Teaching with Research Data

Report to the Australian National Data Service (ANDS)
August 2015

Margaret Henty
About the author

Margaret Henty has been involved with data management and training since 2005 when she joined the Australian Partnership for Sustainable Repositories. She joined ANDS (2009 – 2013) with responsibility for training and outreach and was later associated with policy issues around research data.

Margaret now works as a freelance consultant. Her publications include:

- Investigating Data Management Practices in Australian Universities
- Data sharing in a time of data-intensive research
# Table of Contents

**EXECUTIVE SUMMARY** ................................................................. 4  
**RECOMMENDATIONS** .................................................................. 5  
**INTRODUCTION** ......................................................................... 6  
**THE OPPORTUNITY** .................................................................... 6  
**SOME OVERSEAS DEVELOPMENTS AND INITIATIVES** ....................... 8  
* Data provision for educational use ............................................. 8  
* Commercial availability of data for teaching ............................ 9  
**ENVIRONMENT SCAN** ................................................................. 10  
* TEACHING WITH RESEARCH DATA IN SCHOOLS ....................... 10  
* TEACHING WITH RESEARCH DATA IN UNIVERSITIES .................. 12  
* NINE AUSTRALIAN UNIVERSITY CASE STUDIES ...................... 13  
* Cluster & Cloud Computing at the University of Melbourne: developing computation skills for big data ........................ 13  
* Urban Informatics at the University of Melbourne: analyzing & visualising big data 13  
* Conservation Biology at the University of Queensland: data creation and analysis in the biological sciences .................... 14  
* Climatology at the Australian National University: an example of research-led teaching .................................................... 14  
* History at the University of Melbourne: data creation and text analysis in the humanities .............................................. 14  
* Journalism and information management at the University of Technology Sydney: developing data skills for professional use ......................................................... 15  
* Approaches to Research in Education at Flinders University: learning social science methodology using real data .... 15  
* Archaeology at the University of New South Wales: archaeological practice through data use and analysis .......................... 16  
* eScience at Swinburne University of Technology: the role of IT in scientific discovery ......................................................... 17  
**BARRIERS, ENABLERS AND RECOMMENDATIONS** ................... 17  
* BARRIERS TO DATA USE .............................................................. 17  
* Inaccessibility .............................................................................. 17  
* Need for technical expertise ......................................................... 18  
* Ongoing professional expertise .................................................... 18  
* Unusable data ............................................................................. 19  
* Availability of computer labs ....................................................... 19  
* Issues associated with sharing course materials ....................... 19  
* Overseas findings ...................................................................... 19  
* ENABLERS AND RECOMMENDATIONS ..................................... 20  
* Universities ................................................................................. 21  
* Data creators .............................................................................. 22  
* Data managers and research infrastructure and data providers ................................................................. 23  
* Research funding bodies ............................................................ 24  
* Government ................................................................................ 24  
**ACKNOWLEDGEMENTS** .............................................................. 26  
**APPENDIX 1: SWINBURNE UNIVERSITY - COMPUTER LAB: TRANSFERRING AND STORING DATA** ............................................. 27
Executive Summary

Australia has a world-class research data infrastructure system and a research community which is increasingly able to create, find, analyse and reuse that data to solve the big questions of our time. The opportunity now exists to take further advantage of our data infrastructure, and the huge amount of data which has been amassed in it, to be extended and adapted to provide a comparable infrastructure to support teaching.

This study explores the potential for Australia to meet the demand for graduates who can go on to become business entrepreneurs, innovators, creative thinkers and thoughtful contributors to society in a world dominated by data. It looks at how we develop a conceptual understanding of the role and importance of data and the skills to take advantage of it. It recognizes that there are many skills contributing to data usage, but concentrates on developing data-capable researchers, entrepreneurs and citizens. At the same time, it recognizes, but does not explore in any detail, the associated need for more computation experts and data scientists.

The study identifies a number of initiatives from overseas which show recognition of the need for an increasing focus on education involving data, especially big data. It also identifies and describes a number of outstanding examples of undergraduate teaching in Australia which familiarise students with data analysis and presentation using real data. It identifies, in addition, some of the excellent resources available to support teaching in schools.

Teaching with data at the university level is not without its challenges and the study identified a number of barriers experienced by those designing and teaching courses involving research data. Intellectual property features as a major barrier to making both data and course materials more widely available as does the lack of data sharing and retention policies in universities. There are issues around access to data, as data which is available often has restrictive access permissions. Students without high levels of IT skill are at a disadvantage when there are not the facilities or tools available to serve as a bridge to the data. A lack of uniform computer facilities within a university can make practical sessions difficult to organize. Some data is simply unusable without considerable intervention in the form of cleaning or reformatting. Some of those teaching courses acknowledged their own issues in knowing what suitable data or tools are available. A major issue is course materials: those university teachers interviewed were keen to share their course materials but university copyright policies make that difficult. Other barriers are the result of university structures which separate teaching and research at high levels and cause problems for teachers, who are usually researchers as well.

Teachers would find their ability to incorporate research data into their teaching more straightforward if:

- research data was open, easily discoverable, immediately accessible and usable
- research data came with both the context and tools to assist with analysis and visualisation
- teachers were able to share research-data-related teaching materials and learn from each other.
Enabling these to happen will require input from a number of stakeholders, each with a different role to play: universities, data creators and managers, infrastructure and data service providers, research funding bodies and government.

Recommendations

Recommendations for universities
(1) That universities develop data sharing and retention policies which recognize the value of data and reflect the expectation that the products of publicly funded research be made available for the betterment of society
(2) That universities review their academic structures and policies, especially intellectual property policies, to take advantage of the synergies between research and teaching to enable greater openness of research data and teaching materials
(3) That universities, and course designers and teachers, consider the need for an increasing focus on the inclusion of teaching with and about research data as a means of developing professional competencies in their graduates and develop graduate attributes to support this focus
(4) That universities provide educational programs for their own researchers to encourage them to consider the long-term value and diverse potential uses of their research data at the point of creation and to provide them with skills in the use of emerging tools and facilities that underpin data discovery and reuse
(5) That university data support services recognise the value of research as an input to teaching and restructure their services to encourage use by students and teachers
(6) That universities ensure that IT infrastructure is adequate to meet the classroom needs of courses which include data handling.

Recommendations for data creators
(1) That data creators consider future use of data early in their project and at the same time as developing a data management plan and determining licence conditions.

Recommendations for data managers and infrastructure and data service providers
(1) That data managers provide metadata to support enhanced discoverability and access and provide linkages between the outputs of research to promote discoverability, context and necessary tools
(2) That data managers and infrastructure and data service providers ensure that access to data is provided with minimum barriers wherever possible
(3) That infrastructure providers consider the educational benefit of proposed tools and services used in research at the time of planning, that provision be made in any proposed infrastructure development for teaching materials to be linked to corresponding data sets and that tools be developed to support both basic needs and as well as advanced research
(4) That infrastructure and data service providers provide discovery and other support services optimized specifically for teachers and students.

Recommendations for research funding bodies
(1) That explicit recognition of the value of research data to education be given when grants are considered
That data usage for educational purposes be recognised as an impact measure,
That more encouragement be given for data to be as open as possible, preferably by use of an open licence.

Recommendations for government
(1) That more consideration be given to the synergies between research and education in the development of policy designed to enhance Australian innovation and competitiveness and the ways in which the benefits of research could be leveraged into teaching
(2) That government policy is developed to ensure that access to public sector data and associated teaching materials is provided for the benefit of both research and teaching
(3) That the remit of NCRIS be amended to incorporate education as well as research
(4) That the government provide funding for the development of a facility comparable to the US Oceans of Data.

Introduction
The Opportunity
Midway through the second decade of the 21st century sees Australia in an enviable position to exploit the educational and economic opportunities made available through a world-class research data infrastructure system. Our business, education and research communities are together creating an increasingly sophisticated and interlinked community who can create, find, analyze and reuse data to solve the big questions of our time.

Data is important. We collect it to help solve problems, both minor and major. We need it to support the innovation necessary to take society forward. In recent years we have seen an explosion in the amount of data collected and the means of exploiting it. Data has become more complex, often needing supercomputers, virtual laboratories and other advanced supporting services in order to use and analyse it. A substantial investment has been made in developing infrastructure to support Australia’s burgeoning research requirements through the National Collaborative Research Infrastructure (NCRIS). Global changes in policy around open data and the parallel development of data discovery services have meant that data is now available and discoverable in astounding quantities.

The collection of so much data provides the opportunity to develop further education, not just research and raises the issue of what more can be done develop the skills to support full exploitation of the data as well as to help further develop the infrastructure that underpins it?

