

10 astronomy research data Things

Use

Repurpose

Adapt

Change

10 astronomy research data Things is a self-paced learning program that provides an opportunity to explore issues surrounding management of research data, specifically for people working with astronomy data.

This program was developed from the 23 (research data) Things program and the extensive [ANDS resources and materials](#) related to research data management and re-use.



Australian National Data Service

NCRIS
National Research
Infrastructure for Australia
An Australian Government Initiative



Table of Contents

10 Astronomy Research Data Things.....	4
Why do you need to manage your research data?	4
How can I work through these Things?	4
Thing 1: Getting started with research data	5
What is research data?	5
Data in the research lifecycle	5
How data differs across disciplines.....	6
Thing 2: Issues in research data management	7
Research Data Management in Practice	7
How do you manage “Big Data”?	7
Thing 3: Data Sharing	9
Introduction to ‘open’, ‘shared’, and ‘closed’ data	9
Data sharing practices	9
FAIR data.....	10
Thing 4: What are publishers & funders saying about data.....	12
Learn about journal data policies	12
Data Journals.....	12
Data sharing policies of funders	13
Thing 5: Identifiers for data and people	14
DOI’s are unique (just like you).....	14
Getting to know ORCID	15
Thing 6: Describing data: metadata and controlled vocabularies.....	17
Metadata: your new best friend	17
Controlled vocabularies for data description.....	17
Thing 7: Data citation and publishing	19
The cans and cannots of licensing	19
Licensing for data reuse	20
Getting more out of your citation.....	20
Thing 8: Data Management Plans	21
An introduction to Data Management Plans	21
Templates for Data Management Plans	21
Thing 9: Tools of the trade	22
Dirty data	22

Extracting and scraping data	22
An introduction to VO.....	23
Thing 10: What's in a name?	24
Who's who in the acronym soup?	24
Big data: the sky is not the limit!	25
What's next?	26

10 Astronomy Research Data Things

Why do you need to manage your research data?

Effective management of research data is increasingly recognised as a critical part of the research process. It enables:

- Trust in data you obtain for reuse from other sources
- Reproducibility of research through increasing veracity of data
- Increased quality of your research
- Strengthening of researchers' reputation through increased citations and reach of all research outputs
- Increased connectivity between all research outputs, and researchers
- More efficient use of scarce research funds
- Data description for sharing and collaboration
- Reduced risk of loss or corruption of data

How can I work through these Things?

- All Things have 1 to 3 Activities. You can pick and mix from the Activities to suit your interests.
- You can do as much or as little of the Things and Activities as you want to do, or need to know.
- Some of the Activities are intended as an introduction to a topic, and some delve a little deeper. Choose what interests you and suits your experience.
- You can work through Activities on your own at your own pace, or in a group.

Ideas to reuse and repurpose these activities

This material is licenced with a CC-BY licence, meaning that you can use, repurpose, adapt, or change it to suit your needs.

This program was adapted from the ANDS 23 (research data) Things as well as the re-purposed 10 Medical and Health Things and 10 Marine Science Things.

Please note: this is a snapshot in time: research data as it was in 2017 - you may need to check resources and update resources and links to include more recent initiatives and policy changes.

Thing 1: Getting started with research data

Research data comes in many shapes and sizes and its management changes over time. Kick off your research data journey by exploring different types and forms of research data and how they fit into the research lifecycle.

Activity 1

What is research data?

What "research data" are we talking about?

1. Read an [Introduction to Research Data](#) from Boston University
2. As we have just seen, research data can come in many forms. Some of these are human readable, and some are machine readable. Explore a couple of these types of formats commonly seen in astronomy:
 - a. Observational astronomy images, e.g. [Hubble archive](#)
 - b. Radio Astronomy data, e.g. [NRAO](#)
 - c. Spectral data, e.g. [6dF galaxy survey](#)
 - d. Value-added survey data catalogues, e.g. [SDSS](#)
 - e. Simulations, e.g. [TAO](#)

Consider: make a list of the forms of data you have used or seen in your work. What would people need to know about these data if they wanted to re-use these data?

