

**NCI**  
AUSTRALIA

## What resources are available Internationally and what are the new drivers?

Lesley Wyborn<sup>1, 2, 3</sup>

<sup>1</sup>*National Computational Infrastructure, Australian National University,*

<sup>2</sup>*Australian National Data Service;*

<sup>3</sup>*AuScope,*

ED 32B-03

Contact: [lesley.wyborn@anu.edu.au](mailto:lesley.wyborn@anu.edu.au)

- Chair of the Academy in Science, Data in Science Committee
- On the AGU Data Management Advisory Board
- On the Steering Committee of the ICSU/CODATA Commission on Scientific Standards for Integration of Research Data for Transdisciplinary Science
- On the Steering Committee for the AGU/Research Data Alliance Project for FAIR and Open Data in Research Publications
- On the AuScope Strategic Reference Panel
- AGU Meetings Committee

- FAIR Publishing in the Earth and Space Sciences
- The Linked Open Publication
- IGSN
- AuScope Virtual Environment

- Funded by the Laura and John Arnold Foundation (LJAF)
- LJAF makes strategic investments in criminal justice, education, evidence-based policy, public accountability, and **research integrity**
- Aims to develop a collaborative solution across researchers, journals and repositories that will evolve the Earth and Space Science (ESS) publication process to include all research inputs into a publication to help develop a unified process that supports researchers from grant application through to publishing.



## The Problem...



Optional

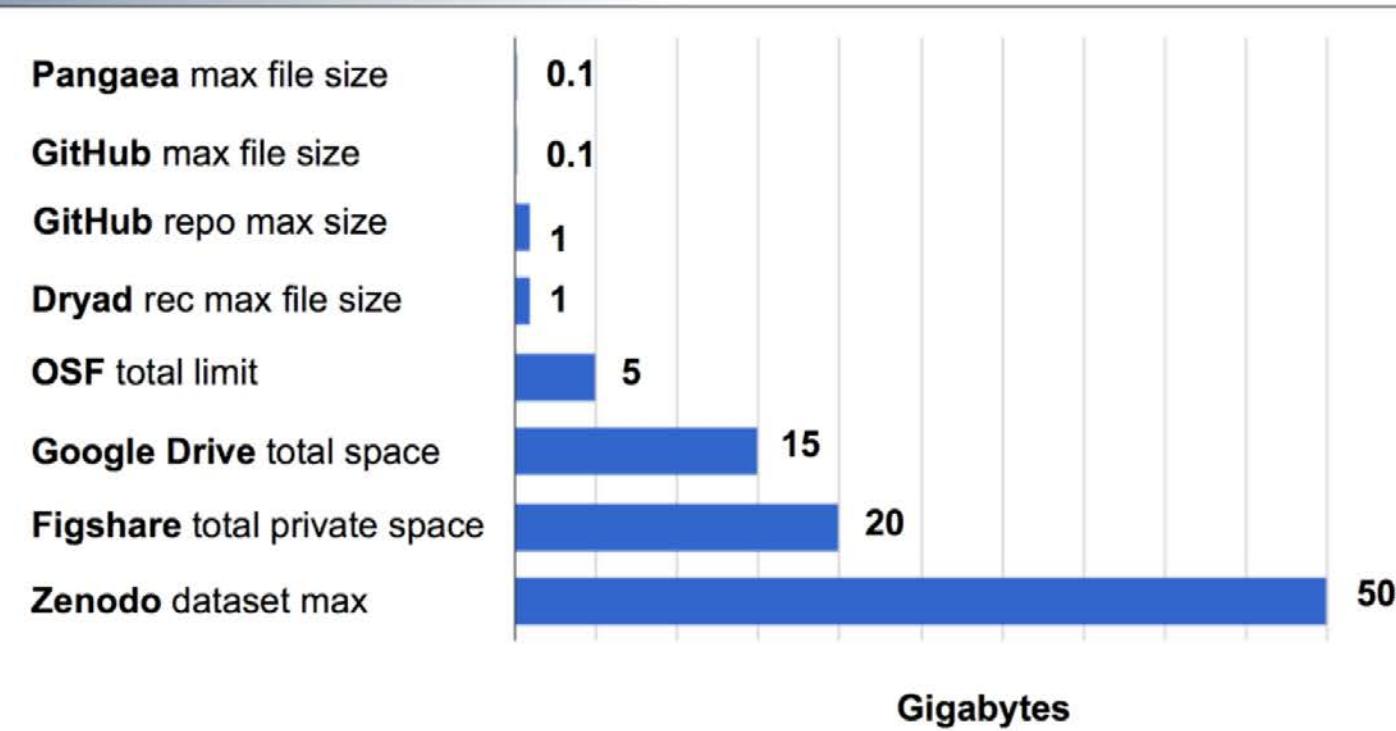
At the time of publication:

- Are the data (samples, software/services) that support the paper properly documented and stored in a persistent, sustainable repository?
- Are the data (samples, software/services) citable with a persistent identifier (e.g., DOI), and support the Nature FAIR Guiding Principles for scientific Data management and Stewardship (<https://www.nature.com/articles/sdata201618>)
- Do researchers have a similar experience with submitting their paper and supporting data (and software/services) no matter the journal?

