDS)

Mr Sun-Tae Kim (Senior Researcher) and Dr Tae-Jung Kim (Researcher) from the Korea Institute of Science and Technology Information (KISTI - http://www.kisti.re.kr/english/) visited ANDS on September 14, 2010. They were undertaking a study tour looking at Australian eResearch infrastructure initiatives. The meeting provided an opportunity for the visitors to ask ANDS staff about the approaches being taken in Australia to encourage better data management and re-use. Following on from the visit, ANDS is exploring how best to work with KISTI.

Gazetteer project – Australia's location infrastructure

ANDS and Geoscience Australia (GA) are combining forces to enable cross-disciplinary discovery of research data, with spatial location playing a vital linkage mechanism in this process. ANDS vision for a research data commons will see non-GIS-experts from arts, humanities and science able to enrich their discipline specific data with standardised spatial information. GA has just signed a contract with ANDS to establish a robust national infrastructure which will allow place names to be validated by both individuals and software systems against an Australian gazetteer service in an efficient manner.

This infrastructure will increase the amount and quality of spatially-marked-up research data. It will enable new kinds of research and innovation based on new data linkage and data merging opportunities. The infrastructure aims to unlock significant innovation and productivity and will bring benefits well beyond the research and innovations sector.

ANDS Online Services

The ANDS Online Services suite includes a number of services and products, such as Research Data Australia, Register My Data, Identify My Data, and Publish My Data.

New services are in the pipeline over the next year, including creation of Digital Object Identifiers (DOI) supporting the citation of data; and services to support sharing controlled vocabularies. A re-design of the Research Data Australia look and feel is also on the schedule.

The existing services are proving popular as seen in the following usage statistics.

ANDS Service	Usage statistics
Research Data Australia	1319 Collections 192,212 page views since Jan 1 2010
Identify My Data Production Service	17,708 handles minted 11 machine to machine agreements
Identify My Data Test Service	35,966 handles minted 114 machine to machine agreements
Production Registry	12 provider organisations (1 OAI-PMH, 11 Direct) 2,906 records (1319 Collections, 3 Services, 894 Parties, 690 Activities)
Sandbox Registry	60 provider organisations (11 OAI-PMH, 49 Direct) 33,258 records (3753 collections, 29 Services, 2,392 Parties, 27,084 Activities)
Publish My Data	51 Collections manually published

Meet the ANDS Staff – Cynthia Love

Cynthia Love joins ANDS from CSIRO as Director, Public Sector Data.

Cynthia's background is in a variety of information areas. Originally trained as a medical librarian, she has worked in electronic information delivery in libraries, web management, management, records and data management. Early in her tenure at CSIRO an activity was training staff in the use of tools on a new innovation called the internet. Co-incidentally, the chair of the steering committee for the initiative was Ron Sandland.

More recently Cynthia has held the position leading CSIRO Library services and in that role initiated the revitalisation of technologies supporting the services, an activity that is critical in an organisation as geographically dispersed as CSIRO. From this position she has been seconded to establish Data Management Services in CSIRO: managing ANDS funded projects, the development of frameworks for data description and policy issues associated with use and publications of data within an information management environment. It is from there that she comes to ANDS.

This range of experience gives her an appreciation of the role of data and the inter-relationships in play in the wider information environment for research. Cynthia says, "There is a wealth of data that has been collected in the public sector that can fuel research that is critical for global innovation and national benefit." Cynthia is looking forward to examining the coverage of the content and working with agencies to further increase the exposure of this data. "Having experienced work on ANDS funded projects, I am also excited to see how the tools and processes being developed can be brought together to enhance capture and re-use."

"There is a wealth of data that has been collected in the public sector that can fuel research that is critical for global innovation and national benefit."

RIF-CS Schema Improvements

November 2010

In response to the experiences of early contributors to the Australian Research Data Commons, ANDS will introduce an enhanced metadata exchange schema, RIF-CS 1.2, from November 2010. The changeover period for harvested metadata will extend to June 2011.

The new schema:

- introduces new dataset citation and collection coverage elements;
- expands the related information element to allow description of links to publications; and
- simplifies the recording of telephone and fax numbers.

ANDS will establish a Community Advisory Board to help to determine future change requirements.

For more information see the RIF-CS Updates page on the ANDS website, http://www.ands.org.au/resource/rifcs_updates.html

ANDS events during August and September

The ANDS event calendar was full during August and September providing three significant opportunities for ANDS staff to be engaged with training and outreach.

The second and last ANDS Boot Camp was held at University House at the Australian National University from August 9 to 16. Twenty-one participants attended, from 17 universities, ANSTO (the Australian Nuclear Science and Technology Organisation) and IMOS, together with two ANDS staff. The success of the week-long program can be seen in a comment from one of the participants on the last day, when he remarked that he had come to the Boot Camp feeling that he was looking down a microscope but was leaving with the feeling that he was now looking through a telescope. All those attending are actively involved in their institution's Seeding the Commons programs, and are now well-equipped to take the ANDS message back home.

The final of the series of ANDS Roadshows was held in Townsville on September 21. There was strong interest from those in attendance in the workshops on ANDS Services and the *Code for the Responsible Conduct of Research*. The ANDS visit coincided with Research month at James Cook University where data management and the *Code* are high on the agenda.

There was a briefing at the Monash Conference Centre on September 2 for staff of the University of Melbourne, RMIT, Monash University, LaTrobe University and the Australian Synchrotron who are engaged in ANDS-funded Data Capture Projects. The briefing was professionally recorded and the videos will be available shortly to assist others engaged in similar projects elsewhere.

L to R: Toby Burrows (University of Western Australia), Joan Moncrieff (Deakin University), Anne Stevenson (CSIRO) and Salim Taleb (Curtin University) at the ANDS Boot Camp.

Forthcoming events

ANDS activities at eResearch Australasia 8 - 12 November 2010

8 November 2010 Data Sustainability: applying archival practice to research data workshop. Time: 1:30pm – 5:00pm

9 November 2010 ANDS Projects Birds of a Feather session. Time: 2:50pm – 3:50pm

9 November 2010 ANDS Poster session: Building a community of research data citation. Time: 5:15pm – 6.30pm

Details at https://ocs.arcs.org.au/public/conferences/1/schedConfs/1/program-en_US.pdf

Registration details at http://www.eresearch.edu.au/registration

Further information about the eResearch Australasia conference is available at http://www.eresearch.edu.au/index.html

Get notified about our forthcoming newsletters via RSS feed: http://ands.org.au/newsletter. For more news, alerts, announcements and discussion subscribe to the ANDS General group: http://groups.google.com.au/group/ands-general/subscribe

ANDS is supported by the Australian Government through the National Collaborative Research Infrastructure Strategy program and the Super Science Initiative.

This newsletter is designed by AK Design (http://www.akdesign.com.au)

ANDS Project Partners:

ANDS Office
C/- Monash University
Clayton, VIC 3800
Telephone: 03 9902 0585
Email: contact@ands.org.au