Activity 2

Data in the research lifecycle

Data often have a longer lifespan than the research project that creates them. Follow-up projects may analyse or add to the data, and data may be reused by other researchers.

A data lifecycle shows the different phases a dataset goes through as the research project moves from "having a brilliant idea" to "making ground-breaking discoveries" to "telling the world about it"

1. Take a look at either one of the links below:
 - a. [UK Data Archive Research Data Lifecycle](#) (if you are new to this concept)
 - b. [DCC Curation Lifecycle Model](#) (if you are familiar with this concept)
2. What is your university's policy on data management and data sharing, how long do they require the data to be stored?

Consider: have you been through all of the steps outlined in this lifecycle? If not, which ones are new to you?

Activity 3

How data differs across disciplines

Choose one of the three specialised data repositories below, or find another data repository of interest - particularly one in a discipline you are unfamiliar with, and spend some time browsing around your chosen repository to get a feel for the data available

1. The [All Sky Virtual Observatory](#) (ASVO) is a growing collection of theoretical and observational datasets
2. [Australian Data Archive](#) (this archive contains Social Science, Historical, Indigenous, Longitudinal, Qualitative, Crime & Justice and International data)
3. [RCSB Protein Data Bank](#) (A Structural View of Biology)

Think about how the data here differs from data you are familiar with. Consider for example, format, size and access method.

Consider how cross disciplinary research could be affected by discipline data conventions, and also one way cross disciplinary data access can be facilitated.

Thing 2: Issues in research data management

Research data is critical to solving the big questions of our time. So what are some of the issues we face in managing research data?

Activity 1

Research Data Management in Practice

Researchers have responsibilities with regards to managing their research data. Governments and universities all around Australia and the world are now encouraging researchers to better manage their data so others can use it. Research data might be critical to solving the big questions of our time, but so much data is being lost or poorly managed.

1. Review the [Policy Statement on F.A.I.R. Access to Australia's Research Outputs](#). What does F.A.I.R. mean?
2. This 4.40mins [cartoon](#) put together by the New York University Health Sciences Library, is about what happens when a researcher hasn't managed their data (at all...). As you watch the cartoon, jot down the data management mistakes made by the researcher.
3. Scan through this guide to [Research Data Management in Practice \(PDF, 0.74 MB\)](#). Look carefully at Figure 1 Key Steps in Research Data Management, Section 3: Steps in Research Data Management.

Consider how just ONE of the data disasters depicted in the cartoon could have been avoided.

Activity 2

How do you manage “Big Data”?

"Big Data" is a term we're hearing about with increasing frequency. Data management for Big Data brings much complexity - citing dynamic data, software, high volume compute, storage costs, transfer of petabytes of data, preservation, provenance, more.

1. Read [this post](#) and presentation titled: "*Big Data: The 5Vs Everyone Must Know*". This article uses 5V's: volume, variety, velocity, veracity and value as a concept for how big data can be managed more successfully.
2. **Consider** whether the concept of 5Vs is useful to support better management and reuse of marine science “Big Data”
3. The [Pawsey Supercomputing Centre](#) located in Perth, Western Australia supports researchers across Australia with an array of capabilities encompassing

supercomputing, data and visualisation services. Review some of the interesting [research projects](#) that involve the use of big data undertaken at the centre.

4. Read more about the [data storage](#) services available at Pawsey.

Thing 3: Data Sharing

Data may be shared in many ways. Here are ways that data can be shared and is currently being shared.

Activity 1

Introduction to 'open', 'shared', and 'closed' data

Repositories enable discovery of data by publishing data descriptions ("metadata") about the data they hold - like a library catalogue describes individual materials held in a library. Most repositories provide access to the data itself, but not always. Data portals or aggregators draw together research data records from a number of repositories, e.g. [Research Data Australia \(RDA\)](#) aggregates records from over 100 Australian research repositories.

1. Watch this 2.5 minute [video](#) from the Open Data Institute titled Open/Closed/Shared: the world of data.
2. Now open [this ANDS open data webpage](#) to see a more in-depth view of why data is sometimes open, shared or closed.
3. If you have time, go to [Research Data Australia](#) portal and try searching for data that is 'open'. Hint: Look for the option to limit your search to data that is **Publicly accessible online**.