New employment paths associated with data have created the need for high-level skills in a number of areas:

- Sophisticated computational skills to support services around data, especially big data, those huge datasets of such size and complexity that they cannot be processed using traditional means.
- Data scientists, those who can extract knowledge from the data, both big and small
- Researchers and professionals with domain-specific skills, such as geologists, demographers, engineers and urban planners, need analytical and visualisation skills to work effectively
Businesses need to be able to analyse data to provide a basis for innovation, identify trends, improve services and communicate with clients. Governments collect huge amounts of data which can be used to guide policy development and improve efficiency. Many individuals become citizen scientists and enjoy the reward of adding to our understanding of the world.

Recognition of the development of special skills to support services around big data was given at a Workshop of the National Academies in the US in 2014. The report of the Workshop quotes McKinsey & Company as having estimated that there is a ‘shortfall of 150,000 data analysts and 1.5 million managers who are knowledgeable about data and their relevance’ and goes on to say:

Training students to be capable in exploiting big data requires experience with statistical analysis, machine learning, and computational infrastructure that permits the real problems associated with massive data to be revealed and, ultimately, addressed ... Analysis of big data requires cross-disciplinary skills, including the ability to make modeling decisions while balancing trade-offs between optimization and approximation, all while being attentive to useful metrics and system robustness.

The National Academies workshop recognised that there is no one-size-fits-all approach to developing the requisite skills around big data but mentions degree programs for novices and certificate programs or boot camps to build on existing skills. This is one area where MOOCs are proliferating.

If the US is in the situation of needing to train more computation and data scientists, then it follows that Australia does also. The Chief Scientist in 2013 recognised our critical need for education in all science, technology, engineering and mathematics (STEM) disciplines, all of which rely heavily on data, urging action to provide such education as Australia runs the risk of being left behind. A recent report from Price Waterhouse Coopers stresses the need to future-proof the economy by developing STEM skills. At an anecdotal level, Professor Lindsay Botten of the National Computational Infrastructure (NCI) has commented on the difficulty of finding appropriately qualified staff to work there.

The increasing availability of data comes at a time when teaching styles in education are changing. Information technology, too, has become a critical underpinning of the changing curriculum. Research, enquiry and technology are intrinsically interwoven, especially at tertiary level where some universities have actively embraced a research-informed approach to developing curricula. Learning is investigative and collaborative, and research

---


2 Training Students to Extract Value from Big Data, p.2.

3 William Howe reported at this Workshop on teaching Introduction to Data Science, a MOOC offered by the University of Washington. While attrition rates were high, about 7,000 passed the course and were issued certificates. Most of those enrolling were computer science professionals. Training Students to Extract Value from Big Data, p.29. [http://www.nap.edu/catalog/18981/training-students-to-extract-value-from-big-data-summary-of](http://www.nap.edu/catalog/18981/training-students-to-extract-value-from-big-data-summary-of)


6 Conversation with the author, 4 March, 2015.
data has become both a key means of enriching the learning process, and a teaching tool. Graduates in many disciplines are expected to be able to identify, access, analyse data and report findings as a part of disciplinary practice and evidence-based enquiry.

As will be seen in the case studies below, some Australian universities are already acting to include data and its usage by using research data as a teaching tool in the classroom. By taking advantage of the synergies between research and teaching, they are utilizing real data to make learning more meaningful.7 Students are becoming better prepared to move into professional careers that require skills in data identification, analysis and presentation. Research could benefit from the participation of students in a number of ways. Teachers at the tertiary level are often researchers who incorporate data they have created into their courses that could in turn benefit from student input and interpretation. Citizen science projects of all kinds can benefit from student contributions while the students enjoy the meaningful experience of contributing real data to a worthwhile undertaking.

At the high school level, the Australian National Curriculum introduced in 2013 places significant emphasis on the integration of ICT capabilities within science subjects. It states that Year 9 and 10 students will:

\[
\text{... employ their ICT capacity to access information; collect, analyse and represent data; model and interpret concepts and relationships; and communicate science ideas, processes and information.}^8
\]

One implication of this change to the high school curriculum is that many of those arriving at university as early as 2016 will be better skilled in identifying and handling data than undergraduates have been in the past.

Some overseas developments and initiatives

A number of overseas initiatives have been designed to improve access to data for teaching purposes and provide useful models. Some of these are outlined below and raise the question of what more could be done in Australia to support data use in the classroom.

Data provision for educational use

There are good examples from overseas of groups with active outreach programs designed to encourage investigation in particular disciplines and making available data resources and tools to support that.

The social sciences have a long history of collecting research data and making it available to others. An understanding of methodology is fundamental to social science research; so making the data available for teaching at undergraduate and school level is an automatic next step. Both the UK Data Service and ICPSR (Inter-University Consortium for Political and Social Research) provide resources for instructors and students. The UK Data Service offers a web page devoted to teaching resources ‘Demonstrating how data can be used in teaching’.9 The extensive offerings from ICPSR include webinars and tutorials on their own

8 Australian Curriculum Assessment and Reporting Authority, Australian Curriculum, F-10 Curriculum http://www.australiancurriculum.edu.au/science/general-capabilities
9 http://ukdataservice.ac.uk/teaching-resources.aspx
YouTube channel, not to mention learning guides, exercise sets, links to appropriate data sets and a crosstab assignment builder. By contrast, the Australian Data Archive (ADA) has been unable to obtain the resources for comparable activities, despite the fact that about 20% of those who are registered to use the Archive are known to be undergraduates.

In the US, the National Science Foundation has made funding available to support Teachingwithdata.org, a partnership of ICPSR, the Social Science Data Analysis Network (SSDAN) and the NSF. This is a service to help teachers bring real data in primary, secondary and post-secondary classrooms with the intention of building quantitative literacy. Instructor resources include lessons, activities, tools, events, pedagogical resources and links to data providers. Their blog, Data in the News, provides regular postings about research on topical issues of likely interest to students and includes links to teaching tools, often within days of an incident being reported in the news.

The Learning and Teaching Division of the Education Development Center, Inc. (EDC) offers a broad range of teaching materials. Their Oceans of Data Institute (ODI) is an initiative designed specifically to prepare students for a data-intensive world by introducing them to big data. They offer services for teaching institutions, teachers and students. Materials are suitable for both undergraduate and school students and ODI staff are available for consultation with teaching bodies to help design and implement training, curriculum, and/or assessment programs. In 2014 the ODI convened a panel of experts to develop a Profile of the Big-Data-Enabled Specialist.

The Ecological Society of America has worked with partners and with funding from the NSF to develop an outreach program directed at the educational community about the National Ecological Observatory Network (NEON). NEON will provide data on the impacts of climate change, land-use change, biodiversity and invasive species through a continental-scale ecological observation system. NEON provides teaching materials as well as Project Budburst, a citizen science program with a strong educational focus to collect data on when plants leaf, flower, and fruit.

The Genomics Education Partnership was established to provide opportunities for undergraduates to participate in genomics research by introducing students to DNA sequencing, gene annotation, genomic analysis and other elements of genomics. They provide course syllabi and curriculum materials from partners as well as a data repository.

Gapminder is a project commenced originally in Sweden and offers a range of activities, guides, games and tools for students to learn how to manipulate and understand statistics relating to social, economic and environmental development at local, national and global levels. This non-profit venture promotes global development and achievement of the United Nations Millennium Development Goals.

Commercial availability of data for teaching
At least one commercial provider is making data available for teaching purposes. Sage Research Methods is an online interactive tool which brings together content of Sage books, journals and reference works to help researchers and students master methodological problems associated with their research. Integrated into this is SAGE Research Methods Datasets, a conveniently packaged collection of

---

10 www.youtube.com/user/icpsrweb
teaching datasets that can be used to support the teaching of quantitative and qualitative analytical methods used in the social sciences. These are datasets taken from larger national and international data sources, cleaned and reduced in size and complexity for teaching and self-study purposes, perfect for researchers, learning a new method, or brushing up on a familiar one.

One can see the attraction for teachers of resources available in one place with no restrictions on use, no copyright issues to deal with, clean data in readily usable formats. There are two disadvantages to this service: the data may not be of immediate relevance to the subject being taught and the need for a subscription to gain access. Two-thirds of Australian universities have current subscriptions to Sage Research Methods Online.¹¹

Environment scan

Teaching with research data in schools

While the primary focus of this report is university teaching, the following examples illustrate some of the resources emerging for high school use.

There has been some impetus to using data in high schools with the release of the National Curriculum in 2013.¹² At the same time, other factors are making it easier for students to connect with data and, in some cases, providing them with practical and engaging opportunities to become involved in data creation. Citizen science projects are proving popular, with some projects actively seeking the participation of school students through outreach programs.

There are notable examples of data collecting programs and other data services which have outreach programs designed specifically for school children. Others have programs which may include school children although not be specifically addressed to them.