Optional

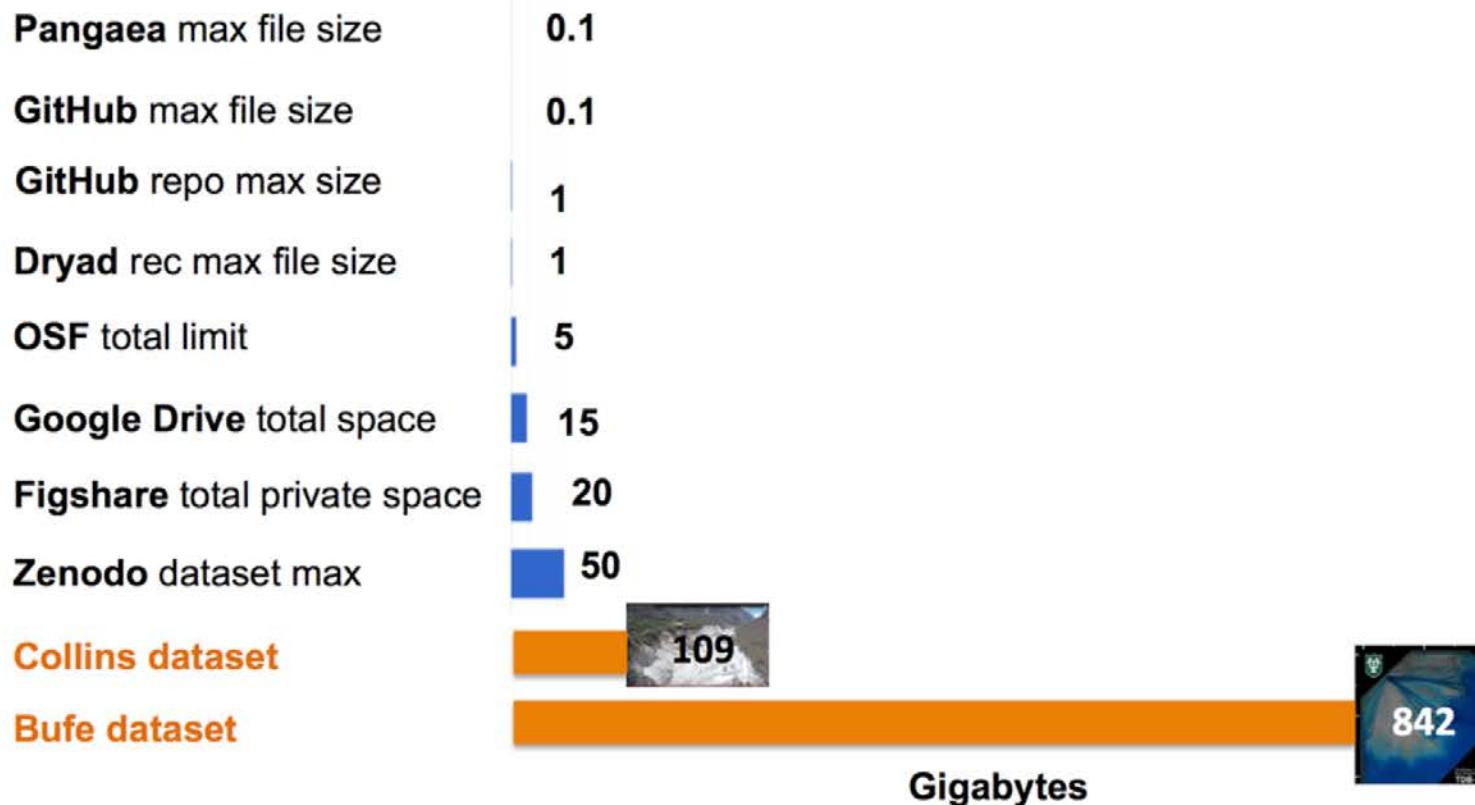
No

## Storage capability of some common repositories (as of Sept 25, 2016)



Slide Courtesy of Dr Leslie Hsu, USGS

## Storage capability of some common repositories (as of Sept 25, 2016)



Slide Courtesy of Dr Leslie Hsu, USGS

## The Solution (1): Proper data documentation and storage



- In support of a publication...there is proper **data documentation** and it is stored in and accessible from a trustworthy repository.
- Need to:
  - require data be Accessible from a repository with the paper as the default option (No data as Supplements to papers)
  - Need to engage domain repositories to ensure proper curation as much as possible



- In support of a publication...proper **data citation with a persistent identifier** supporting the FAIR Guidelines.
  - Need to require use of repositories that use persistent identifiers.
  - Need to require use of repositories that have landing pages that support citation.
  - Need to require use of repositories that can support community agreed standards

- In support of a publication...provide a **similar experience for a researcher** when submitting their paper and supporting data (and software/services) no matter the journal.
  - Journals and repositories need to define and adopt recommendations and align policies.



- Earth Science Information Partners
- Research Data-Alliance
- *Science*
- *Nature* Research
- *PNAS* (NAS)
- Center for Open Science (COS)
- COPDESS
- AuScope
- National Computational Infrastructure, ANU
- Australian National Data Service (ANDS)