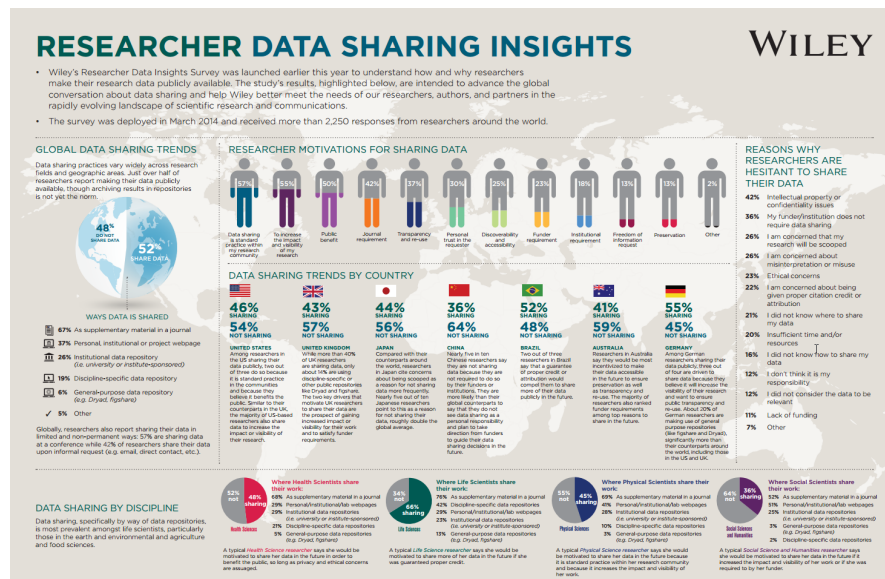
Consider: Why more data isn't publicly accessible or more 'open'? What are the policies or 'customs' in your field? Do you consider your data to be open?

Activity 2

Data sharing practices

Repositories are one means by which research data may be shared but in order to get data into repositories, research teams must be willing to publish their data: there are huge differences between data sharing practices by country and by discipline.

1. Take a look at this 2014 infographic from Wiley titled [Research Data Sharing Insights](#) [PDF, 2.08MB]. It provides a succinct overview of current data sharing practice and perceptions.



Research Data Sharing Insights (Wiley, 2014)

- Now look closely at the sections titled 'Global Data Sharing Trends' and 'Data Sharing By Discipline'.
- Have a look at this [article](#): How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers (PLoS One. 2014; 9(8): e104798 doi: 10.1371/journal.pone.0104798)

Consider: Why do you think there are differences between disciplines and countries - what changes to these statistics would you expect between 2014 and now?

Activity 3

FAIR data

To be able to effectively share data the FAIR principle should be employed:

- Findable
- Accessible
- Interoperable
- Reusable

Astronomy is already doing a good job with this as there are, across many sub-disciplines:

- Common data formats since the 70s, most notably the FITS file system
- Strong tradition of international collaboration
- Open data in general (often after a proprietary period)
- Driven by community needs (on-line observation archives, on-line services)

1. Have a look at this overview on FAIR data <https://www.dtls.nl/fair-data/fair-principles-explained/>
2. See also the FAIR Guiding Principles for scientific data management and stewardship <http://www.nature.com/articles/sdata201618>
3. As an example, explore the All-Sky Virtual Observatory ([ASVO](#)) and its growing collection of theoretical and observational datasets. How do the different nodes present the data, do they follow the FAIR principles?
4. If you have the time, explore the Documents and Standards section on the International Virtual Observatory Alliance (IVOA) [webpage](#). Consider how well a FAIR data standard can be achieved using these guidelines.

Consider: Are you planning to share your data? Is it following the FAIR principles?

Thing 4: What are publishers & funders saying about data

Data sharing policies are becoming increasingly common in Australia and internationally. Learn why research funders and journal publishers are particularly influential when it comes to encouraging data availability.