• **Pulse@Parkes** is an initiative of CSIRO designed to provide high school students with an opportunity to control the Parkes radio telescope for the purpose of studying pulsars under the guidance of astronomers and possibly identifying a previously undiscovered one. Students can work from a number of different observing facilities around Australia or they can download and analyse data already collected. All data collected is made available for ongoing research.

• **Census at School** was a nation-wide project conducted by the Australian Bureau of Statistics that collected information about the characteristics, attitudes and opinions of students based on completed questionnaires. This information was then made available for schools to support the teaching of statistics, mathematics, social science and other subjects. Suitable resources were provided for teachers of students in year 4 through to senior levels. Teachers were able to contribute activities they had developed to support use in the classroom and these were made available for the benefit of other teachers. The program was discontinued in late 2014.

¹¹ Email from Diane Costello, Executive Officer of the Council of Australian University Librarians, 24 March, 2015.
The Atlas of Living Australia (ALA) is one example of the many projects which provide access to large amounts of data while at the same time encouraging the contribution of new discoveries from students and members of the general public. ALA collects biodiversity data and brings together data from a wide variety of Australian sources such as the Australian National Insect Collection (ANIC), the Tasmanian Herbarium and the Royal Botanic Garden in Melbourne. New sightings, once verified, are added to the database. Software is available for local communities to record findings and an education resources page provides activities and step-by-step teacher guides for different age groups. One example of school use, by Merici College in the ACT, is described on their website:

*Students use the spatial portal to determine the number of species within a 1km radius of their school and compare it to the number within a 5km radius that takes in part of the nearby Mount Ainslie nature reserve. From there, they construct food webs using the species they find. The students also look at a local endangered species, the Golden Sun Moth, and suggest ways they could set up a conservation plan for the moth in the school grounds.*

There are many examples of other citizen science activities in the biodiversity field of likely interest to school students of which ALA is but one. Others include RedMap, School of Ants and Microblitz. There is great potential for these to be developed further to incorporate facilities for teachers, such as activities, course materials related to the curriculum or provision for a teacher discussion forum.

Fungimap is one of the citizen science initiatives contributing data to ALA and is dedicated to the identification and documentation of Australian fungi. Fungimap actively encourages school participation and feedback from both students and teachers is positive. Resources and activities are linked to the national curriculum. Its popularity with children from the age of three upwards is described by Jasmin Packer, who works with the project in South Australia:

*Children are intuitively drawn to things in their world that are physically scaled to their size (e.g. small toys they can role play with), and this means when they are given the opportunity to spend time in nature they are delighted to discover they are the first ones to find the fungal! Fungi are mostly at their physical level, and often missed by adults. The tiniest, most cryptic, specimen will be sought out with great glee!*  

GovHack is an annual open data competition described on its website as being for ‘web and application developers, open data & visualisation gurus, user experience folk, graphic artists, data journoes, augmented reality-ists, mobile maesters, and open government enthusiasts’ both Australian and, more recently, New Zealanders. It is not specifically designed for school students, but has attracted entries from schools and has prizes in a special category for them. Its value for students is described by Matthew Purcell of Canberra Grammar School whose students have participated in GovHack since its inception in 2012:

---


14 Email communication May 6, 2015.
Unlike other events in which students participate, GovHack is not a school event – it is for hackers from all backgrounds ranging from professional developers and designers through to hobbyists. This provides relevant and authentic context for our students to further develop and apply their skills. Additionally, the community nature of GovHack, with the focus being open data, provides the students with exposure to working collaboratively, engaging with other teams, seeking assistance from subject-matter experts, and being involved in a major nation-wide event.\textsuperscript{15}

- **N3xGen Unleashed** was introduced in 2015 as part of Unleashed, South Australia’s annual open data competition which is part of GovHack. N3xGen Unleashed is designed for competitors aged from 13 to 18 and runs simultaneously with the main event. The event runs over 46 hours and competitors are asked to provide a proof of concept and a video telling the story of how the data can be used.

- The University of Southern Queensland (USQ) has an active program of providing data to schools and the general public in association with the development of practical activities for the classroom designed to demonstrate the dangers of UV radiation to the skin. The University is making data on UV radiation, cloud cover and photosynthetic reactive radiation, collected since 2003, available though QCIF to support the project. In 2014 they received a grant from the [auDA Foundation](http://www.govhack.org/2014/10/06/govhack-and-high-schools/) to develop an online interactive platform to educate children and members of the public. Staff of the University work together with teachers in developing curriculum materials to underlie the project. The USQ Library has also received funding from ANDS to assist with data description and management to support this program.

### Teaching with research data in universities

In preparing this report, case studies of courses offered by Australian universities and involving the use of research data as a tool for teaching were identified. These were either at undergraduate or master-by-coursework levels. Interviews were held with the teacher responsible for the course. In all cases, the teachers are also researchers and in some cases, they were involved with the course design.

By research data is meant all data created or used in the course of research.\textsuperscript{16} For the purposes of this investigation, only research data in digital form is included. While ANDS is responsible for Research Data Australia (RDA), a registry of Australian research data from higher education and government, the report is not limited to data available via RDA.

The case studies included here cover a number of data usage scenarios:

- Where data is downloaded to a simple spreadsheet for analysis or other activities such as visualization or data-mining
- Where data from an external source might be compared or augmented with data collected by the student
- Where the student might be asked to work with big data in ways which involve the development or use of sophisticated tools and analysis

\textsuperscript{15} [http://www.govhack.org/2014/10/06/govhack-and-high-schools/](http://www.govhack.org/2014/10/06/govhack-and-high-schools/)

\textsuperscript{16} For a fuller description of what might be covered by the term ‘research data’ see the ANDS Guide [What is Research Data?](http://ands.org.au/guides/what-is-research-data.html)
• Where a student might be learning how to prepare data for curation and subsequent identification and use by others
• Where a student is asked to design a system to facilitate use of data by others.
• Where a student might be asked to do a textual analysis of a literary work

Nine Australian university case studies

Nine case studies of courses that incorporate data usage in one form or another are outlined below. Interviews were held by phone or in person with those responsible for teaching or developing the course. The course descriptions below are brief as many of the issues arising during the interviews were common to more than one course and are dealt with in the Enablers and Barriers sections of this report.

Cluster & Cloud Computing at the University of Melbourne: developing computation skills for big data

Professor Richard Sinnott is Director of eResearch at the University of Melbourne and, in addition, is a professor in the Department of Computing & Information Systems in the Melbourne School of Engineering. For the past three years he has been teaching Cluster and Cloud Computing (COMP90024), a course designed for postgraduate and upper level undergraduate students in software engineering. He describes himself as ‘very much an applied computer scientist’ who builds software systems for researchers, one example being AURIN, the Australian Urban Research Information Network which provides access to over 1,000 datasets from government and commercial sources distributed across Australia.17

This is the third year the course has been offered and student numbers have grown each year, reaching 120 in 2015. The course offers students the opportunity to work with ‘big data’ by developing technologies which familiarise them with cluster and cloud computing. The emphasis is not so much on the research data itself but on giving students the capacity to develop middleware and applications which permit access to the data and analysis of it.

The course utilises the facilities of the University of Melbourne and the NecTAR cloud. The data used for analysis comes from many sources: Twitter, AURIN, the Complete Works of Shakespeare and others, enabling students to become familiar with the requirements of different types of data.18

Urban Informatics at the University of Melbourne: analyzing & visualising big data

Associate Professor Chris Pettit is the Strategic Implementation Coordinator of the Australian Urban Research Infrastructure Network (AURIN). In addition he has been teaching Urban Informatics (ABPL90366) at the University of Melbourne. However, he will be moving to Sydney later in 2015 to take up the Inaugural Chair in Urban Science at the University of New South Wales where he will be establishing a data-focused masters program in urban science.

Urban Informatics is a course for higher-level undergraduates and masters-level students in architecture, urban planning, and geospatial information technology. While not essential to these disciplines—the course is elective rather than compulsory—students regard it as a valuable addition to their skill sets. The course has only been taught for two years and was

17 http://aurin.org.au/
most recently offered as a one-week intensive over the summer. Like Cluster and Cloud Computing at the University of Melbourne, data usage is based primarily but not exclusively on use of AURIN. The students learn to use tools for mapping, graphing and other forms of visualisation, as well as analysis. In addition, they learn about metadata, data management principles and spatial data infrastructures so that at the completion of the course, they have gained a broad understanding of the data, know how to add value and to use it in novel ways.

**Conservation Biology at the University of Queensland: data creation and analysis in the biological sciences**

Dr Ross Dwyer is a Professorial Fellow who has been associated with the development of Oz-Track, online software for storing, analysing and visualising animal location data. Oz-Track provides a suite of tools that make it possible for those without software development skills to analyse and map animal tracking data. It has proved useful as a teaching tool for students at both undergraduate and high-school level in various parts of Australia.