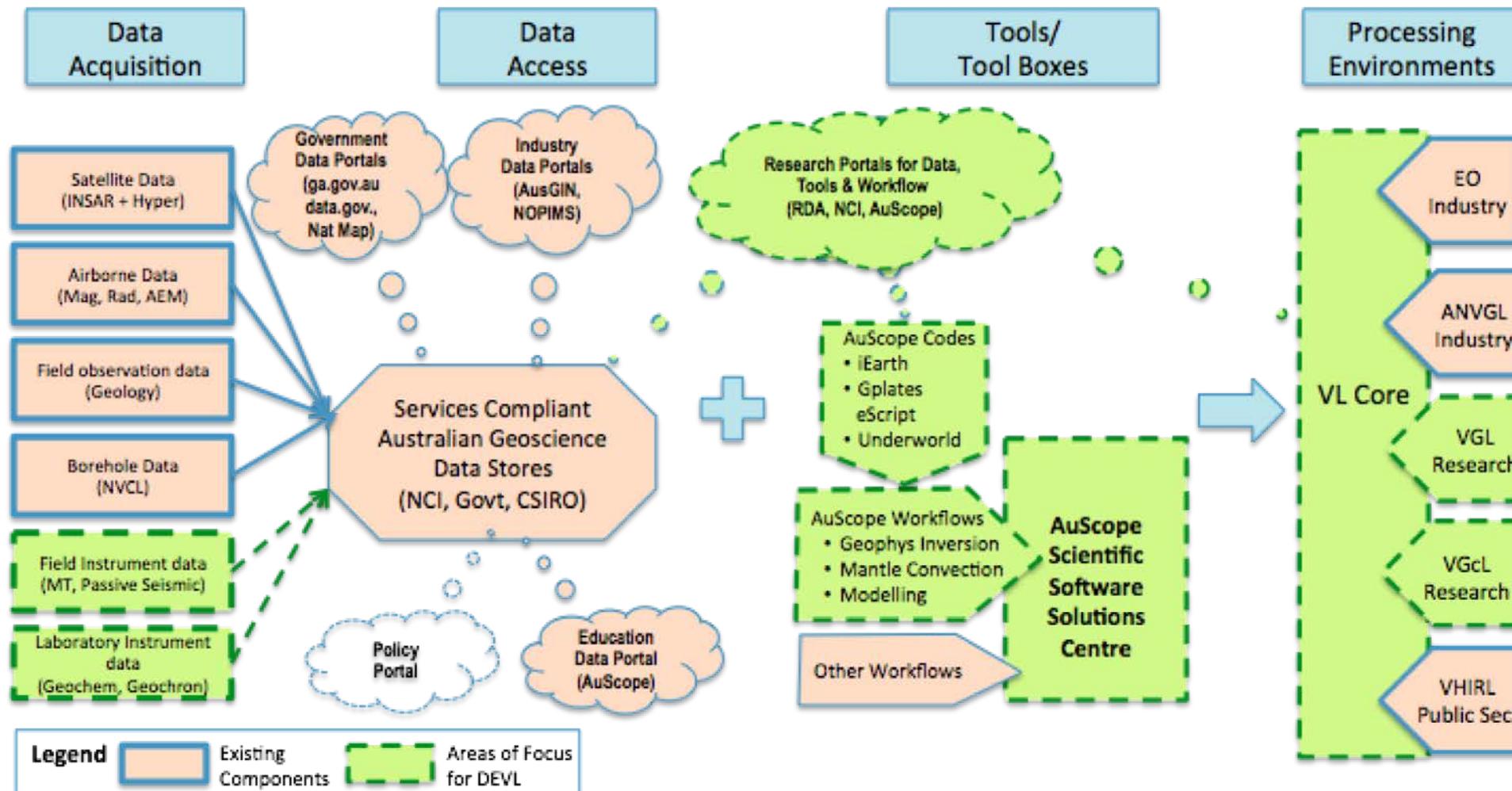
Focus is on the Earth and Space science journals, repositories, and researchers.



- Across Earth and Environmental areas they are patchy
- More common in Geophysics
- For Machine Learning/Deep Learning we need more consistent and agreed vocabularies, semantics and ontologies?
- Enter the ICSU/CODATA Commission on Scientific Standards for Integration of Research Data for Transdisciplinary Science

- Growing attempts on standards for geochemistry
- Certified domain repositories are coming on line (e.g., NSF IEDA)
- Proposed online registry of reference materials for isotopes
  - Proposed meeting at Goldschmidt
- US and Germany are combining to share developments and develop standards
- Will be done through International Geochemistry Society
  - Proposed splinter meeting at EGU Vienna

## AuScope (data enhanced) Virtual Research Environment (AVRE)



- National **Collaborative** Research Infrastructure Strategy
- AuScope Geosciences DeVL (Data-enhanced Virtual Laboratory)
  - IGSN identifiers for the Australian ~~Geoscience~~ NCRIS Research community
  - A FAIR MagntoTelluric (MT) repository for research data at NCI
  - Enable a software registry that indexes code in GitHub repositories
  - Revamp the AuScope Virtual Laboratory infrastructure to enable it to make connection to more geochemistry data sets and more software – i.e. support the Virtual GeoChemistry Laboratory (VGCL)

# What is a Virtual Research Environment (VRE)?



A Virtual Research Environment, Science Gateway or Virtual Laboratory is an on-line system supporting collaborative research that enables harnessing of the power of the Internet to support a more dynamic, online approach to collaborative working

## Key features:

- Provide access to data resources that are accessible online
- Enable online use of discipline-specific tools, such as data analysis, visualisation, or simulation management
- Online access to compute resources
- Collaboration support (Web forums and Wikis)
- May include publication management and teaching tools

VRE's are important in fields where research is primarily carried out in teams spanning multiple institutions and even countries

