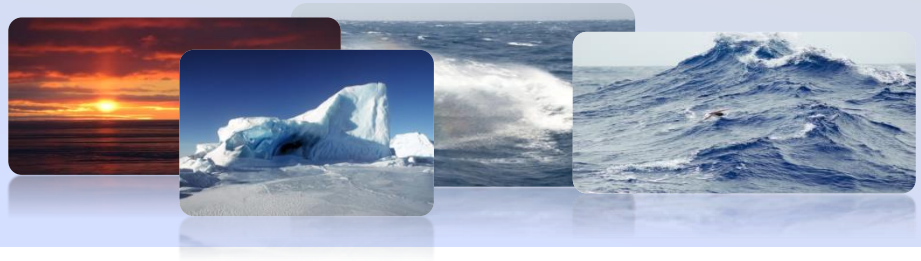
Australian Government
Department of the Environment and Energy
Australian Antarctic Division

The Three Legged Stool of Antarctic Data Management

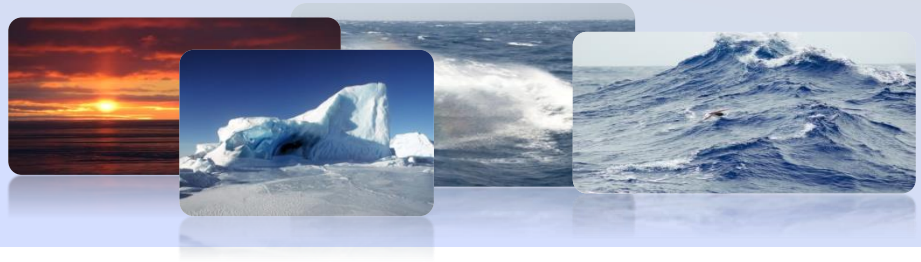
Dave Connell – Australian Antarctic Data Centre



The Australian Antarctic Division

- Established 1947
- Coordinates Australian scientific involvement in the Antarctic, sub-Antarctic and Southern Ocean
- Has responsibility for management of the Australian Antarctic program (AAp)

Image - https://goo.gl/images/d04dLU



The Australian Antarctic Division

- Maintains three Antarctic stations, a sub-Antarctic station, an intercontinental air system, and a shipping system
- Conducts broad-themed, multi-disciplinary science
  - Approximately 60 science projects running each season

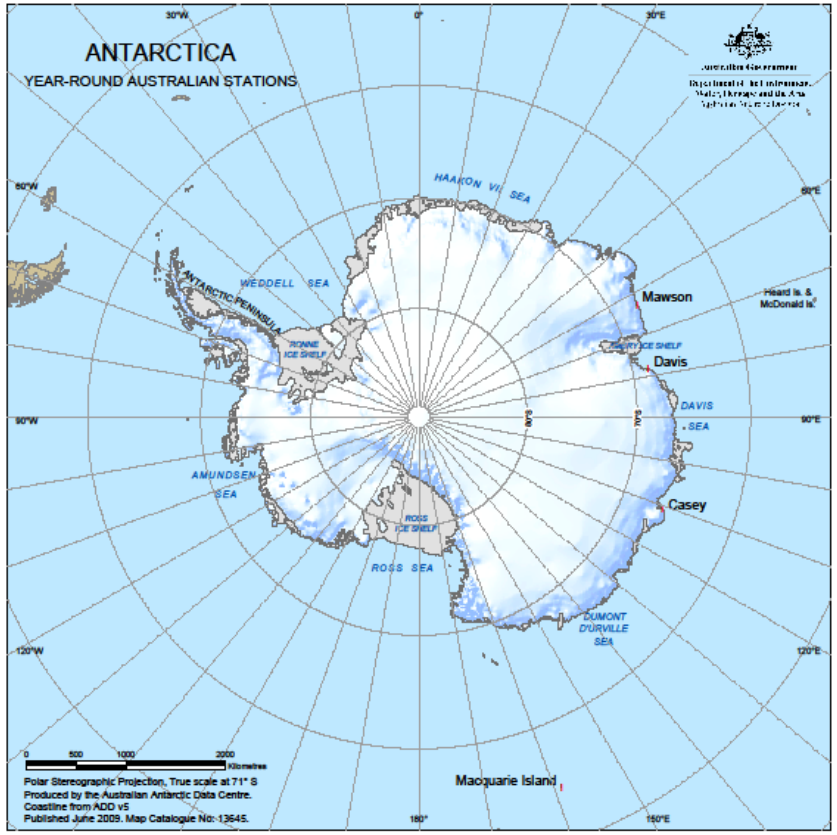
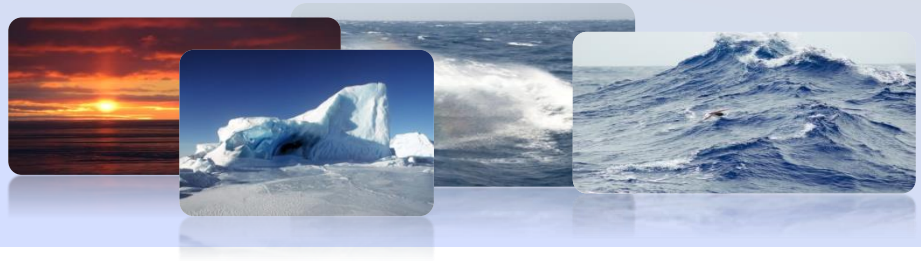