*Conservation and Wildlife Biology* (CONS6009), is a course offered at master’s level which combines both the theory and practice of conservation science and using the Oz-Track facility. Ethics permission is obtained for tagging animals—in this case cane toads—which are then released and tracked as part of the research projects developed by the students. The data is analysed to answer questions posed by the student projects.

All data collected is uploaded into OzTrack so that it is available to others interested in the topic. By using real, rather than synthesised, data students are exposed to a wider range of issues associated with conducting research, such as the creation of metadata for their data files, the ethical considerations around animal research and the need for appropriate sampling.

**Climatology at the Australian National University: an example of research-led teaching**

Professor Janette Lindesay is Associate Director Education & Deputy Director of the Fenner School of Environment and Society at the Australian National University. She teaches both undergraduate and postgraduate courses in climate science in addition to her research activities. Being able to manipulate and interpret data is seen as part of courses that are designed to introduce the student to all aspects of research.

*Weather, Climate and Fire* (ENVS2004 & ENVS6204) is a basic scientific introduction to atmospheric science and focuses on the fundamentals of meteorology and bushfire science. Students learn to analyse and manipulate data provided to them from the Bureau of Meteorology (BOM) or other sources and may be required to compare it with data collected as part of their practical work.

*Climatology* (ENVS3013 & ENVS6303) involves the students developing their own research questions to which they then have to find data to provide an answer. The data used may come from a variety of sources, most commonly the Bureau of Meteorology but US and other international sources as well. Data may be non-numeric as well as numeric. Professor Lindesay regards the opportunity to set their own research questions as empowering for the student.

**History at the University of Melbourne: data creation and text analysis in the humanities**

David Goodman is Professor in the Department of Historical and Philosophical Studies at the University of Melbourne. He teaches, among other courses, the compulsory capstone
course, *Making History* (HIST30060). This course is designed to have a digital component in addition to research in a traditional paper-based archive. To quote Professor Goodman:

*There are certain skills that a twenty-first century graduate in history should have and this subject is one way in which we ... try to embed them in it.*

The course includes an introduction to big data. The creation of vast textural databases means that it is possible not just to find information but to analyse and interpret in ways not previously available for teaching or for the general public. Students are encouraged to use databases of text such as Google Books (using NGram) or the NLA’s Trove newspapers (using QueryPic) to generate graphs of word-use or phrase-use frequency to see whether their findings accord with other evidence or to otherwise add depth to their assignments.

Students are required to open a Zotero account, and use it to collect and present information they have collected. Use of Omeka is currently under investigation: this is an open-source web-publishing platform to display online exhibitions of scholarly materials and has been one of those resources publicised by the University of Melbourne Research Bazaar.

*Journalism and information management at the University of Technology Sydney: developing data skills for professional use*

Maureen Henninger teaches at the University of Technology, Sydney where she is Senior Lecturer in the IKM and Digital Studies Program and Program Coordinator for the Faculty of Arts and Social Sciences. In the IKM and Digital Studies Program there are a number of courses in the BA (Communications), at either undergraduate or masters-by-coursework level, which involve students identifying and manipulating data. These include Digital Assets Management (57204), a course currently offered at both levels and Investigative Research in the Digital Environment (57152) offered only (at the moment) to masters students. Further courses involving digital data are under development.

Digital Assets Management concerns the management of digital assets, text, sound, still and moving images, through their lifecycle. Investigative Research in the Digital Environment concerns the discovery and use of all types of data, but especially that from the public sector, for data mining, analysis and presentation. Data sources used include ABS, the World Bank, data.gov.au (and overseas equivalents) and may be textual or numeric. Both of these courses incorporate the use of research data as there is an immediate relevance to the professional careers in information management, journalism or related fields for which students are being trained.

Both of these courses are seen as relating closely to the UTS graduate attributes such as professional readiness, a capacity for critical and creative enquiry and developing students to be life-long learners.

*Approaches to Research in Education at Flinders University: learning social science methodology using real data*

David Curtis is Associate Professor, Educational Research at Flinders University where he also teaches a number of subjects to do with statistical methods and research techniques as part of masters-degree courses in education.
One subject in particular, *Approaches to Research* (EDUC9761) has a focus on using authentic data, most often PISA data from the OECD\(^{19}\) and LSAY data collected by ACER\(^{20}\). The primary object of the course is to introduce both qualitative and quantitative approaches to research, including the identification of problems, literature review, developing questions and hypotheses, collecting and analysing data and reporting and evaluating research. In this context, use of authentic data provides an opportunity to explore real issues within the field and to integrate the students’ studies with their own research interests. Using real world data means that students confront issues such as incompleteness, measures which are ‘not quite as good as you might wish they were’ and other problems usually not encountered with manufactured data sets.

Professor Curtis designs his courses to take into account the Flinders University graduate attribute concerning effective communication. He sees the need to ensure that students not only can analyse data but also are able to report their findings in ways that demonstrate that they can communicate effectively:

> ... I get students to write a report of their analyses and I see that their report is just one opportunity to ensure that students can communicate ideas effectively and I put quite a bit of emphasis on how to write readable reports. So I provide guidelines, obvious things like how to construct tables, how to use figures, but it’s also about being selective in the information that you do need to present. Because there is the temptation to do an analysis in ... a stats program, to get an output table, or just copy the thing into a Word document, when most of the output is irrelevant.

**Archaeology at the University of New South Wales: archaeological practice through data use and analysis**

Shawn Ross is Associate Professor at Macquarie University where he will be teaching later in 2015 in their Bachelor of Archaeology degree, having recently moved from the University of New South Wales. He spoke with enthusiasm about the value of courses which involve active student participation in the practical aspects of archaeology, including the collection and interpretation of data. This, he feels, is essential if students are to be adequately prepared to take up work in archeological consultancies or in cultural heritage management or as researchers.

In one instance, an ARC Linkage Grant allowed him and his team to take fifteen honours students to Bulgaria to experience fieldwork at first hand. They found that the students were taken aback to discover the extent to which technology is involved in archeology and how data intense it is. Being in the field allowed students to observe the data being collected and to see preliminary results quickly available, providing an authentic experience of what archeology is like. The University of New South Wales includes in its list of graduate attributes the need to be both ‘information literate’ and ‘digitally literate’, both of which have relevance to the teaching and practice of archaeology.

Macquarie University is now establishing a new Bachelor of Archaeology degree with a strong methodological and practical focus. He is hoping to be able to provide access to data collected in Bulgaria through Open Context, a US-based site that reviews, edits, and publishes archaeological research as well as archiving data. He feels that if students are unable to have that practical experience, then having access to real datasets can allow them

---

to work with data, apply different approaches to analysing and interpreting the data and
gaining a better idea of archaeological practice than simply reading a printed report.

**eScience at Swinburne University of Technology: the role of IT in scientific discovery**

Chris Fluke is Associate Professor in the Centre for Astrophysics and Supercomputing at Swinburne University of Technology where he teaches *Introduction to E-Science* (ICT10007), a compulsory second-semester unit introduced for first year science students in 2014.

The course is intended to show students what a career in science might look like in five to ten years when many will move from being laboratory-based to become scientists who do most of their discovery and analysis by dealing with very large data sets. The aims and the objectives of the unit were to introduce students to the fundamentals of e-science and the role that IT in its broader sense plays in scientific discovery. The unit tries to prepare students for some of the big data challenges of modern and future science careers, equip them with some basic practical strategies for data analysis, develop a basic capability around data modelling analysis and visualisation and build an understanding of how e-science influences some of the research activities at Swinburne. Students come from all scientific areas, meaning that some find this approach an unexpected challenge as their background has not focused on computers in the same way.

The focus of the course is big data and the approach to it is largely conceptual rather than hands-on. However, there are practical sessions involving participation in a citizen science project, data mining, data storage and transfer and other topics as well as weekly discussion groups on particular questions. Appendix 1 is an outline of a computer lab session on transferring and storing data.

**Barriers, enablers and recommendations**

**Barriers to data use**

From the perspective of university teachers who are designing and delivering courses, there are many factors seen as barriers to the use of data in the curriculum. From a broader perspective, one can see that there are structural issues here that might also have an impact. Major barriers include:

- inaccessibility,
- need for technical expertise on the part of the student,
- ongoing professional development of teachers,
- usability of the data,
- availability of computer facilities within the university, and
- inability to share course materials.

**Inaccessibility**

Unfettered access to data was seen as a key requirement in all of the nine case studies. Accessibility was mentioned frequently as an issue and most of the courses outlined above utilise data already known to the lecturer as being straightforward to access. For some courses, the data is downloaded and made available through the university’s learning management system rather than students having to locate it and download it themselves. In only one case study were students expected to identify data resources themselves: this was the *Investigative Research in the Digital Environment* course at UTS where students are expected to become familiar with sources of government or other official data likely to be
of value to journalists. For most courses, the focus is on analysis and presentation of the data, with discovery seen as secondary.