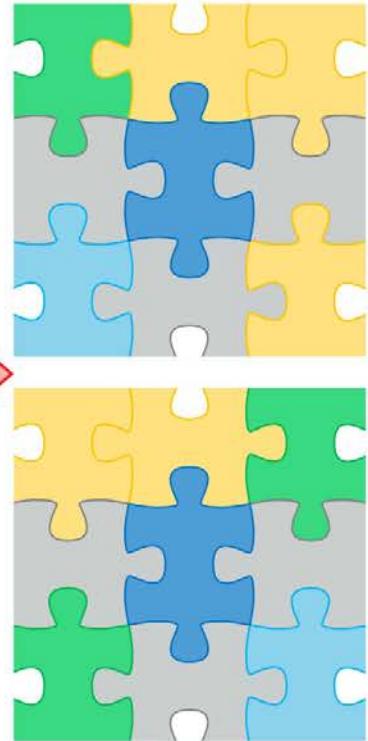
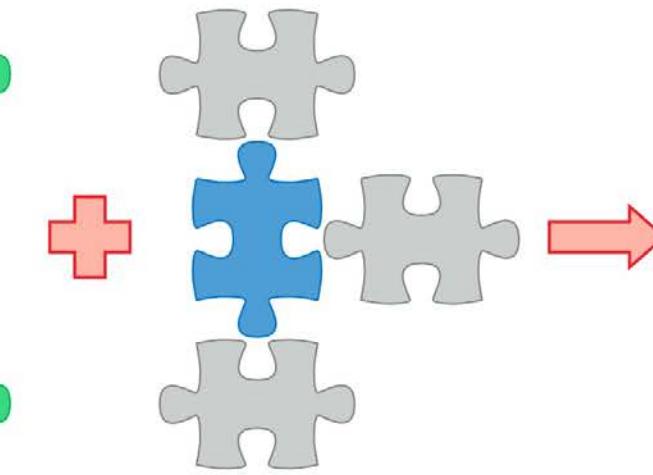
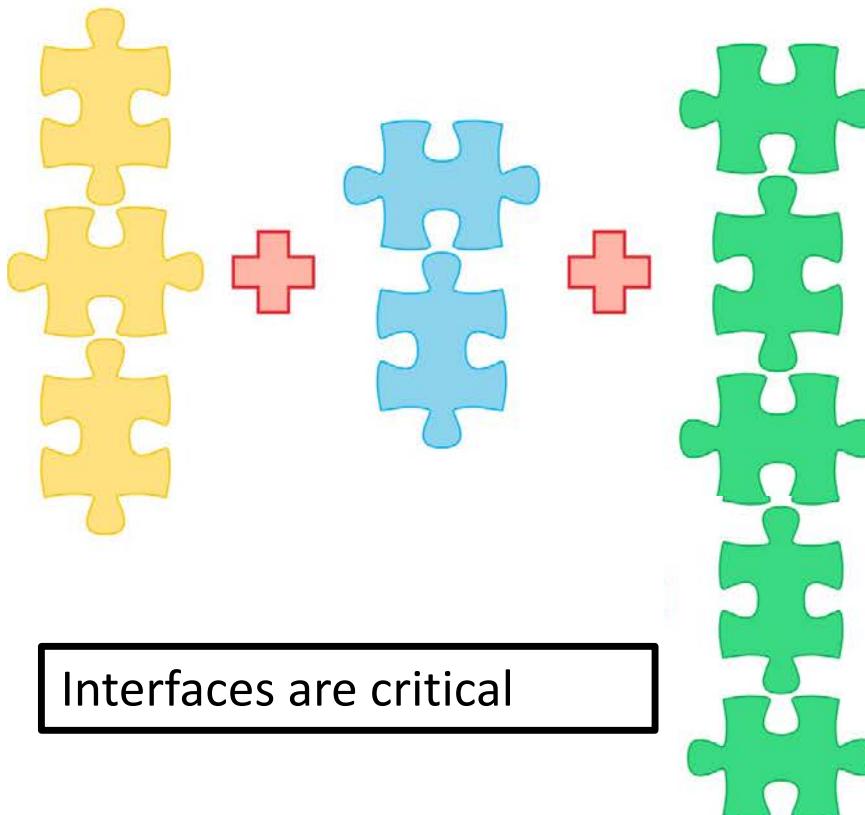