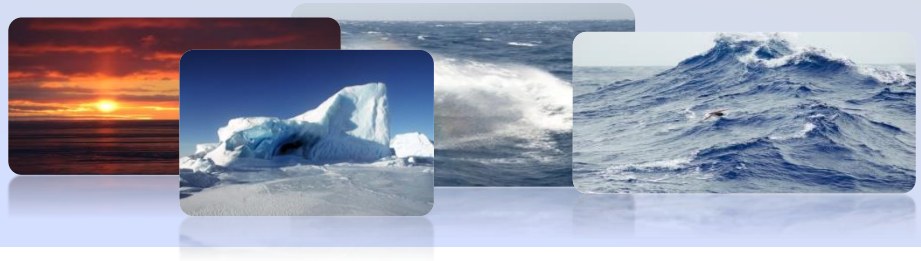
Image – Australian Antarctic Division



The Australian Antarctic Data Centre

- Established 1995
- Has responsibility for the *data* management of the Australian Antarctic program (AAp)
- Fulfils Australia's obligations under Article (III).(1).(c) of the Antarctic Treaty – “Scientific observations and results from Antarctica shall be exchanged and made freely available.”

Image - https://goo.gl/images/ex1606

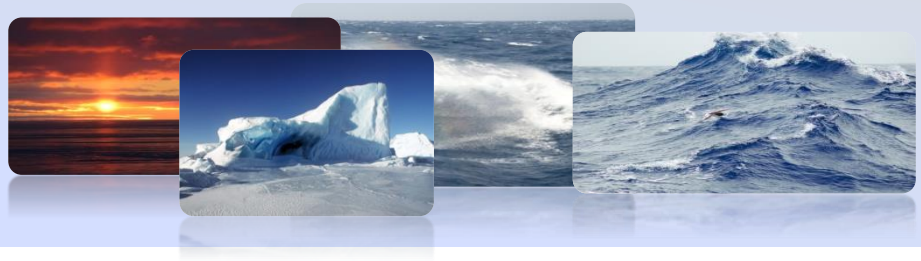


The AAp Project Cycle

- Project application
- Project approval
- Data Management Plan
- Conduct scientific research
- Catalogue, archive and publish data (DOIs)
- Write papers

- Underpinned by the AAp Data Policy - http://data.aad.gov.au/aadc/about/data_policy.cfm

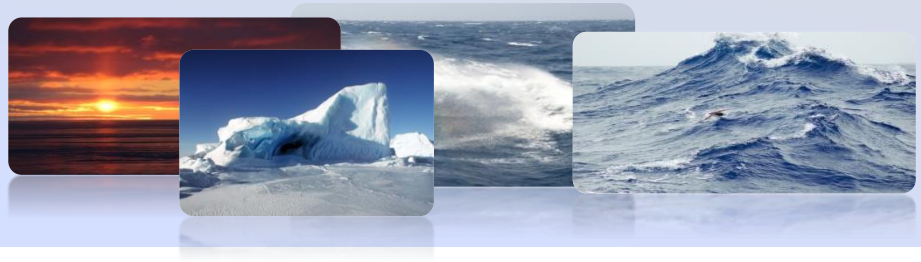
Image – Australian Antarctic Division



The AAp Project Cycle

- Review process
- Scientists are scored on their data management practices
  - Can have an impact on approval of future projects

Image - https://goo.gl/images/CoI1CG



The Three Legged Stool

- How can the Australian Antarctic Data Centre make the data management process easier?