There are different reasons why data may not be readily accessible.

- **Copyright** Where there are copyright restrictions on data access, students must spend time seeking permission: unrealistic for students on tight time schedules to complete assignments. While it is possible for teachers to seek prior permission, and some do, this too was seen as an impediment: ‘I think … if you had to make individual agreements with research organisations, that kind of thing, it starts to get quite time consuming and potentially messy – not impossible but not desirable.’

- **Ease of Access** It may not be possible to download data directly via the web because of the need to request data from an individual, register interest or otherwise face delays in access. This was cited as a major barrier to asking students to discover data for their own use and was mentioned in particular in relation to registries such as RDA where students are delighted to locate data only to be disappointed when they cannot get access.

As more online courses are developed, and in particular, MOOCs, having a large corpus of openly available data would increase opportunities for innovative learning tasks for students and course design and outcomes for teachers.

**Need for technical expertise**
Most courses are domain-focused, which means that students do not necessarily have the technical expertise to handle the data. As one teacher put it:

> Freer access to data through online portals, web-based, would be very desirable and for the sorts of students we teach here who would not be mathematics majors or have computer programming skills … it’s not just you can ftp this dataset off that website, you can actually generate the map or the graph or whatever online … I’m trying to minimize the need for the technical expertise. We want to be able to use simple statistics with a package.

A number of those interviewed made specific mention of the need for easy to use tools to support analysis and presentation of complex or large datasets. This applies especially where students are handling big data and need access to virtual laboratories or online tools. **Oz-Track** is one example of a tool developed originally as a data depository but then extended to include analysis and presentation tools. This has in turn led to an increase in use and also a broadening of use into the teaching sector, including at least one high school.

**Ongoing professional expertise**
Many of the teachers commented on their own needs for ongoing professional training to help them keep up with developments. This related particularly to a lack of familiarity with the variety of data now becoming available and the need for proficiency in handling the many tools associated with data reuse.

One initiative of interest here is the **Research Bazaar** developed at the University of Melbourne. This program is directed at training researchers to use tools such as MATLAB, Python and Omeka and has been well received. As a side benefit of the training, **Omeka** is already being considered for undergraduate teaching in History.
**Unusable data**

Data once retrieved may not be usable. This can occur

- if the data is in proprietary formats or requires proprietary software for analysis. Encouraging use of open source formats when depositing data may relieve this.
- if the data has not been properly cleaned, meaning that the student must clean. Some tools are now available to assist data cleaning, such as [OpenRefine](https://openrefine.org).
- if it is in a format not suitable for analysis. Some numeric data available from data.gov.au and other repositories are in pdf format that has to be reformatted as csv or another readily usable format. Greater awareness of the impact of format on data reuse would increase the stock of data availability for teaching.

**Availability of computer labs**

Sometimes a barrier is quite unrelated to the data itself. One teacher, when asked what the major barrier to data use is, commented ‘not enough computer labs in the university’. While it is possible for students to bring their own laptops to class, it is preferable to have standard setups available for all students when analysing data.

**Issues associated with sharing course materials**

All of the university teachers interviewed for this report agreed that they would benefit from the sharing of course materials and also indicated a willingness to share their own materials. However, all could see logistical issues in sharing:

- where to house the materials
- ownership of the IP
- university policy concerning sharing, and
- how to make them known to others.

The two most intractable of these issues are ownership of the IP and university policy towards sharing. In no Australian university is copyright in teaching materials held by the creator but instead it is held by the institution. In order for course materials to be shared, university policies towards such valuable assets would have to be changed.

Course materials are usually housed in a learning management system (LMS), such as Moodle or Blackboard. These systems are designed for use within the institution, so course materials held there are unavailable more widely, a technical barrier that could be overcome in a variety of ways. However, there is at present no impetus to do so. There is no learning object repository in Australia, although the possibility of developing one has been the subject of discussion for some years. Most LMS systems, including Blackboard and Moodle, have course cartridge interoperability which means, in theory, course materials can be shared across LMS systems.

There has been a suggestion that teaching materials associated with particular datasets could be identified via metadata inserted into the record for that data held in RDA or other data registries. Some data services already integrate course materials as part of an outreach program. [AURIN](https://www.aurin.org.au) is one such example, as is the [Atlas of Living Australia](https://www.allenstim.gov.au), although the latter at this stage only has materials suitable for school use.

**Overseas findings**

Problems similar to those reported locally have been reported from overseas. In relation to big data in ecology, in the US, Langen reports:

> How practical are big data activities for undergraduate institutions? Our working group was surprised at how difficult it was to find the specific data needed for an exercise online (if they existed at all). Creating adequate data templates and
documents to guide students was also very challenging; we found that students required considerably more technical help than is typical for an ecological exercise. We needed to carefully “scaffold” concepts and tasks to make these exercises doable by our students. Additionally, at some institutions, we had to deal with imposed cyber-security blocks on downloading computer applications and data onto instructional computers.²¹

A 2001 report from the UK Data Archive, An Enquiry into the Use of Numeric Data in Teaching and Learning²² included a discussion of barriers to using data. The report is now over ten years old but it appears that many of the barriers still remain:

The survey [of university teachers] uncovered a number of barriers experienced by teachers in the use of these services, namely a lack of awareness of relevant materials, lack of sufficient time for preparation, complex registration procedures, and problems with the delivery and format of the datasets available. A compounding problem is the lack of local support for teachers who would like to incorporate data analysis into substantive courses. A majority of the survey respondents said that the level of support for data use in their own institutions was ad-hoc. Peer support was more common than support from librarians and computing service staff, and over one-third received no support whatever. The top three forms of local support needed were data discovery/locating sources, helping students use data, and expert consultation for statistics and methods (for staff).

Enablers and recommendations
Incorporating research data into teaching would be more straightforward if:

- research data was open, easily discoverable, immediately accessible and usable
- research data came with both the context and tools to assist with analysis and visualisation
- teachers were able to share what they are doing and learn from each other.

Researchers would appreciate the first two of these as well. Only the third relates solely to teaching.

Identifying those factors that might enable and facilitate the use of data as a teaching tool in undergraduate education involves looking at the policies and practices of government, research funders, universities, infrastructure creators, providers of data services and data managers.

The enablers and recommendations below will be discussed in terms of the different responsible stakeholder groups:

- universities
- data creators
- data managers and research infrastructure and data providers

http://www.researchgate.net/publication/267276603_Engaging_undergraduate_students_in_ecological_investigations_using_large_public_datasets_lessons-learned_by_a_teaching_with_big-data_working_group

• research funding bodies
• government.

**Universities**

Universities could take a number of actions to facilitate the use of research data as a teaching tool, while at the same time improving the potential to increase the overall value of their data for research and innovation.

**University policies** could be enhanced to facilitate the discovery of and accessibility to data for the benefit of teaching, research and innovation through:

• A data sharing and retention policy which would recognise the value of data to the community as a whole and reflect the growing expectation that all products of publicly funded research be made available for the betterment of society. Such a policy would identify the wide potential reuse for data (government, education, business, citizens), make clear the need for open licences and offer incentives to follow the policy. It would assume that intellectual property exists in creative data and clarify ownership of the data.
• University intellectual property policies amended to take into account the benefit of openness and clarify ownership.
• Development of policies to permit the sharing of course materials would extend the range of materials available to teachers and course designers and would be welcomed.

Organisational structures in universities are not always best suited to maximise the leverage between research and teaching and could be reviewed from this perspective.

**Professional competencies.** Universities, together with course designers and teachers, could consider the role of research data in education in terms of the contribution this makes to professional skills development and graduate attributes that take into account the acquisition of research skills.

In this fast changing environment, universities could consider the value of programs to extend skills in using emerging tools and facilities for data identification, analysis and visualisation, for research and teaching purposes for students. While the case studies included above can be seen as exemplary, it is not always the case that consideration is given to the inclusion of research data as a tool in university courses.

**Initiatives to realise the potential of data beyond research**

There is still much to be done to educate researchers to consider the long-term value of their data beyond their own project and discipline. Open data practices are not uniform across disciplines.

---

24 Griffith University, for example, mentions data as one of several descriptors under the broader heading ‘Innovative and Creative, with Critical Judgement’. *Ability to analyse and critically evaluate arguments and evidence appropriate to their disciplines (e.g. collect analyse and interpret data and information, generate and test hypotheses, synthesise and organise information)* *http://www.griffith.edu.au/__data/assets/pdf_file/0004/290659/Critical-evaluation-skills.pdf*
Most universities now have research data support services and infrastructure in place, and an increasing number are providing programs which encourage “Plan to Publish” data management practices from the point of data collection and as part of developing a data management plan. Few university support services, however, make the case for research data as an input into education courses. For this to happen, universities may need to consider educational programs for their own researchers to realise the long-term value and diverse potential uses of their research data. Researchers with strong data management practices and up-to-date skills in the use of emerging tools and facilities that underpin data identification, analysis and visualization, will also have the skills and understandings necessary to make data available for education, business and the broader community.