Activity 1

Learn about journal data policies

More and more journal publishers are asking authors to make the data underpinning a journal article available. It's all about ensuring that the research being described in the article is based on solid, reproducible science. Thinking back to [Thing 3: Data Sharing](#), remember that “available” can be “open” or “shared” through mediated access.

Have a look at the resources below:

1. Data policies of journal publishers: [SpringerNature](#), [A&A](#), [MNRAS](#)
2. [Figshare](#) and [Dryad](#) are data repositories which integrate data and articles. They facilitate submission of your research data to journals.
3. Look up a journal you know and see what the advice the journal gives on related data.

Consider: How easy, or hard, it was for you to understand what you had to do in regard to research data?

Activity 2

Data Journals

Explore this relatively new form of data publishing: the data journal. Data journals focus on data, rather than discuss an analysis of the data (as in traditional journals).

1. Read this short introduction: [What are data journals?](#)
2. Browse this [data paper](#) published in the data journal Scientific Data.
 - a. Note the extensive exposure of the data through maps, links to full tables, and diagrams etc. and how to cite this article.

Consider: Why do you think authors might choose to share their data in data journals rather than, or in addition to, traditional journal formats?

Activity 3

Data sharing policies of funders

The Australian Research Council (ARC) provides funding to various research programs in Australia.

1. Take a look at the [Australian Research Council \(ARC\) requirements](#) on research data management.
2. Review the [ANDS Guide to filling in the data management section in ARC grant applications](#). What are some of the aspects of research data management a researcher can look into when describing how to manage his/her datasets when applying for a grant?

Also consider what the data sharing and publishing policies of different telescopes are:

1. Look at point 9 of the Australian Time Allocation Committee (ATAC) [Policies and Procedures](#) for proprietary times on AAT data
2. Hubble Telescope data [proprietary rights](#)
3. [ESO data access policy](#)
4. [Gemini Observatory archive](#)
5. [NRAO data access](#)

Consider: Will you be using telescope data or simulations? What are the data access policies of your chosen data provider? What part of your data are you required to make public, the raw information or your final data product?

Thing 5: Identifiers for data and people

What are DOIs and ORCIDs? These unique identifiers support data citation, metrics for data and related research objects, disambiguation of people, accurate attribution and impact metrics.

Activity 1

DOI's are unique (just like you)

Digital Object Identifiers (DOIs) are unique identifiers that provide persistent access to published articles, datasets, software versions and a range of other research inputs and outputs. There are over 120 million Digital Object Identifiers (DOIs) in use, and last year DOIs were “resolved” (clicked on) over 5 billion times!

Each DOI is unique but a typical DOI looks like this: <http://doi.org/10.4225/08/50F62E0D359D5>

DOIs can be used to collect citation metrics about the use of a dataset or article.

1. Start by watching this short 4.5min video [Persistent identifiers and data citation explained](#) from Research Data Netherlands. It gives you a succinct, clear explanation of how DOIs underpin data citation.
2. Have a look at the poster [Building a culture of data citation](#) (also shown below) - follow the arrows to see how DOIs are attached to data sets.



3. Let's go to a data record which shows how DOIs are used. Click on this DOI to 'resolve' the DOI and take us to the record: <http://dx.doi.org/10.4225/22/55BAE9DBD9670> (Population Health data collection for the City of Greater Bendigo)

The same record has been [syndicated to Research Data Australia](#). Click on the DOI at the bottom of the page, under 'Identifiers'. No matter where the DOI appears it always resolves back to its original dataset record to avoid duplication. i.e. many records, one copy.

If you have time: Want to know more about DOIs? Flick through the [ANDS DOI Guide page](#).

Consider: should DOIs be routinely applied to all research outputs? Remember that DOIs carry an expectation of persistence (maintenance costs etc.) but can be used to collect metrics as well as link articles and data (evidence of impact).

Activity 2

Getting to know ORCID

What about identifiers for people? Think about the many forms a person's name may take or common names. Is the author JK Rowling the same person as Joanne Rowling and Jo Rowling? More than 38,000 Americans have the name James Smith!