- MyScience – project management
- Metadata – create metadata records
- Data submission – reliably upload data

https://goo.gl/images/s7YoTu

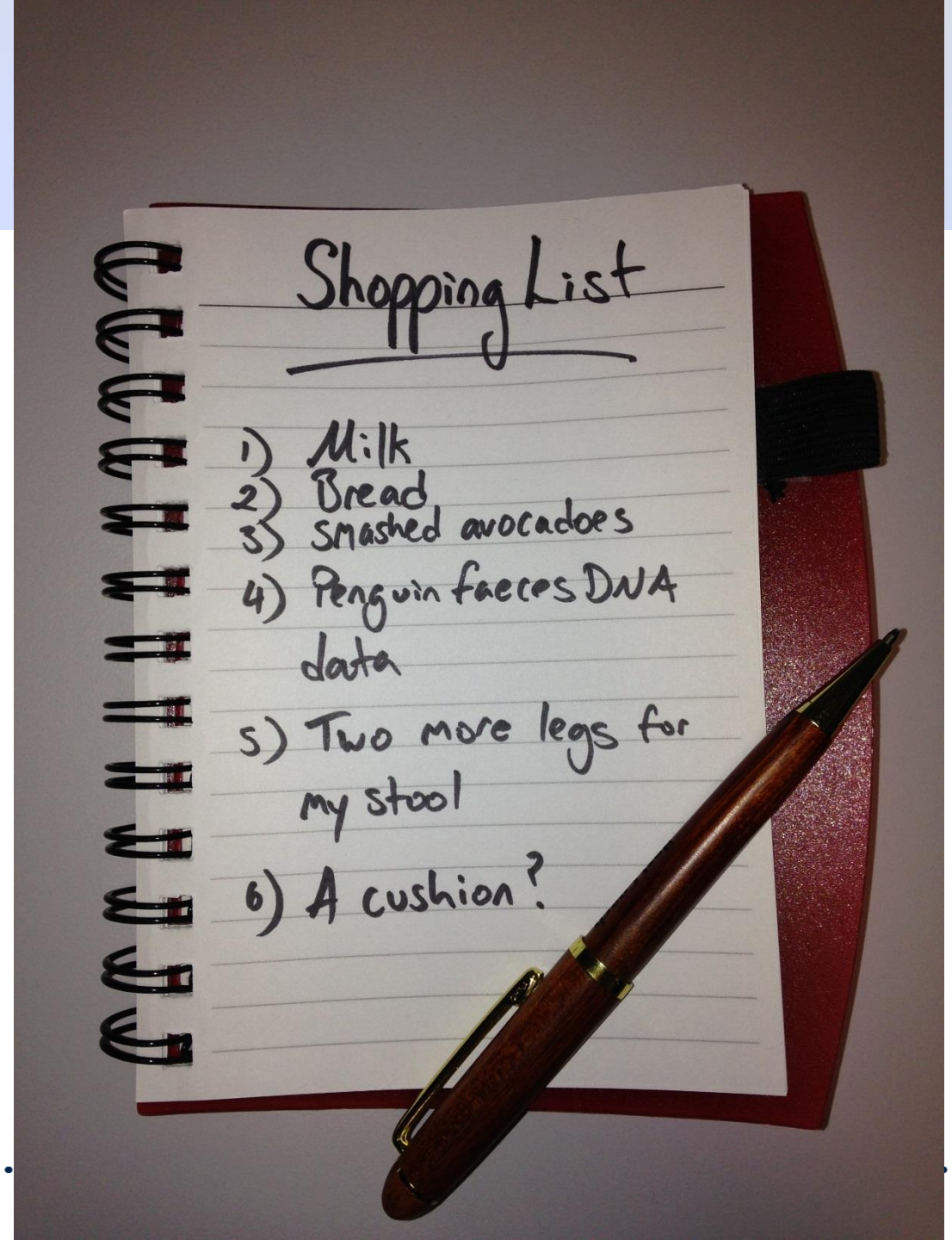
The Three Legged Stool – Leg One

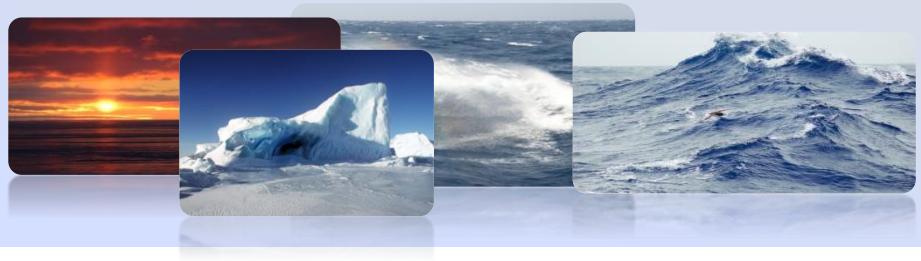
- MyScience – project management tool
- Introduced to the AAp in 2012
  - Designed for scientists and AADC staff to keep track of the data management progress of each project
  - Linked to the (old) metadata catalogue
  - Could instantly see what data have been archived with each project
  - Could instantly see the data status of each project (e.g. *in progress*, *complete*)
  - Utilised Data Management Plans (DMPs)

https://goo.gl/images/mAfRJZ

The Three Legged Stool – Leg One

- DMPs the main feature of MyScience
  - We knew what data to expect
  - We knew when to expect the data
  - We knew who to expect the data from
  - We knew how much storage space we would need (estimate)
  - We knew when to stop asking for the data