More could be done at both the university policy level and the individual teacher level to change this.

**Recommendations for universities**

1. That universities develop data sharing and retention policies which recognize the value of data and reflect the expectation that the products of publicly funded research be made available for the betterment of society.
2. That universities review their academic structures and policies, especially intellectual property policies, to take advantage of the synergies between research and teaching to enable greater openness of research data and teaching materials.
3. That universities, and course designers and teachers, consider the need for an increasing focus on the inclusion of teaching with and about research data as a means of developing professional competencies in their graduates and develop graduate attributes to support this focus.
4. That universities provide educational programs for their own researchers to encourage them to consider the long-term value and diverse potential uses of their research data at the point of creation and to provide them with skills in the use of emerging tools and facilities that underpin data discovery and reuse.
5. That university data support services recognise the value of research as an input to teaching and restructure their services to encourage use by students and teachers.
6. That universities ensure that IT infrastructure is adequate to meet the classroom needs of courses which include data handling.

**Data creators**

Consideration of future data use at the point of creation would greatly facilitate reuse for education purposes. Data cannot be reused if it cannot be found, is missing vital metadata or if there are technical or other barriers to access. Data creators can expedite reuse of data in education by determining data management practices which support reuse early in the research project. This is also the time to consider such matters as data formats, required software, the inclusion of metadata to support use by others and other factors that may inhibit subsequent use.

Many teachers in the case studies remarked on the difficulty of accessing data. This means appropriate licensing, preferably using an open access licence rather than providing no licence at all or applying conditions to limit access.\(^{25}\) In addition, data creators and

\(^{25}\) It is also important to recognise that there may be ethical considerations that might genuinely create a need to ensure that the data is not open.
managers can design access systems that encourage data access via a single url or other simple means.

**Recommendations for data creators**

1. That data creators consider future use of data early in their project and at the same time as developing a data management plan and determining licence conditions.

**Data managers and research infrastructure and data providers**

Data managers within universities and other research institutions have responsibility for providing services which can greatly enhance the reuse of data, by students and teachers as well as other researchers. These services could be improved significantly by ensuring:

- the inclusion of metadata to enhance discoverability, access and reuse
- linkages between all outputs of research (software, articles, visualisations, data etc) to promote discoverability, context and all the tools necessary for the data to be reused by students and teachers

Research infrastructure providers are engaged with the research community in developing priorities for infrastructure development and determining how the development will proceed. There would be benefit from consideration of the educational benefit of initiatives at the time of planning and the capacity to include teaching materials that could be used by both researchers and students.

One way of providing benefit to teaching as well as research is through the provision of virtual laboratories. These provide data and tools embedded within a context that provides a broader understanding of the subject area. The Biodiversity and Climate Change Virtual Laboratory (BCCVL) developed by NecTAR is already being used by later year ecology students at Griffith University. Plans are in hand for the Marine Virtual Laboratory (MVL) to be used in the same way. We have already seen that OzTrack is used in both universities and high schools.

Tools developed by research data infrastructure providers are, rightly, designed for research use. There are, however, simpler tools that might be suitable for basic data analysis and presentation but which may not be adequate for advanced research. Examples of existing tools fitting this category include Wordle, NGram and the Nesstar online analysis system which is available from the Australian Data Archive. Such simple tools can be of benefit to researchers, teachers and students.

Data services, such as ALA, have the potential to take on a stronger outreach role, taking a more active role in working with universities to develop documentation, including teaching materials, linked to related datasets. Both AURIN and OzTrack, as mentioned above, already provide access to teaching materials. The Australian Data Archive, used extensively by students, has never had the resources to develop the kinds of instructional materials available from their equivalent organisations in the US or UK. Teaching resources could be developed either by staff of the data service concerned or others could contribute them.

Citizen science projects are known to be popular and provide students with the opportunity to take an active role in the collection of data. At this stage, few in Australia offer opportunities specifically targeted at the tertiary level.

**Recommendations for data managers and infrastructure and data service providers**
(5) That data managers provide metadata to support enhanced discoverability and access and provide linkages between the outputs of research to promote discoverability, context and necessary tools
(6) That data managers and infrastructure and data service providers ensure that access to data is provided with minimum barriers wherever possible
(7) That infrastructure providers consider the educational benefit of proposed tools and services used in research at the time of planning, that provision be made in any proposed infrastructure development for teaching materials to be linked to corresponding data sets and that tools be developed to support both basic needs and as well as advanced research
(8) That infrastructure and data service providers provide discovery and other support services optimized specifically for teachers and students.

**Research funding bodies**

Research funding bodies have an important role in developing policy around research data creation and use. There would be benefit in providing more explicit recognition of the value of research data to education at the point where grants are considered. Applicants are currently asked about the value of their research to the community, but not specifically to education. Impact measures are seen as an important indicator of the value of research but at present do not include recognition of data used for educational purposes.

Funders of research have an influence over the way that research data is made available for future use. They are in a position to encourage the need for data to be made available in as open a manner as possible by the use of appropriate licensing.

**Recommendations for research funding bodies**

1. That explicit recognition of the value of research data to education be given when grants are considered
2. That data usage for educational purposes be recognised as an impact measure,
3. That more encouragement be given for data to be as open as possible, preferably by use of an open licence.

**Government**

Government has a role to play when setting policy directions around research, learning and teaching and to harness the synergies between them. More could be done to recognise the benefits both business and research would derive from a greater availability of skills in data creation, computation, data science, data analysis and visualisation. The need for skills is not limited to information technology graduates but covers entrepreneurs and those with professional qualifications in data-dependent disciplines.

Policy initiatives should encourage the benefits of research to be leveraged into teaching so as to unlock the potential of resources developed for another purpose. Putting more emphasis on research-informed teaching is one way of developing needed skills while at the same time enriching the student learning experience. Practical means of doing this would include changing the remit of National Collaborative Research Infrastructure (NCRIS) to include education as well as research and to develop a facility comparable to the Oceans of Data Institute mentioned above.
Governments, both federal and state, have an important role to play in making data available for research and for teaching and there have been significant advances in the provision of public sector data in recent years.

Recommendations for government

(1) That more consideration be given to the synergies between research and education in the development of policy designed to enhance Australian innovation and competitiveness and the ways in which the benefits of research could be leveraged into teaching

(2) That government policy is developed to ensure that access to public sector data and associated teaching materials is provided for the benefit of both research and teaching

(3) That the remit of NCRIS be amended to incorporate education as well as research

(4) That the government provide funding for the development of a facility comparable to the US Oceans of Data.
Acknowledgements

Thanks are due to all of the following who have helped in various ways with this project.

Staff of the Australian National Data Service
Professor Richard Baker, Pro Vice-Chancellor Student Experience, The Australian National University
Professor Lindsay Botten, Director, National Computational Infrastructure
Diane Costello, Executive Officer, Council of Australian University Librarians
Associate Professor David Curtis, Educational Research, Flinders University
Dr Nathan Downs, School of Agricultural, Computational and Environmental Sciences
Dr Ross Dwyer, School of Biological Sciences, University of Queensland
Dr Dominique Estival, MARKS Institute, University of Western Sydney
Minky Faber, Communications Advisor, Atlas of Living Australia
Associate Professor Christopher Fluke, Centre for Astrophysics and Supercomputing, Swinburne University of Technology
Professor David Goodman, School of Historical and Philosophical Studies, University of Melbourne
Maureen Henninger, IKM and Digital Studies Program, University of Technology, Sydney
Taryn Johnson, Education Officer, Atlas of Living Australia
Dr Penny King, College of Physical and Mathematical Sciences, Australian National University
Professor Janette Lindesay, College of Medicine, Biology and Environment, Australian National University
Alison Makins, Library Compliance Officer, University of New South Wales
Dr Steve McEachern, Director, Australian Data Archive, Australian National University
Dr Stuart Minchin, Chief of Division, Environmental Geoscience Division, Geoscience Australia
Linda O’Brien, Pro-Vice Chancellor, Griffith University
Associate Professor Shannon Notley, College of Physical and Mathematical Sciences, Australian National University
Shane Nuesler, Teaching and Learning Centre, University of Canberra
Jasmin Packer, School of Biological Sciences, University of Adelaide
Associate Professor Chris Pettit, Strategic Implementation Coordinator, Australian Urban Research Infrastructure Network (AURIN).
Dr Roger Proctor, Director, IMOS, University of Tasmania
Dr Joanna Richardson, Division of Information Services, Griffith University
Sue Robbins, Coordinator, Research Services, Information Services, University of Western Sydney
Dr Glen Roe, College of Arts and Social Sciences, Australian National University
Associate Professor Shawn Ross, Ancient History and Archaeology, Macquarie University
Professor Richard Sinnott, Director e-Research, University of Melbourne
Maureen Sullivan, Director, Information Management, Griffith University
Edwina Tanner, University of Sydney Institute of Marine Science
Fiona Tweedie, Research Community Coordinator at University of Melbourne
Professor Deb Verhoefen, School of Communication and Creative Arts, Deakin University
Dr Nigel Ward, Deputy Director for Research Software Infrastructure, NecTAR
Jacquie Wolstenholme, Research Services Librarian, James Cook University
Dr Lyle Winton, Deputy Director, Research Platforms, NecTAR
Malcolm Wolski, Director, eResearch Services & Scholarly Application Development, Griffith University
Dr Lesley Wyborn, National Computational Infrastructure
Appendix 1: Swinburne University - Computer lab: Transferring and storing data.
Permission to include this by Swinburne University is gratefully acknowledged.
Overview