Universities, funders and publishers worldwide now use ORCID to differentiate between people with the same name by assigning individuals with a unique identifier.

1. Let's start by going to [ORCID](https://orcid.org). In the search box at the very top of the page, enter *John David Burton* to search the ORCID registry. Scan the list of results to find the entry for John David Burton. How many versions of his name do you see?

2. Now enter *Toby Burrows* into the search box. Open his ORCID record to see a wonderful example of a rich ORCID record. Note he has combined his ResearcherID and his Scopus Author ID with his ORCID. Scroll through his list of works and look closely at *Source* to see the wide range of sources of his publications.

You can now choose from 3 activities that will get you in touch with ORCID.

Option 1. Don't have an ORCID record but would like one?

Use this time to create your ORCID profile and make it as complete as possible.

1. Visit [ORCID](https://orcid.org) and follow steps 1 and 2.
2. When you're done, add your ORCID iD to your email signature, LinkedIn profile and blog
3. Send your new ORCID iD to a colleague and ask for some feedback on your profile

Option 2. Already have an ORCID?

When was the last time you logged in to update or enhance your profile? You may be surprised at the additional functionality now available.

1. Read Alice Meadow's blog post [Six Things to do now you have an ORCID iD](#).
2. Now go to your ORCID profile and update it to be as current and complete as possible
3. When you're done, add your ORCID to your email signature, LinkedIn profile and blog

4. Consider using the new [QR code feature](#) for your ORCID iD in new and uncharted ways

Option 3. Don't want an ORCID?

Get up to date with the latest features, functionality and news on the [ORCID blog](#) and explore the [Australian ORCID Consortium](#) (most Australian universities are members).

Consider how ORCID can be used to enhance your online profile.

Thing 6: Describing data: metadata and controlled vocabularies

Metadata elements are essential for finding and reusing research data. Data is only as valuable as the metadata which describes and connects it. In addition to selecting a metadata standard or schema, whenever possible you should also use a controlled vocabulary. A controlled vocabulary provides a consistent way to describe data.

Activity 1

Metadata: your new best friend

Metadata is structured information about a resource that describes characteristics such as content, quality, format, location and contact information. Creating metadata to describe research data is very similar to the process for descriptive cataloguing of library resources.

Metadata schema are sets of metadata elements (or fields) for describing a particular type of information resource. Numerous metadata schema exist for describing research data across different disciplines.

1. Read the short ANDS [Introduction to Metadata](#) to understand what metadata is and why is it the lifeblood of research data sharing!
2. Let's have a look at the [second data release](#) of the Galaxy And Mass Assembly (GAMA) survey. Have a closer look at one of the tables found under the [schema browser](#). **Consider** both the type and quality of information provided. What metadata included in this record help discovery and reuse of the data?
3. Explore the UK Digital Curation Center's [Directory of Disciplinary Metadata](#). You might find a schema that is applicable to your research!

Consider: Why, if metadata is the lifeblood of data discoverability and reuse, is it often neglected or not richly done when data is published?

Activity 2

Controlled vocabularies for data description

In addition to selecting a metadata standard or schema, whenever possible you should also use a controlled vocabulary. A controlled vocabulary provides a consistent way to describe data - location, time, place name, subject.

Controlled vocabularies significantly improve data discovery. It makes data more shareable with researchers in the same discipline because everyone is 'talking the same language' when searching for specific data.

1. We are going to see some controlled vocabularies in action in the Atlas of Living Australia (ALA).
 - a. Click [here](#) to do a search on the ALA datasets: type "whale" in the search box. Choose your favourite whale species and click on the (red text) **View record** link.
 - b. Any metadata field where you see *Supplied...* tells you that the information supplied by the person who submitted the record (often a 'citizen scientist') has been changed to the controlled vocabulary being used in metadata fields e.g. Observer, Record date and Common name.
2. Now have a look at the IVOA Unified Content Descriptors (UCD) [documentation](#). These UCD are used to describe content fields for data found in the virtual observatories. If you look back at the schema used by GAMA (previous activity) you will see that they also follow these descriptors.
 - a. You can build your own UCD for your data [here](#).