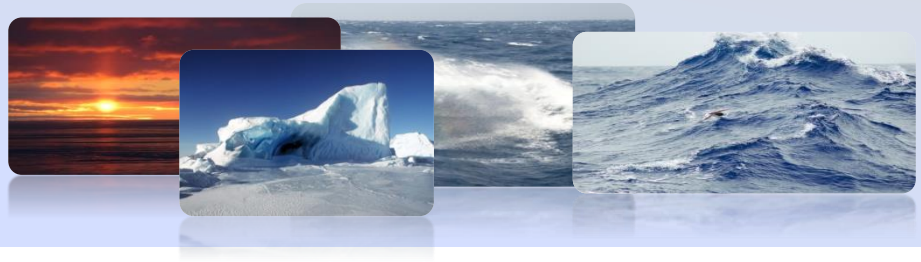




The Three Legged Stool – Leg Two

- Metadata

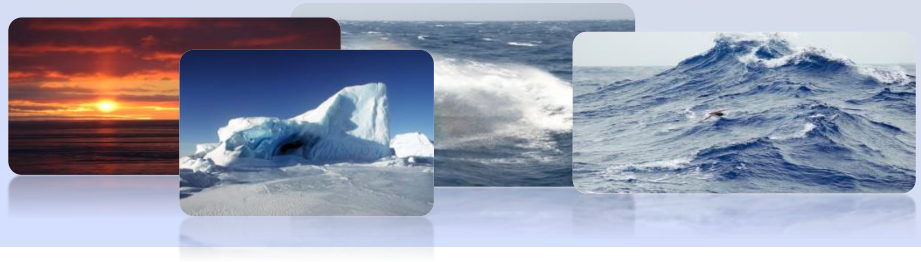
https://goo.gl/images/bFXxZ6



https://goo.gl/images/F5y5gK

The Three Legged Stool – Leg Two

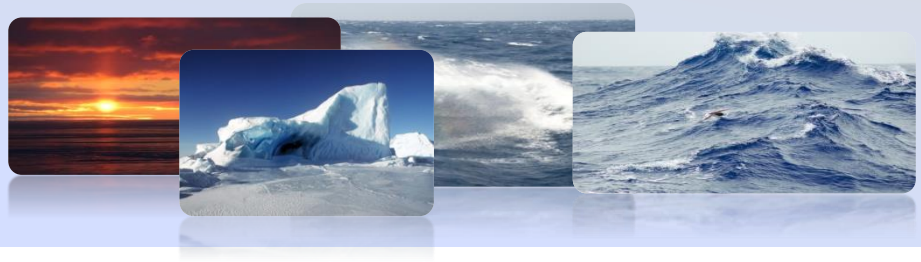
- 1995-1999 – ANZLIC
- 1999-present – DIF, ISO 19115, RIF-CS
  - DIF (Directory Interchange Format) was developed by the GCMD (Global Change Master Directory) of NASA.
  - Standard used by the international Antarctic community
  - Metadata also made available in ISO 19115 and RIF-CS formats (DIF metadata are converted automatically)
  - All metadata placed in WEBDAV folders for harvesting



The Three Legged Stool – Leg Two

- Metadata Creation
- The problem was that the AADC did not have a reliable, easy-to-use authoring tool
- The DIF tool was:
  - Complex
  - Used unfamiliar metadata jargon (to a scientist)
  - Stand alone tools – not integrated with AADC systems (labour intensive)
  - Required significant assistance from metadata officer
- Mostly produced poor quality metadata that required AADC time to fix