# Elements of A Virtual Research Environment

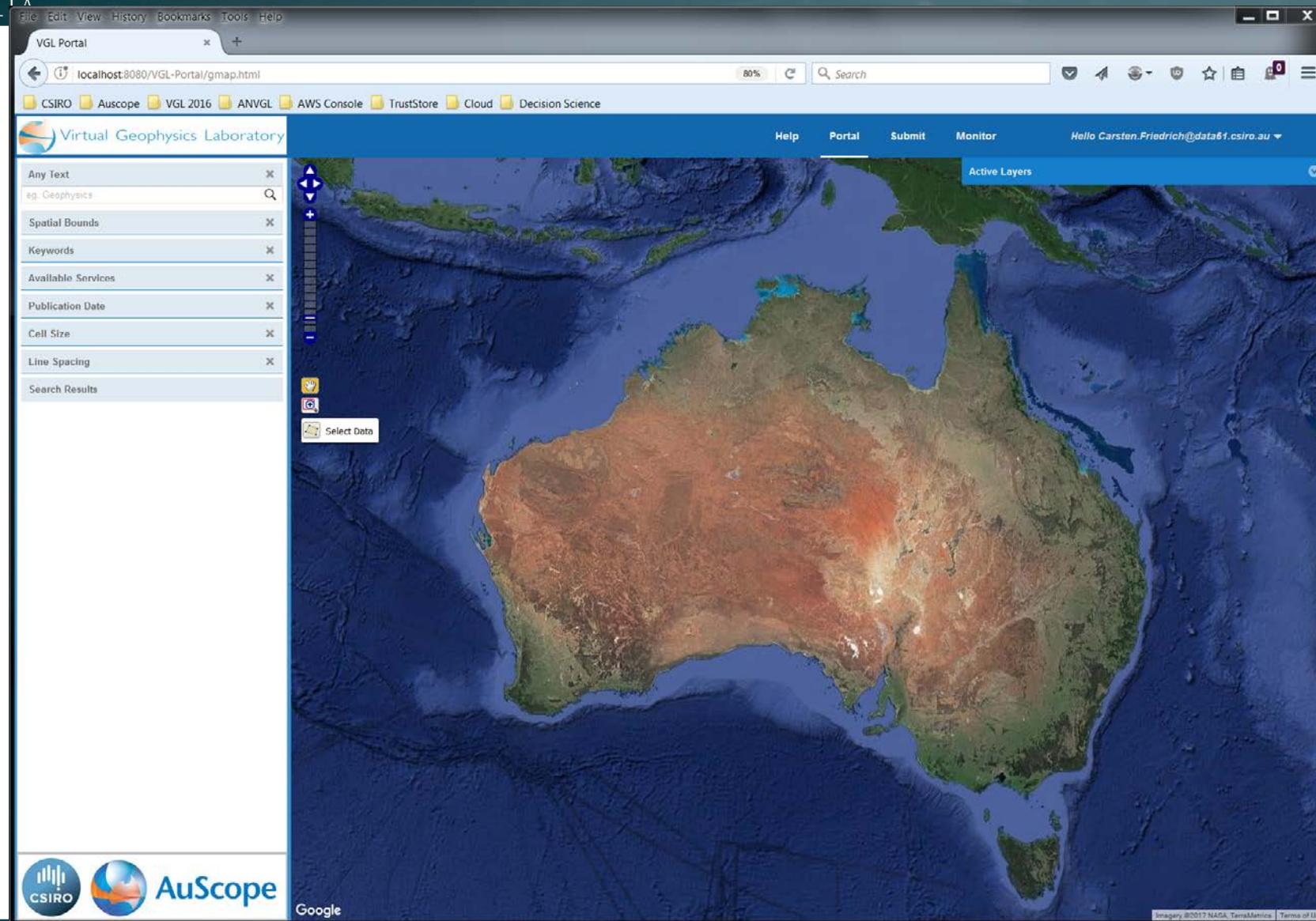
- Text goes here

**Data Services** + **Processing Services** + **Compute Services** + **Enablers (eg. OGC "Glue")** → **Virtual Laboratory**



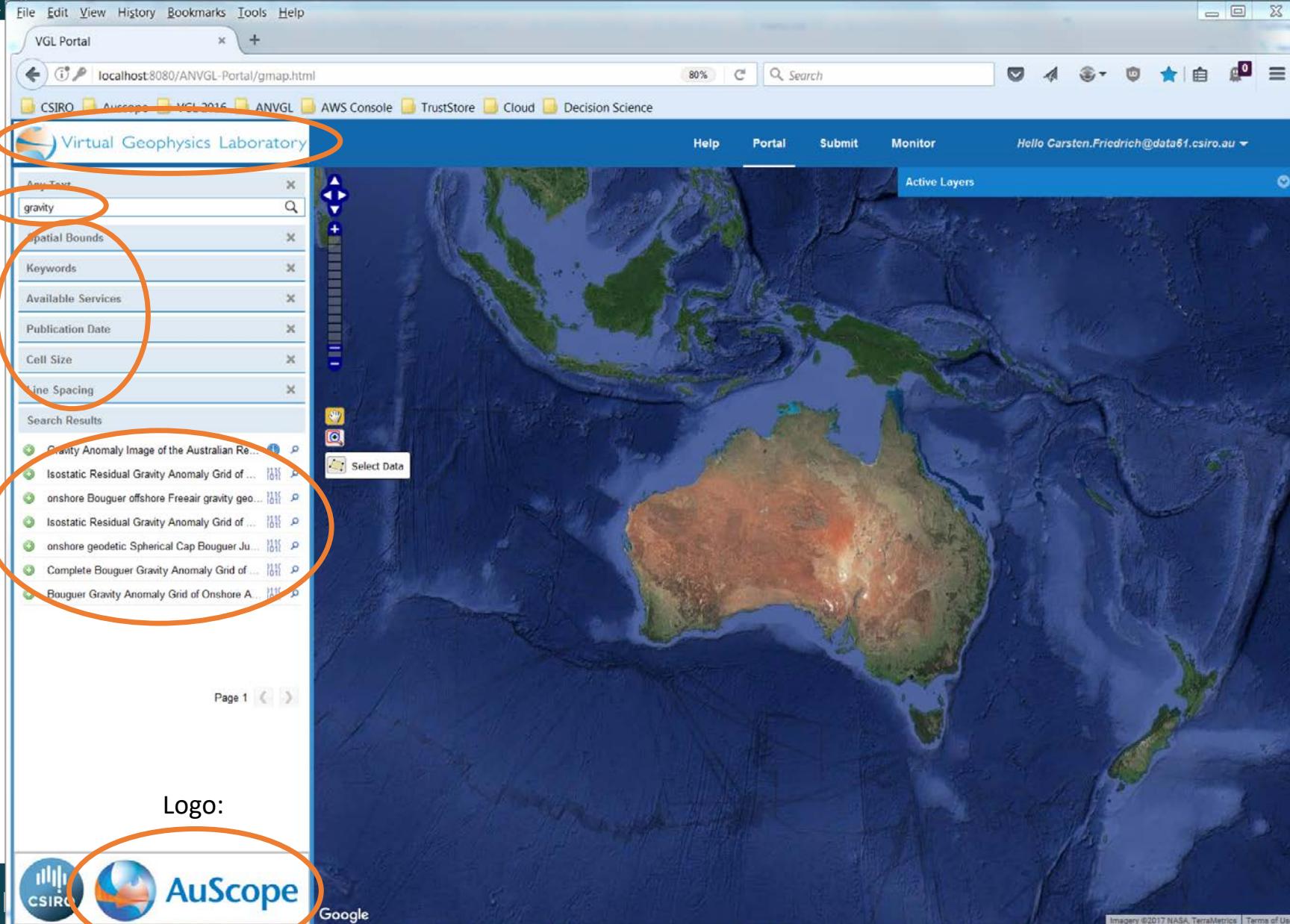
Sharing core components across multiple VL's enhances sustainability and is more cost efficient

# Introducing the Virtual Geophysics Laboratory



# Data Select Screen

Logo:

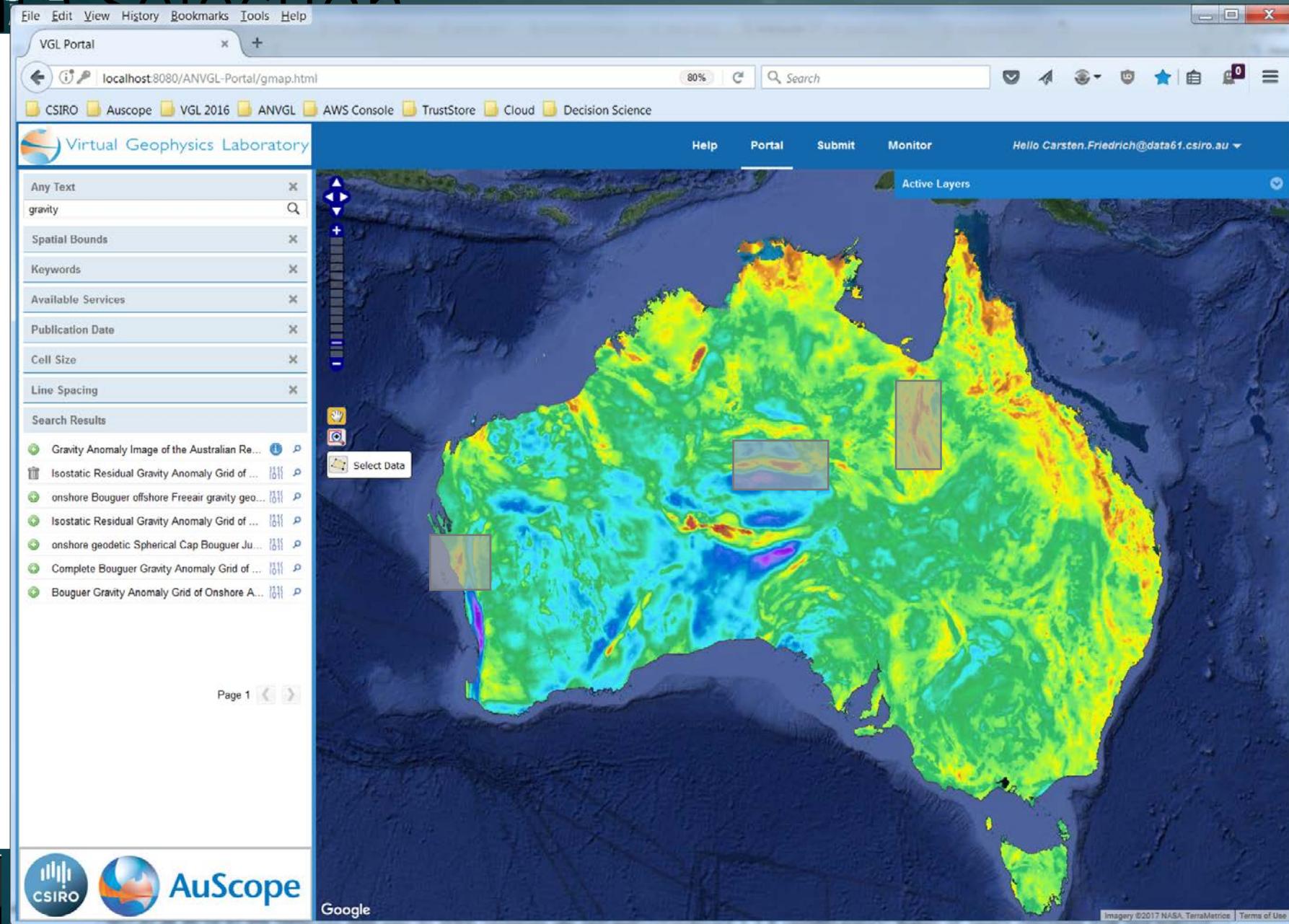


The screenshot shows a web browser window for the "VGL Portal" at [localhost:8080/ANVGL-Portal/gmap.html](http://localhost:8080/ANVGL-Portal/gmap.html). The interface includes a top navigation bar with links for File, Edit, View, History, Bookmarks, Tools, Help, and a search bar. A sidebar on the left contains a tree view of categories like CSIRO, AuScope, VGL 2016, ANVGL, AWS Console, TrustStore, Cloud, and Decision Science. Below the sidebar is a search panel with a search term "gravity" and facets for Spatial Bounds, Keywords, Available Services, Publication Date, Cell Size, and Line Spacing. The main area features a world map with a focus on Australia, showing geological and topographical data. A list of search results is displayed on the left, including items like "Gravity Anomaly Image of the Australian Re..." and "Isostatic Residual Gravity Anomaly Grid of ...". A "Select Data" button is visible near the bottom of the search results. At the bottom of the page, there are logos for CSIRO and AuScope, along with Creative Commons BY and Google branding.