In this Computer Lab, you will undertake three e-Science-related research projects. These projects will give you an opportunity to think about some of the challenges relating to storing and transferring scientific data. You will also reinforce your understanding of the processing capabilities of the typical computing resources you access on a day-to-day basis, compared to the High Performance Computing facilities operated on your behalf at National and International facilities.

Learning Objectives

During this Computer Lab, you will develop your skills in:

- Explaining the limits of computers in storing and representing scientific data
- Applying e-Science strategies and approaches to science discipline-specific requirements
- Problem solving, analysis and tackling unfamiliar problems

Introduction

The following text from Wu & Chin (2014) was used as motivation in the Week 7 lecture, Big Data Challenges:

“The era of big data is approaching as large and complex data sets are being collected for diverse reasons through all kinds of technologies…. The big data sets tend to be more unstructured, distributed and complex than ever before. The mainly concerned characteristics of big data can fall into three dimensions:

(a) the volume of information that systems must ingest, process and disseminate;
(b) the velocity at which information grows or disappears;
(c) the variety in the diversity of data sources and formats.

This situation poses significant challenges on traditional data processing applications and data management tools. The … challenges include collecting, storing, transferring, and visualizing all kinds of big data. More importantly, we need effective ways of turning “big data” into “big insights”…new theories, novel methods and right analytics tools are needed to help scientists … make sense of the volumes of data.”

- This article is the introduction to the first issue of the new Elsevier journal Big Data Research.

In this Computer Lab, you will experiment with storing and transferring data. You will also investigate the processing capabilities of the computers available to you in the Swinburne University computer labs, and compare this to several national and international High Performance Computing (HPC) facilities.

This Computer Lab exercise is deliberately open-ended. One of the goals is for your Lab Group to discuss how the problems might be solved, and then you need to perform some research to test whether the Group was right. It is not expected that all of the e-Science research projects will be completed during
your Lab Session, however, if you choose to write up and submit this Lab Exercise for your Lab Report 2, then you must provide your solutions to all of the research questions.

**Background Reading**

- Lecture 7: *Big Data Challenges*

**Related Online Activity**

You will be making use of the data collected by other students as part of Online Activity 7.1 *The Size of a Song:*

Randomly select 10 songs from your music collection and report the following information, preferably in a table, for each song:

a) Music file format  
b) File size of song (in MB)  
c) Duration of song (in seconds)

**Assessment**

There are two assessment items related to this Computer Lab. The Online Test must be completed during your Computer Lab session.

<table>
<thead>
<tr>
<th>Assessment Item</th>
<th>Description</th>
<th>Marks</th>
<th>Due</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online Test</td>
<td>You will be required to complete a short multiple-choice test relating to material covered in lectures during Week 5, 6 and 7.</td>
<td>30 (3% of final grade)</td>
<td>End of your Computer Lab</td>
</tr>
<tr>
<td>Laboratory Report</td>
<td>You must select either Computer Laboratory 3 or Computer Laboratory 4 and write up your results for submission and assessment.</td>
<td>100 (10% of final grade if chosen for submission)</td>
<td>5pm, Friday 31 October 2014</td>
</tr>
</tbody>
</table>

Full instructions on the requirements for the Laboratory Report, including the required Cover Sheet, assessment criteria, and grading rubric, may be found on the ICT10007 Blackboard site.
Laboratory Exercise

When conducting any laboratory work, you should take sufficiently clear notes of what you did and make sure you have answered all questions. You should work under the assumption that this is the Laboratory Exercise that you will write up and submit for assessment...even if you do go on and choose Laboratory Exercise 4.

<table>
<thead>
<tr>
<th>Task</th>
<th>Description</th>
<th>Approx. duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>Arrive, power up the computers and login!</strong></td>
<td>5 minutes</td>
</tr>
<tr>
<td></td>
<td>Read through the Laboratory Exercise and familiarize yourself with the</td>
<td></td>
</tr>
<tr>
<td></td>
<td>location of the Background Reading resources (if you have not already done</td>
<td></td>
</tr>
<tr>
<td></td>
<td>so).</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Your Lab Demonstrator will spend a few minutes talking over what you will</td>
<td></td>
</tr>
<tr>
<td></td>
<td>do during this Laboratory Exercise.</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td><strong>e-Science Research Projects</strong></td>
<td>5 minutes</td>
</tr>
<tr>
<td></td>
<td>As a group, discuss your initial thoughts on the 3 e-Science research</td>
<td></td>
</tr>
<tr>
<td></td>
<td>projects:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A. Characterise your Computer</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B. The Cost of Transferring Data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C. The Consequences of Storing Data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• How might you approach finding solutions to each of the problems?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• As a class, decide on how much time you will devote to complete these</td>
<td></td>
</tr>
<tr>
<td></td>
<td>projects during your lab session. You need to leave yourself with 30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>minutes at the end of the lab session to complete the online Computer Lab</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Test 3.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• At the conclusion of each project, discuss as a class what progress you</td>
<td></td>
</tr>
<tr>
<td></td>
<td>made, what you learnt, and what the next steps are.</td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>Approx. duration</td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>-----------------</td>
<td></td>
</tr>
<tr>
<td><strong>A1</strong></td>
<td><strong>A. Characterise your Computer</strong></td>
<td></td>
</tr>
<tr>
<td><em>As an e-Scientist, the Computer is your main research tool, so it is important that you are able to determine, interpret and understand its most important processing characteristics.</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Complete the entries in Table 1 for the computer you are using to complete this Computer Lab Exercise.</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>The line speed should be determined using the Oz Broadband Speed test:</em></td>
<td>10 minutes</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.ozspeedtest.com">http://www.ozspeedtest.com</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>[Note: If you are not able to run the Oz Broadband Speed test yourself, then use the line speed and values provided in Lecture 7 for the test run on 17 September 2014. Choose either the Cable or WiFi result.]</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Record how to determine each of the properties in Table 1 for your computer (this should be included in your Method if you write up this Exercise as Computer Laboratory Report 2).</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Note: As a group, you should discuss how to find out each of the characteristics of your computer.</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>A2</strong></td>
<td><strong>High Performance Computing</strong></td>
<td></td>
</tr>
<tr>
<td><em>Fill in the details in Table 2, and compare the computer you are using to one facility (your choice) from each of the following three categories:</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>1. An Australian national computing facility</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>2. A computer on the Top 500 Supercomputer List</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>3. A computer on the Top 500 Green Supercomputer List</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Question:</strong> Why do e-Scientists need access to High Performance Computing (HPC) facilities?</td>
<td>10 minutes</td>
<td></td>
</tr>
</tbody>
</table>
### B1. The Costs of Transferring Data

As data sets grow in size, it becomes increasingly difficult to move them from one location to another. Transferring data across the Internet costs money, even though this is often hidden from researchers (and students!). Transferring large quantities of data across the Internet takes time.

**Question 1:** So, is it **quicker** and **cheaper** to send data on a hard drive via Australia Post rather than to transfer it over the network?

**Question 2:** And how does this depend on the properties of the network, the volume of data to be transferred, and the transfer speeds?

To answer these two Questions, you should consider the following three scenarios:

1. A user who is only able to transfer data at the speed of the computer you characterised in Task A using a standard ADSL broadband plan.
2. A user who has access to the fastest possible National Broadband Network (NBN) download speeds (assume a typical NBN Tier 5 bundle price and the maximum available download speed). You should assume that the NBN roll out is complete and fully accessible from any destination! [You may wish to check if any of the locations are currently part of the NBN]
3. A user who has access to the fastest National network speeds available through AARNET (assume 100 Gbps and $0.05 per GB).