https://goo.gl/images/y5VpYk

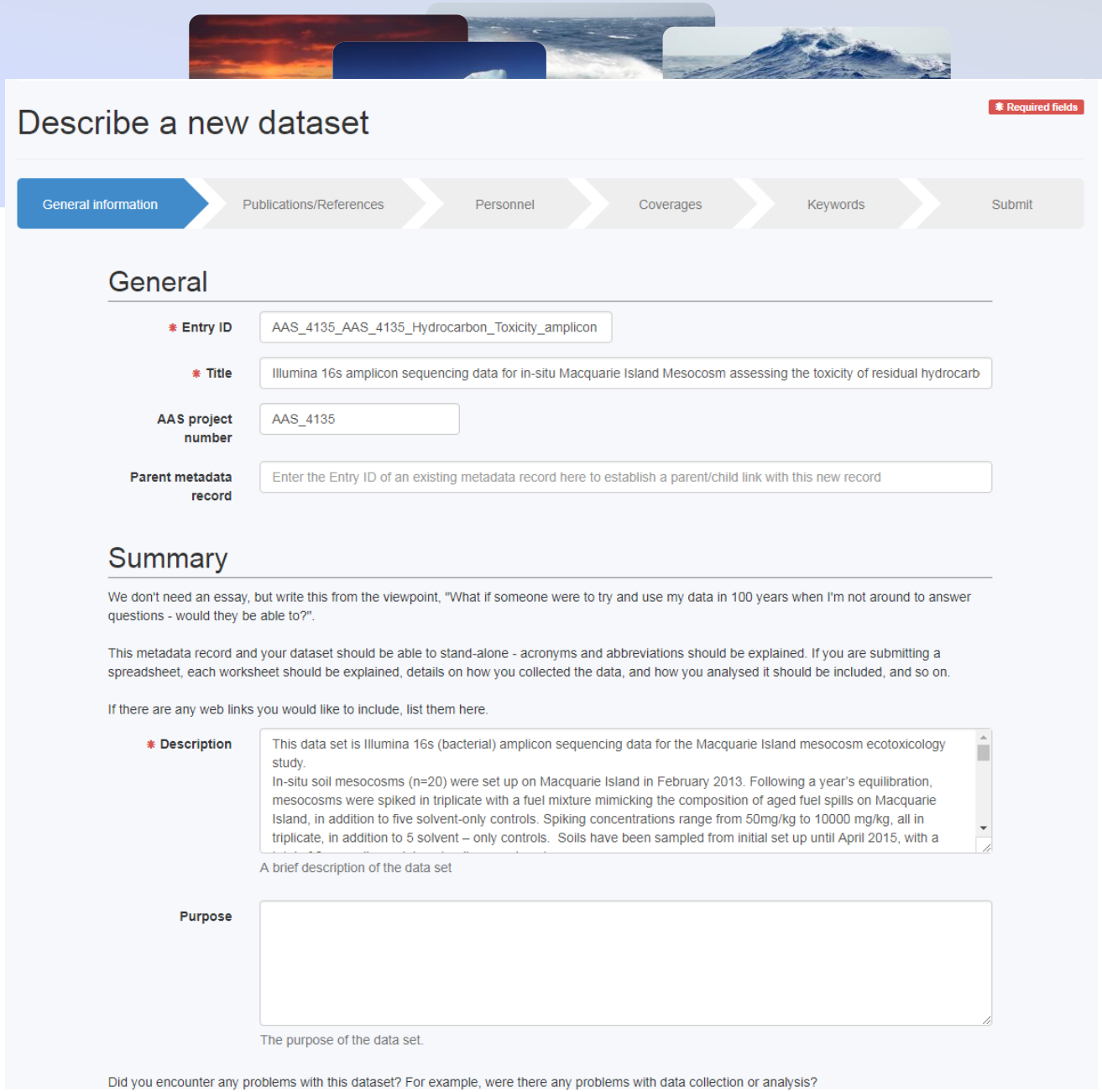


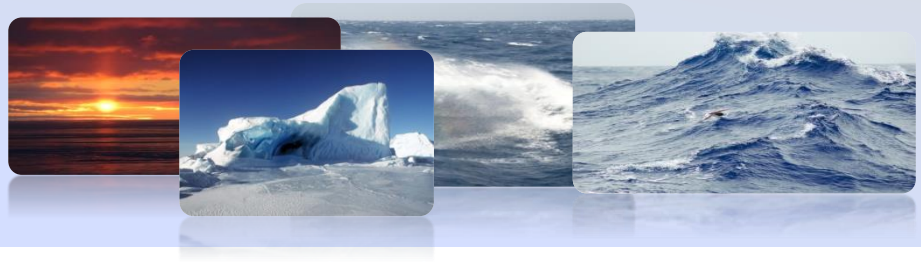
The Three Legged Stool – Leg Two

- 2012 - Development of a Word template
  - Only 8 questions
  - Familiar interface, simple
  - Could be used offline (e.g. on Antarctic voyage)
  - Enormously successful – high quality metadata produced
- Lot of work for the AADC to then convert into a DIF metadata record
  - Time consuming
  - Manual process
  - DIF tool used as an admin interface