Consider: How do you think we could encourage people to use controlled vocabularies in their data descriptions?

Thing 7: Data citation and publishing


Understand the importance of data licensing, learn about Creative Commons and find out where data fits in the citation picture.

Activity 1






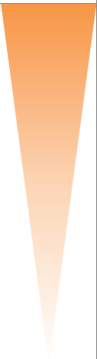




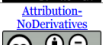





The cans and cannots of licensing

Consider this scenario: You've found a dataset you are interested in. You've downloaded it. Excellent! But do you know what you can and cannot do with the data? The answer lies in data licensing. Licensing is critical to enabling data to be reused and cited.

1. Start by reading this [brief introduction](#) to licensing research data.
2. Now have a closer look at the [poster](#) from creativecommons.org.au. Click on the descriptions for more information. Notice they have used CC BY as the licensing information at the bottom of the poster so you know what you can do with the poster itself.




Know Your Rights: Understanding CC Licences

Licence	Licence conditions	Author can:	User can:	User can:	User can:	User can:	User can:	HowOpenIsIt?
		<ul style="list-style-type: none"> generally retain copyright grant a non-exclusive licence enter into other publishing agreements archive in an institutional repository, subject archive or personal website 	quote and cite in research	Share copies of articles with attribution	User can: create modified versions including abridgments, annotated versions, excerpts and figures	Redistribute commercially	Release modified versions under terms of their choosing including CC licence	
		✓	✓	✓	✓	✓	✓	
		✓	✓	✓	✓	✓	✗	
		✓	✓	✓	✓	✗	✓	
		✓	✓	✓	✓	✗	✗	
		✓	✓	✓	✗	✓	✗	
		✓	✓	✓	✗	✗	✗	
All Rights Reserved		✗	✓*	✗	✗	✗	✗	

*"HowOpenIsIt" is a trademark and has been used with permission. The spectrum is used in this context to illustrate how open-ness is enabled by CC licences. "HowOpenIsIt? Open Access Spectrum" (c) 2014 SPARC and PLOS, licensed [CC BY](#)

✓* limited by scope of available copyright exceptions

 Find this poster on the ccAustralia website at <http://creativecommons.org.au/know-your-rights/>. Unless otherwise noted, this material is licensed under a Creative Commons Attribution 4.0 licence. You are free to copy, communicate and adapt the work, so long as you attribute Creative Commons Australia.

3. Browse through the (21) [slides](#) from a presentation ANDS gave in June 2016 about licensing data.

Activity 2

Licensing for data reuse

Enabling reuse of data can speed up research and innovation. Licensing is critical to enabling data reuse.

1. Start by watching this 4.30mins [video](#) in which Dr Kevin Cullen from the University of New South Wales explains their approach to licensing which aims to strengthen the University's relationship with business and industry.
2. Now read the [Australian Government Public Data Policy Statement](#) (2 pages) that was released in December 2015. Note in particular, the last dot point.
3. If you have questions, ANDS has a list of [research data licensing FAQs](#).

Consider: if you would use the Creative Commons Non-Derivative license for your data.

Some research institutions and research funders now require researchers to submit a Data Management Plan (DMP) for new projects. What should a DMP cover? Could you help with one?

Activity 3

Getting more out of your citation

Data citation continues the tradition of acknowledging other people's work and ideas. Along with books, journals and other scholarly works, it is now possible to formally cite research datasets and even the software that was used to create or analyse the data.

1. Scan through the ANDS introduction to [data citation](#).
2. Then have a look at the Force11 Declaration of [Data Citation Principles](#).
3. Now have a think about how astronomy datasets are typically cited.
 - a. Have a look at the Hubble legacy archive [policy](#), or the ESO [policy](#).
Do other surveys or simulations work in a different way?

Consider: Data citation is a relatively new concept in the scholarly landscape and as yet, is not routinely done by researchers, or expected by most journals. What could be done to encourage routine citation of research data and software associated with research outputs?

Thing 8: Data Management Plans

Some research institutions and research funders now require researchers to submit a Data Management Plan (DMP) for new projects. What should a DMP cover? Could you write or help with one?