For each of these users, you should consider the three following data volumes:

1. 420 Gigabytes
2. 4.2 Terabytes
3. 42 Terabytes

Finally, for each of the combinations above, you should consider the following locations as the source of the data that is being transferred to Swinburne University:

1. CSIRO, Highett, Victoria
2. Parkes Observatory, New South Wales
3. The Pawsey Centre, Western Australia

That means there are \(4 \times 3 \times 3 = 36\) possible scenarios to investigate.

Before you commence, discuss with the rest of your class how you will distribute the work, and any assumptions you are going to make.

[**Hint:** Don’t forget you will need to “buy” something to store your data on and a box to safely pack it inside if you intend to use Australia Post. You can
assume that the source and destination institutions can load and unload data at no additional cost.]

You may find the following web-sites to be helpful:


[Note: You will need to type these addresses in – copy and paste to your browser might not work]

Use Table 3 to record the class results (make sure you report your individual results to the Lab Demonstrator).

Once you have calculated/obtained all of the required data for the two methods of transferring data (network vs. Australia Post):

- Compare and contrast the costs associated with transferring the data.
- Compare and contrast the time taken to transfer the data.
- What conclusions do you draw about the two methods of transferring data?

<table>
<thead>
<tr>
<th>Task</th>
<th>Description</th>
<th>Approx. duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td><strong>C. The Consequences of Storing Data</strong></td>
<td>30 minutes</td>
</tr>
</tbody>
</table>

Using the data collected from Online Activity 7.1, determine a relationship between the duration of songs and their file sizes.

**Questions:** How does this depend on the music file format?

Analyse and characterize the song file data, identifying and indicating any steps you took to inspect, validate, or prepare the data.

As a class, you should decide on how you will extract and share the data from the Discussion Groups, and which data characteristics you need to calculate.

**Note:** This investigation must be based on at least 100 songs (in total) and at least two different music file formats.

**Question:** Based on your results, which is the preferred music file format to use in order to maximize the number of files that can be stored on a file-system with a fixed size?
**Table 1.** Processing characteristics of your computer. Complete this table for Task A1.

If you use any additional computers to complete any part of this Computer Lab Exercise, then please be sure and characterise those computers as well.

You can extend this table to include other relevant properties, such as the type of Graphics card, bit-length of the operating system (usually 32-bit or 64-bit), etc.

<table>
<thead>
<tr>
<th>Property</th>
<th>Quantity</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operating System</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Processing speed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>System Memory</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disk space (Total, Used, Available)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethernet speed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Line speed (measured)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Download speed (measured)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upload speed (measured)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2.** High Performance Computing Facilities. Complete this table for Task A2 (and extension Task A3 – see below).

Complete this table for each of the High Performance Computers you chose to investigate from the three categories:

1. Australian National High Performance Computing Facility
2. A computer on the Top500 Supercomputer
3. A computer on the Green500 Supercomputer

<table>
<thead>
<tr>
<th>Name</th>
<th>Host Institution</th>
<th>Web-site</th>
<th>Description [Include position on Top500/Green500 as relevant]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak processing speed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total storage space</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total system memory</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other [Include any other relevant system characteristics]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Table 3.** The cost of transferring data. Use this table to record the results for Task B1 for your Lab Group.

**CSIRO, Highett, Victoria to Swinburne**

<table>
<thead>
<tr>
<th></th>
<th>Australia Post</th>
<th>Local network user</th>
<th>NBN user</th>
<th>AARNET user</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cost</td>
<td>Time</td>
<td>Cost</td>
<td>Time</td>
</tr>
<tr>
<td>420 GB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.2 TB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42 TB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Parkes Observatory, New South Wales to Swinburne**

<table>
<thead>
<tr>
<th></th>
<th>Australia Post</th>
<th>Local network user</th>
<th>NBN user</th>
<th>AARNET user</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cost</td>
<td>Time</td>
<td>Cost</td>
<td>Time</td>
</tr>
<tr>
<td>420 GB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.2 TB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42 TB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Pawsey Centre, Western Australia to Swinburne**

<table>
<thead>
<tr>
<th></th>
<th>Australia Post</th>
<th>Local network user</th>
<th>NBN user</th>
<th>AARNET user</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cost</td>
<td>Time</td>
<td>Cost</td>
<td>Time</td>
</tr>
<tr>
<td>420 GB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.2 TB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42 TB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 4.** Are you costing the University money? Complete this table for extension Task B2 (see below).

Complete this table classifying the type of network connection for the 10 (legal!) web site you access most regularly while using Swinburne computing resources.

<table>
<thead>
<tr>
<th>Number</th>
<th>Network address</th>
<th>Domestic or International?</th>
<th>On-net, off-net or peering?</th>
</tr>
</thead>
<tbody>
<tr>
<td>E.g.</td>
<td><a href="http://www.google.com">www.google.com</a></td>
<td>Domestic</td>
<td>Peering</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Laboratory Report

Parts of the tasks that you do not complete during the Laboratory Session can be completed in your own time, particularly if you intend to submit this as your assessable Laboratory Report 2.

The following extensions tasks are to be completed if you choose to submit this Lab Exercise as Computer Laboratory Report 2.

<table>
<thead>
<tr>
<th>Description</th>
<th>A3: Additional Characterisation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1. Complete Task A1 for at least one other computer and compare the characteristics of the two computers.</td>
</tr>
<tr>
<td></td>
<td>2. Complete Task A2 for a second computer from each of the three categories. Compare and contrast the properties of the 6 HPC facilities you have identified.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Description</th>
<th>B2: Are you costing the University money?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>As an institution that uses AARNET, Swinburne is responsible for paying for your network usage. AARNET distinguishes between three main types of web-sites:</td>
</tr>
<tr>
<td></td>
<td>• On-net (domestic and international)</td>
</tr>
<tr>
<td></td>
<td>• Off-net (domestic and international)</td>
</tr>
<tr>
<td></td>
<td>• Peering (domestic and international)</td>
</tr>
</tbody>
</table>

**Question:** What does each of the three classifications mean?

AARNET provides a tool for you to check the type of connection available for a chosen network address: [http://lg.aarnet.edu.au/cgi-bin/traffic.cgi](http://lg.aarnet.edu.au/cgi-bin/traffic.cgi)

Choose the 10 (legal!) web sites that you access most regularly while using Swinburne computing resources. Enter the network address (host URL) without the http component, and fill in the entries in Table 4.

**Example:**
Enter network address: [www.google.com](http://www.google.com)
74.125.237.180 is **domestic peering** network 74.125.237.0/24 originated by AS15169 via path AS15169

**Question:** Based on your results, are you a responsible user of Swinburne’s network resources?

<table>
<thead>
<tr>
<th>Description</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1. Investigate the methods used by at least two of the music file formats to store the music data. What compromises do these music file formats make between sound quality and file-size?</td>
</tr>
<tr>
<td></td>
<td>2. Now suppose that these were data files relating to an experiment conducted in your scientific discipline. What compromises do scientists have to make when choosing file formats, and what are the potential consequences of a poor choice?</td>
</tr>
</tbody>
</table>
If you choose this Laboratory Session to write up for assessment as Laboratory Report 2 (due 5pm, Friday 31 October 2014), please ensure your write-up includes the following items:

**Cover Sheet**
- Please download and fill in relevant sections of the Laboratory Report Cover Sheet

**Background**
- What are the three Vs of Big Data?
- What are the main ICT challenges for Big Data?
- What are the following projects and what role do they play in supporting Australian e-Science projects?
  - Australia’s Academic and Research Network (AARNET)
  - Australian National Data Service (ANDS)
  - Research Data Australia
  - The Research Data Storage Infrastructure (RDSI) project

**Method**
- Present the sequence of steps you undertook to complete each of the Tasks. Do not just repeat word-for-word what is listed in the Task sections.

**Results**
- Answer all of the questions throughout the Laboratory Exercise, including filling in all of the entries for Tables 1-4.

**Discussion**
- What have you learnt about the challenges associated with processing, storing and transferring data for “Big Data” e-Science projects?

**Conclusion**
- What are the most important results you have obtained from completing this Laboratory Exercise?

**References**
- Make sure you include references to any resources that you used to complete this Laboratory Exercise. References should be formatted using the Swinburne Harvard style.

**Hints**
You should aim to incorporate the questions and answers directly into your writing, rather than just presenting them as questions and answers! For example:

**Good:**
There were some clear strengths and weaknesses between using R and Software X. While it was easier overall to calculate the “kurskewdian” with SoftwareX because of the in-built `kurskewdian()` function, R gave a choice of using both the “kurskew” and the “skewdian” algorithms independently.

**Not so good:**
**Question.** Compare your experiences in using R with one other data analysis program?
**Answer.** Strengths of R: good for “kurskew” and “skewdian”. Weaknesses: not good for “kurskewdian”.