The Three Legged Stool – Leg Two

- 2015 – release of new metadata tool
  - DIF based
  - Streamlined – many fields removed or automatically set
  - Wizard interface
  - Produces “complete” DIF XML that requires minimal modification by AADC
  - Integrated with AADC applications
  - Proven to be very successful

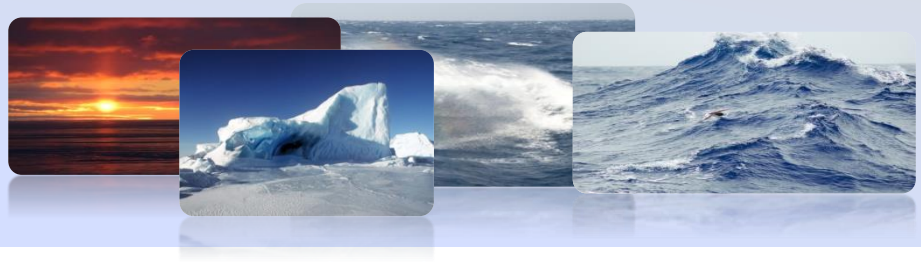




The Three Legged Stool – Leg Three

- Data submission
  - Version one released in 2008
  - It was serviceable, but poorly coded
  - Grew evermore buggy before becoming irretrievably broken in 2015
  - Work on a replacement tool began in 2013, but was continually delayed due to resourcing problems – work finally began in earnest in 2016

https://goo.gl/images/YoLXbg



The Three Legged Stool – Leg Three

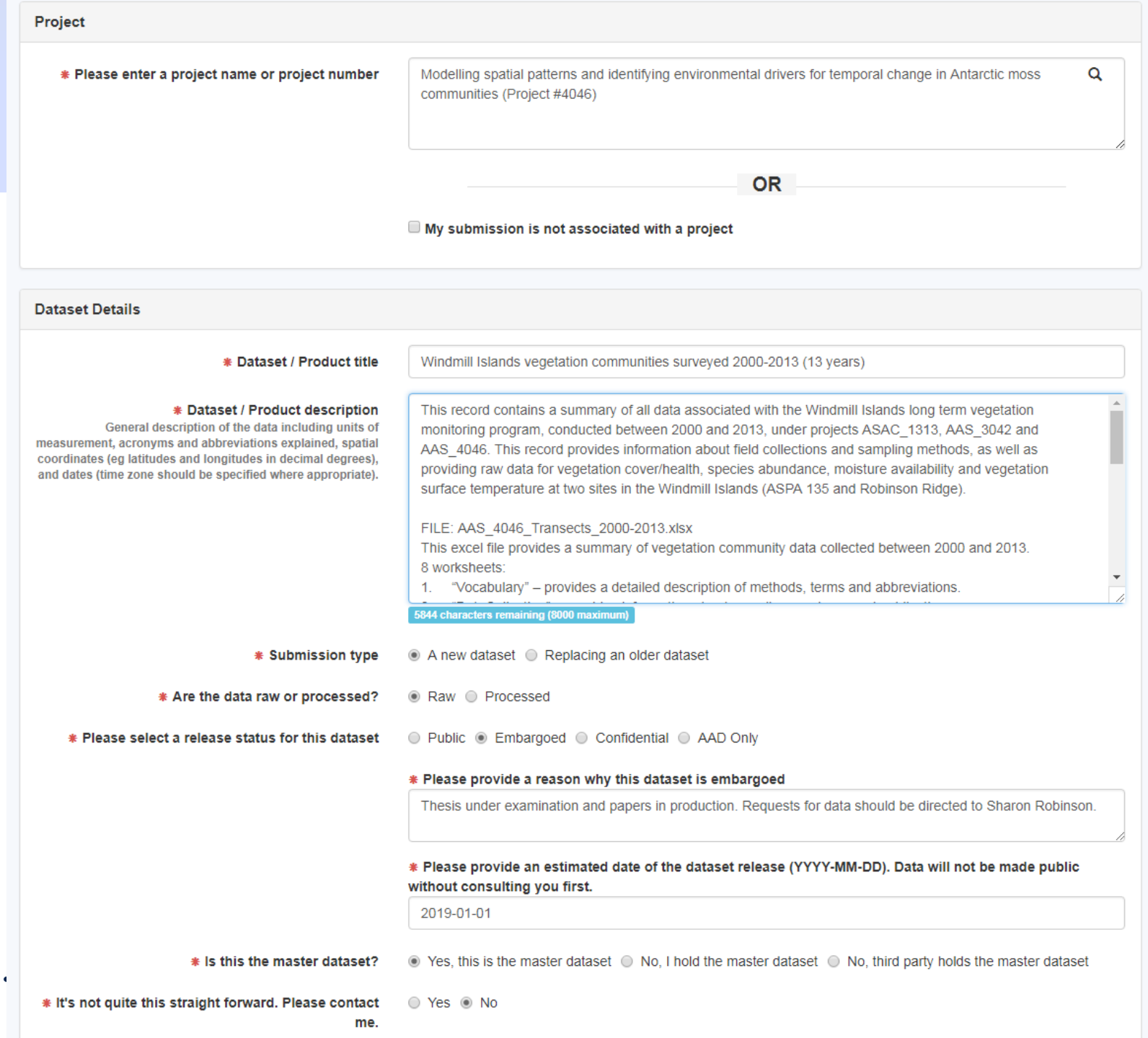
- Data submission
  - Version two released in 2017
  - Stable
  - Functional
  - Far more useful than the original (can upload larger datasets)
  - Linked to DMPs for easier administration
  - Linked to metadata tool