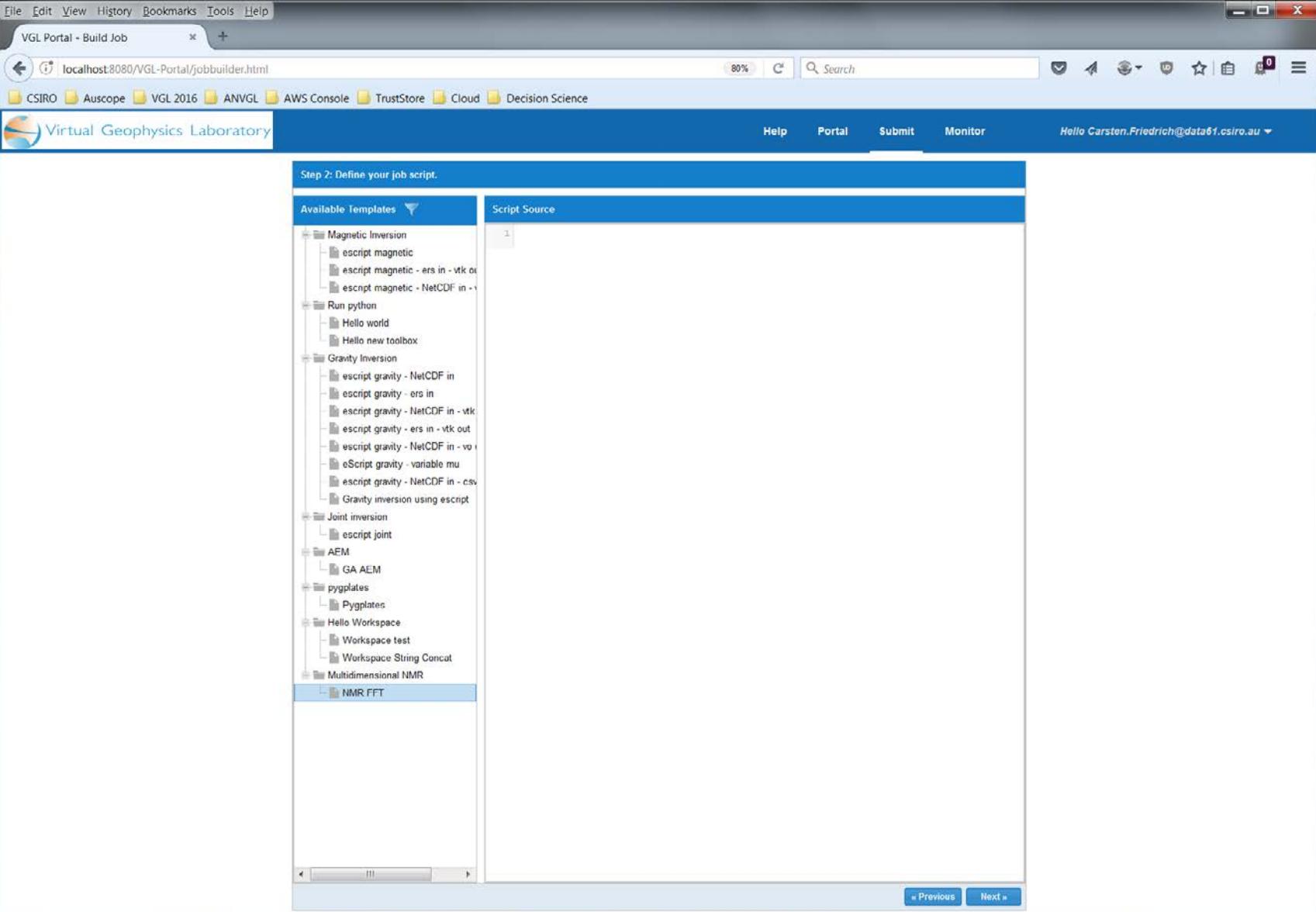
Search term:

Search facets:

Result datasets:



# Browsing available science codes



The screenshot shows a web browser window titled "VGL Portal - Build Job" with the URL "localhost:8080/VGL-Portal/jobbuilder.html". The page is titled "Step 2: Define your job script." and displays a sidebar titled "Available Templates" and a main panel titled "Script Source".

**Available Templates:**

- Magnetic Inversion
  - eScript magnetic
  - eScript magnetic - ers in - vtk out
  - eScript magnetic - NetCDF in - vti out
- Run python
  - Hello world
  - Hello new toolbox
- Gravity Inversion
  - eScript gravity - NetCDF in
  - eScript gravity - ers in
  - eScript gravity - NetCDF in - vtk
  - eScript gravity - ers in - vtk out
  - eScript gravity - NetCDF in - vti out
  - eScript gravity - variable mu
  - eScript gravity - NetCDF in - csv
  - Gravity inversion using eScript
- Joint inversion
  - eScript joint
- AEM
  - GA AEM
- pygplates
  - Pygplates
- Hello Workspace
  - Workspace test
  - Workspace String Concat
- Multidimensional NMR
  - NMR FFT