Activity 1

An introduction to Data Management Plans

A Data Management Plan (DMP) documents how data will be managed, stored and shared during and after a research project. Some research funders are now requesting that researchers submit a DMP as part of their project proposal.

1. Start by scanning this short introduction to [introduction to Data Management Plans](#)
2. Take a look at the Monash University [data planning](#) website and review what are some of the benefits of data planning highlighted.
3. Now browse through some public DMPs from either the CDL or DataOne, and open up one or two of the DMPs to see the type of information they capture:
 - a. [California Digital Library](#)
 - b. [DataOne](#)

Consider: You will have noticed that DMPs can be very short, or extremely long and complex. What do you think are the 2 or 3 pieces of information essential to include in every DMP and why you chose those?

Activity 2

Templates for Data Management Plans

Preparing a Data Management Plan (DMP) can be complex and should be done at the start of the project. This makes sure that all the elements are in place by the time the research team needs to publish their data. DMP templates are now freely available for reuse by other institutions.

1. See more examples of DMP template or guides from the bottom of the [ANDS DMP webpage](#)
2. Choose one DMP template or guide from the Australian, International or Discipline examples at the bottom of the [ANDS DMP page](#)
3. Spend 5-10 minutes starting to complete the template, based on a research project you have been involved with in the past.

Consider: what are the strengths and weaknesses of your chosen template.

Thing 9: Tools of the trade

Dig in to dirty data. What is it? Why should we care? Try your hand at using an open source data cleansing tool.

Activity 1

Dirty data

Data horror stories: how does it happen!?

Why is “clean” data important? Public policy, changes to medical protocols and economic decisions all depend on accurate and complete data. Thing 9 looks at the why and what of “dirty data.”

1. Browse down the [Bad Data Guide](#) list of commonly encountered data quality issues (with possible solutions). This list is aimed at journalists but it shows who is responsible for cleaning up dirty data.
Click into a few of the causes and solutions to dirty data - many of us contribute information to reports or do our home accounts in spreadsheets, and maybe it's time to think about how clean our own data is!
2. Check out one of the School of Data's [Data Fundamentals course \(a Gentle Introduction to Data Cleaning\)](#).

Consider: What are the wide ranging implications of how dirty data can impact on your research?

Activity 2

Extracting and scraping data

How often have you found data that looks interesting, but is in a PDF or webpage... how do you get the data into spreadsheets so you can work with them?

The School of Data has fantastic, easy to follow tutorials working with real data.

1. Let's start extracting tabular data from text-based PDFs. The [Extracting Data From PDFs](#) module provides a brief overview of the different techniques used to extract data from PDFs, with a focus on introducing Tabula, a free open-source tool built for this specific task.
2. As much as we wish everything was available in CSV or the format of our choice – most data on the web is published in different forms. How do you extract data from HTML? Use a Scraper!

- a. Go to [Making data on the web useful: scraping](#) and follow the two 'recipes' to learn code-free Scraping in 5-10 minutes using Google Spreadsheets & Google Chrome (Note: Use the Google Chrome Extension "Scraper, by dvhtn")

If you have time or just love data dabbling:

Extracting data from PDFs will inevitably result in some dirty data creeping into your dataset. The School of Data have some really interesting [Data Cleansing](#) modules.

Consider: strategies for encouraging data to be published in more re-usable forms rather than PDF.

Activity 3

An introduction to VO

Getting started - http://www.ivoa.net/astronomers/getting_started.html

Using the VO - http://www.ivoa.net/astronomers/using_the_vo.html

VO compliant tools - <http://www.ivoa.net/astronomers/applications.html>

Especially the section on manuals and how to's

(A more detailed overview will be implemented after our ADACS workshop in November)

Thing 10: What's in a name?

Learn about the key players in Australia's research data management ecosystem.

Activity 1

Who's who in the acronym soup?

Data is at the heart of the Australian national innovation agenda. The key players who help enable the innovation agenda float in an acronym soup! Let's find out who's in the soup....

Read [this article](#) that explains the Australian Government initiatives to foster innovation through publishing and sharing data.