https://goo.gl/images/F2P8m3

Australian Government
Department of the Environment
Australian Antarctic Division

The Three Legged Stool – Leg Three

- Data submission
  - Simple to use
  - Automatically puts data in correct area of server
  - Allows for versioning of datasets
  - Sends automated email when embargoed data are due for release



Project

* **Please enter a project name or project number**: Modelling spatial patterns and identifying environmental drivers for temporal change in Antarctic moss communities (Project #4046)

OR

☐ **My submission is not associated with a project**

Dataset Details

* **Dataset / Product title**: Windmill Islands vegetation communities surveyed 2000-2013 (13 years)

* **Dataset / Product description**
General description of the data including units of measurement, acronyms and abbreviations explained, spatial coordinates (eg latitudes and longitudes in decimal degrees), and dates (time zone should be specified where appropriate).

This record contains a summary of all data associated with the Windmill Islands long term vegetation monitoring program, conducted between 2000 and 2013, under projects ASAC_1313, AAS_3042 and AAS_4046. This record provides information about field collections and sampling methods, as well as providing raw data for vegetation cover/health, species abundance, moisture availability and vegetation surface temperature at two sites in the Windmill Islands (ASPA 135 and Robinson Ridge).

FILE: AAS_4046_Transects_2000-2013.xlsx
This excel file provides a summary of vegetation community data collected between 2000 and 2013.
8 worksheets:
1. "Vocabulary" – provides a detailed description of methods, terms and abbreviations.

5844 characters remaining (8000 maximum)

* **Submission type**: ◉ A new dataset ○ Replacing an older dataset

* **Are the data raw or processed?**: ◉ Raw ○ Processed

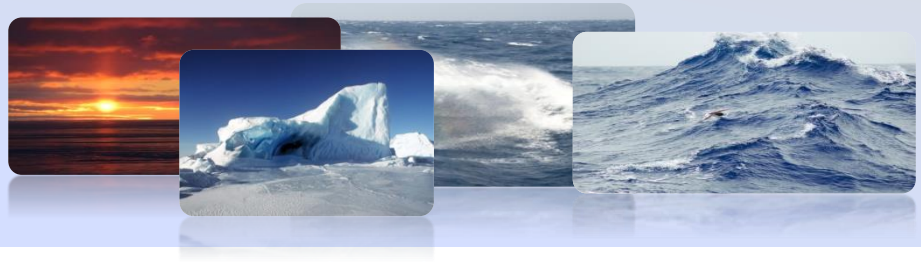
* **Please select a release status for this dataset**: ○ Public ◉ Embargoed ○ Confidential ○ AAD Only

* **Please provide a reason why this dataset is embargoed**: Thesis under examination and papers in production. Requests for data should be directed to Sharon Robinson.

* **Please provide an estimated date of the dataset release (YYYY-MM-DD). Data will not be made public without consulting you first.**: 2019-01-01

* **Is this the master dataset?**: ◉ Yes, this is the master dataset ○ No, I hold the master dataset ○ No, third party holds the master dataset

* **It's not quite this straight forward. Please contact me.**: ○ Yes ◉ No

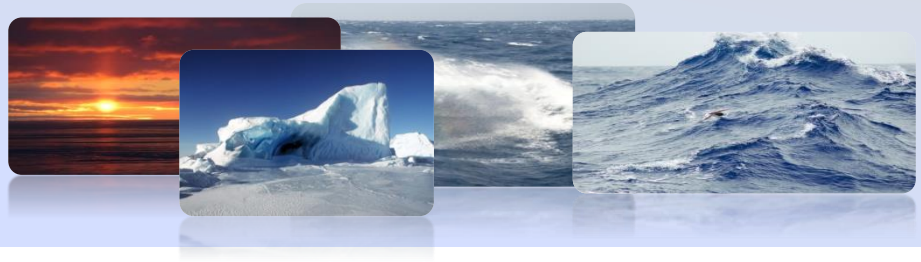


The Three Legged Stool

- MyScience
- Metadata
- Data submission

- All linked together for ease of use by administrators as well as general users

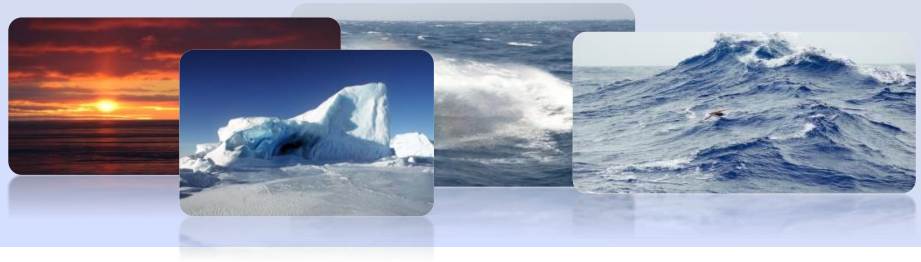
https://goo.gl/images/mA2rTq



The Three Legged Stool – With a cushion

- DOIs and increased citation
- Searching and data access
- Lots of exposure due to integration with other catalogues
- Value adding with applications
- Nature approved repository
- WDS certification (hopefully soon)
  - Data Seal of Approval

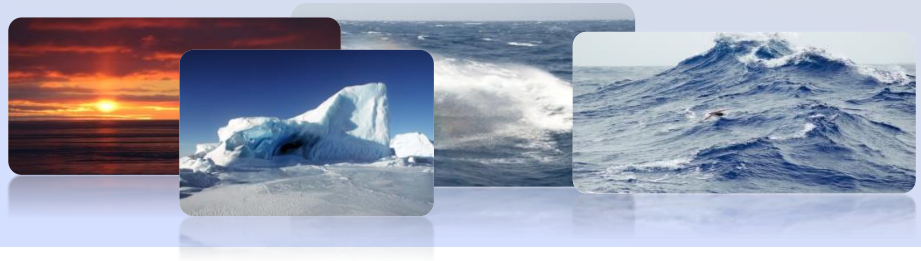
https://goo.gl/images/Jo2McU



The Three Legged Stool – With a whip

- Reporting back to funding office
- Data scores for scientists
- What happens to those that score badly?

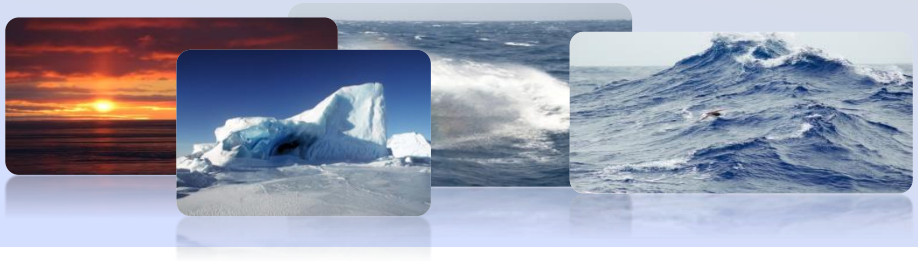
https://goo.gl/images/jRfmoZ



Conclusions - The Three Legged Stool – is it wobbly?

- Better linkages – automatically update DMPs (improvement)
- Some policy loopholes need to be tightened (non-science projects)
- Transferring very large datasets (> 100GB) (Cloudstor site says that 2TB is the limit – new? Not tested yet)
- Some projects still don't archive data (can we expect 100%?)
- App development can bit a bit stop-start (due to a lot of competing pressures)

https://goo.gl/images/qrbZvf



Conclusions - The Three Legged Stool – is it wobbly?

- Standards always changing/updating
  - 19115-1 (metadata)
  - 19165 (new draft standard for data centres)
- Responsibilities to international and national needs
- We don't really integrate our data, etc. like the AODN does
  - Would like to
  - Will hopefully soon be pushing data out to AODN for inclusion in their services
- Need better reporting mechanisms

https://goo.gl/images/qrbZvf

Questions?

https://goo.gl/images/nr5SR5