Put (very!) simply the main players in the Australian research landscape (a.k.a. research alphabet soup...) are:

- Universities - our 41 universities generate data, graduate and train new researchers (at ANU, UWA, UQ, UTas, UNSW etc etc)
- CSIRO – for example the CSIRO DAP (Data Access Portal)
- Funders - ARC and NHMRC
- Governments - state and federal departments fund research and produce data eg BoM (Bureau of Meteorology for weather), GA (Geoscience Australia for minerals, maps, more), ABS (Australian Bureau of Statistics) etc etc
- Medical Research Institutes ([AAMRI](#)) and hospitals - conduct research and produce data
- Businesses reuse research and government data and also generate their own data. BCA (Business Council of Aust) brings together the big names in data production and use.
- NCRIS (National Collaborative Research Infrastructure Strategy) is funded by the Australian government to drive research excellence and collaboration between 35,000 researchers, government and industry to deliver practical outcomes.

Let's focus on **NCRIS - it's amazing**. NCRIS is designed to take a national approach to providing the world's best research infrastructure for Australia. NCRIS facilities provide storage for data ([RDS](#)), research computer networking across Australia ([AARNET](#)), tools and virtual labs for researchers ([NeCTAR](#)), very, very big data crunching ([NCI](#)), as well as lots of specialised research facilities. ANDS, who put together 23 Things, on which these 10 Astronomy Things are based, is one of the 27 NCRIS facilities.

1. Browse over some [NCRIS case studies](#) to get an idea of what data and activities are produced by NCRIS facilities.
2. **Just for fun:**
 - a. Fancy some Snakes and Ladders? Find out how some of the Australian infrastructure players fit together: [Print the game and start playing!](#)
 - b. Play the Australian Research Data [Acronym](#) quiz and get your acronym literacy score! Created by the ORCID group for a bit of fun.

Activity 2

Big data: the sky is not the limit!

We often hear terms such as “big data” and “data deluge”. And it doesn’t get much bigger than astronomy and satellite data! Let’s look at 2 big data projects that are only possible because of national collaborations - most of whom love acronyms!

ASKAP - reaching for the stars

The Australian Square Kilometre Array Pathfinder (ASKAP) telescope is made of 36 identical 12-metre wide dish antennas. These produce some 2.5 gigabytes of data per second, equivalent to 75 petabytes per year!

1. Learn about managing this big data project by hearing from the researchers who use ASKAP - watch this stunning 3 min [video](#).

Storing, analysing, managing and publishing such big data usually requires a collaborative effort across a number of organisations. In the case of ASKAP data for example:

- the data is captured at the Murchison Radio-astronomy Observatory (MRO), approximately 315 km northeast of Geraldton in WA.
- It is then transmitted 730 km to the Pawsey Supercomputing Centre in Perth where it is processed and stored.
- Processed data is then published via the CSIRO DAP (Data Access Portal) with metadata records harvested to RDA (Research Data Australia) to enhance discovery.

It’s hard to imagine just how many organisations work together to make this apparently simple workflow possible. Consider for example:

- funding bodies (e.g. NCRIS, federal and state governments)
- research organisations (e.g. CSIRO, Swinburne University, University of WA)
- infrastructure providers (AARNET, Pawsey Supercomputing Centre, ANDS)

Learn more about the [ASKAP telescopes](#).

Consider: why big data is often publicly available, yet the so called ‘long tail of research data’ (smaller data sets) are often not published.

What's next?

Reflect on the changes you could, and perhaps should, make in research data management practices which will enable the ethical and efficient publication of health, medical and clinical data for reuse by the research community.

Consider

1. Learning more about research data management by browsing the [23 \(research data\) Things program](#) which includes data management issues not covered here.
2. Making connections to other people who 'know data' in your institution e.g. Librarians, Repository managers, IT, Researchers
3. Reviewing your technical skills through either:
 - Online courses such as those mentioned in [Thing 21: Tools of the \(dirty\) data trade](#)
 - [ADACS](#) training, [ResBaz](#) or one of the [Carpentry courses](#)