**Script Source:**

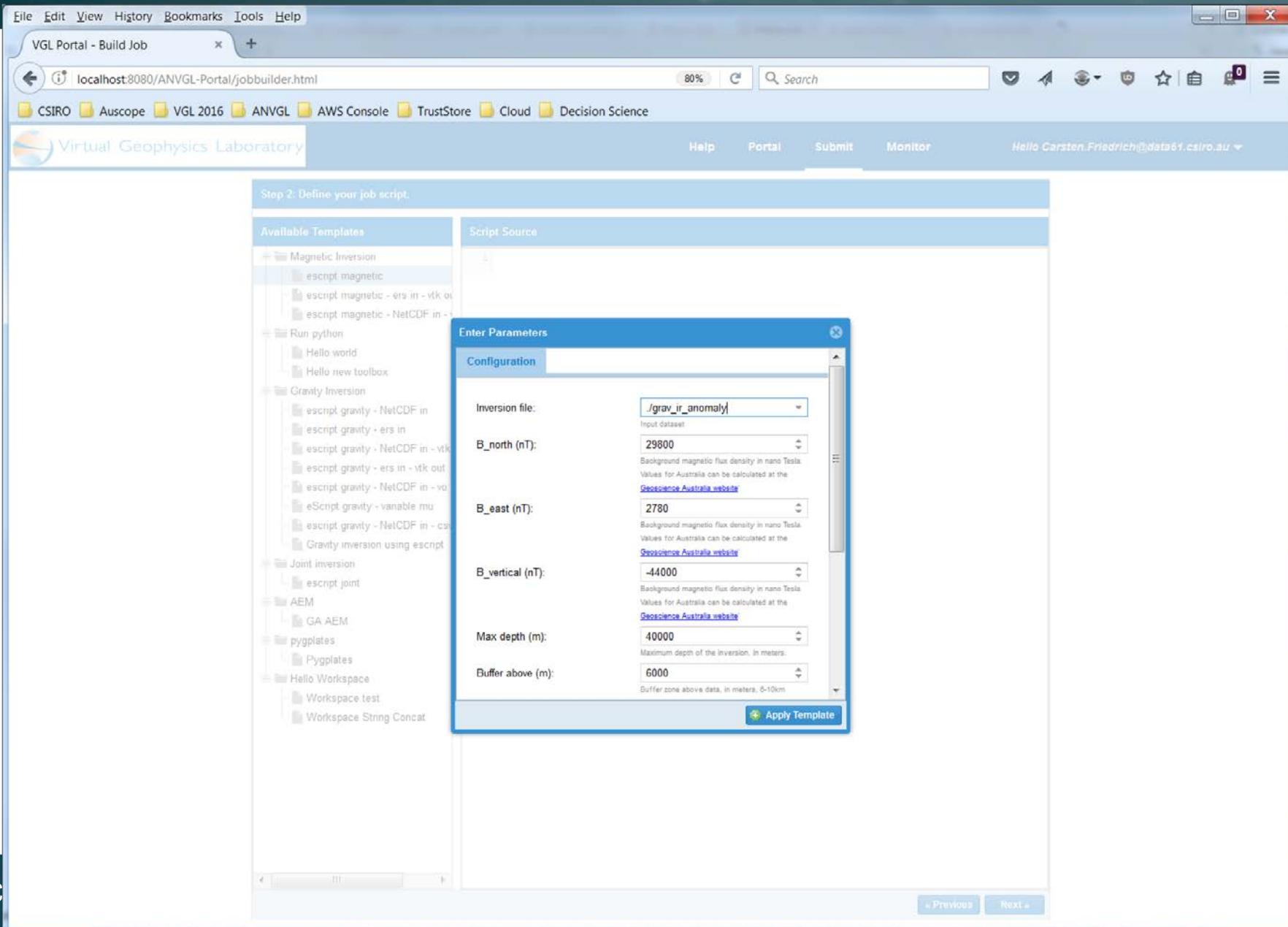
```
1
```

**Navigation:**

- « Previous
- Next »



# Task Parameterization

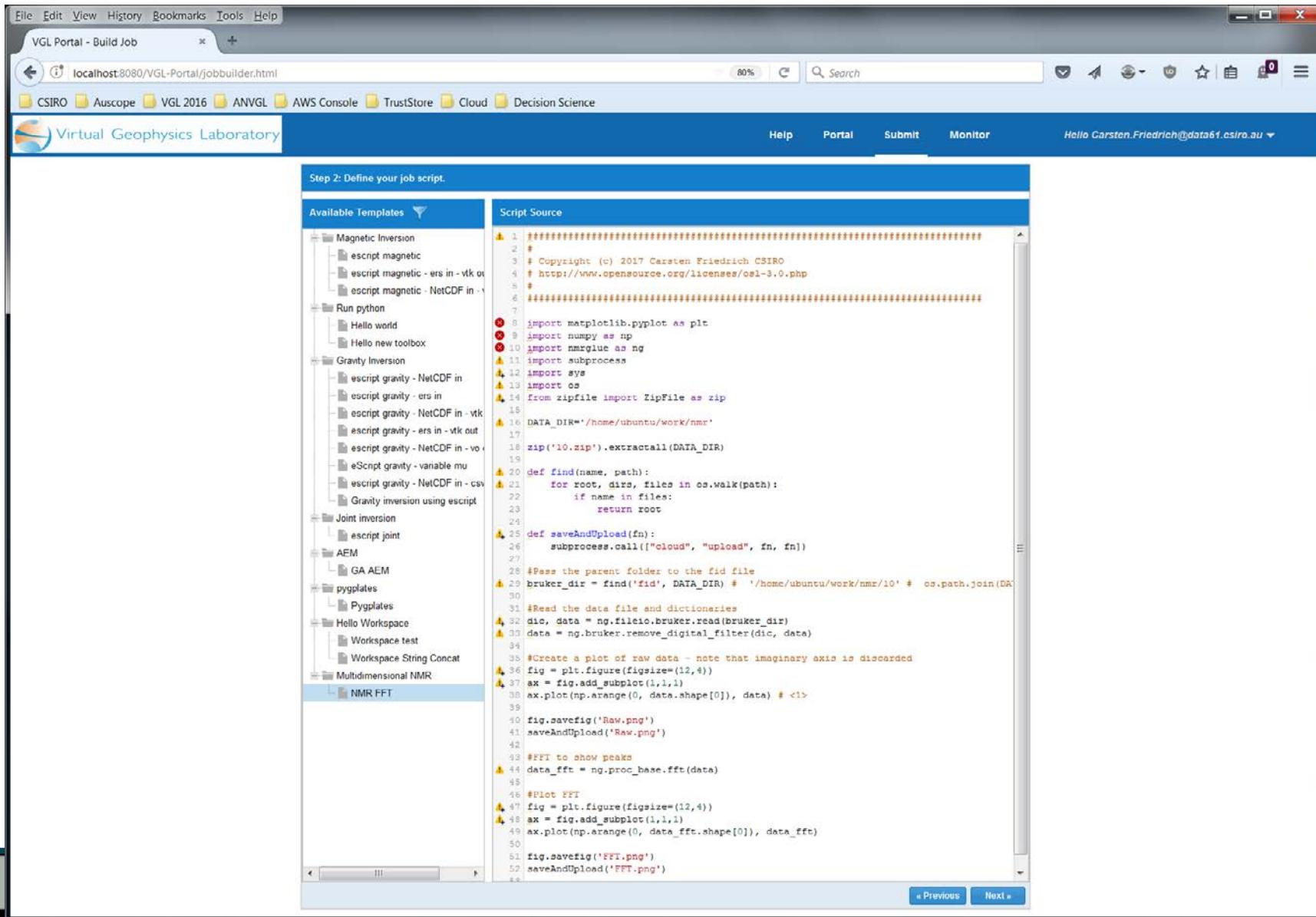


The screenshot shows a web browser window titled "VGL Portal - Build Job" with the URL "localhost:8080/ANVGL-Portal/jobbuilder.html". The page is titled "Step 2: Define your job script." and displays an "Available Templates" sidebar and a "Script Source" editor. A modal dialog box titled "Enter Parameters" is open, showing configuration settings for a gravity inversion task. The parameters are:

- Inversion file: /grav\_ir\_anomaly/
- B\_north (nT): 29800
- B\_east (nT): 2780
- B\_vertical (nT): -44000
- Max depth (m): 40000
- Buffer above (m): 6000

The "Apply Template" button is at the bottom right of the dialog.

## Allowing power users to make last minute adjustments

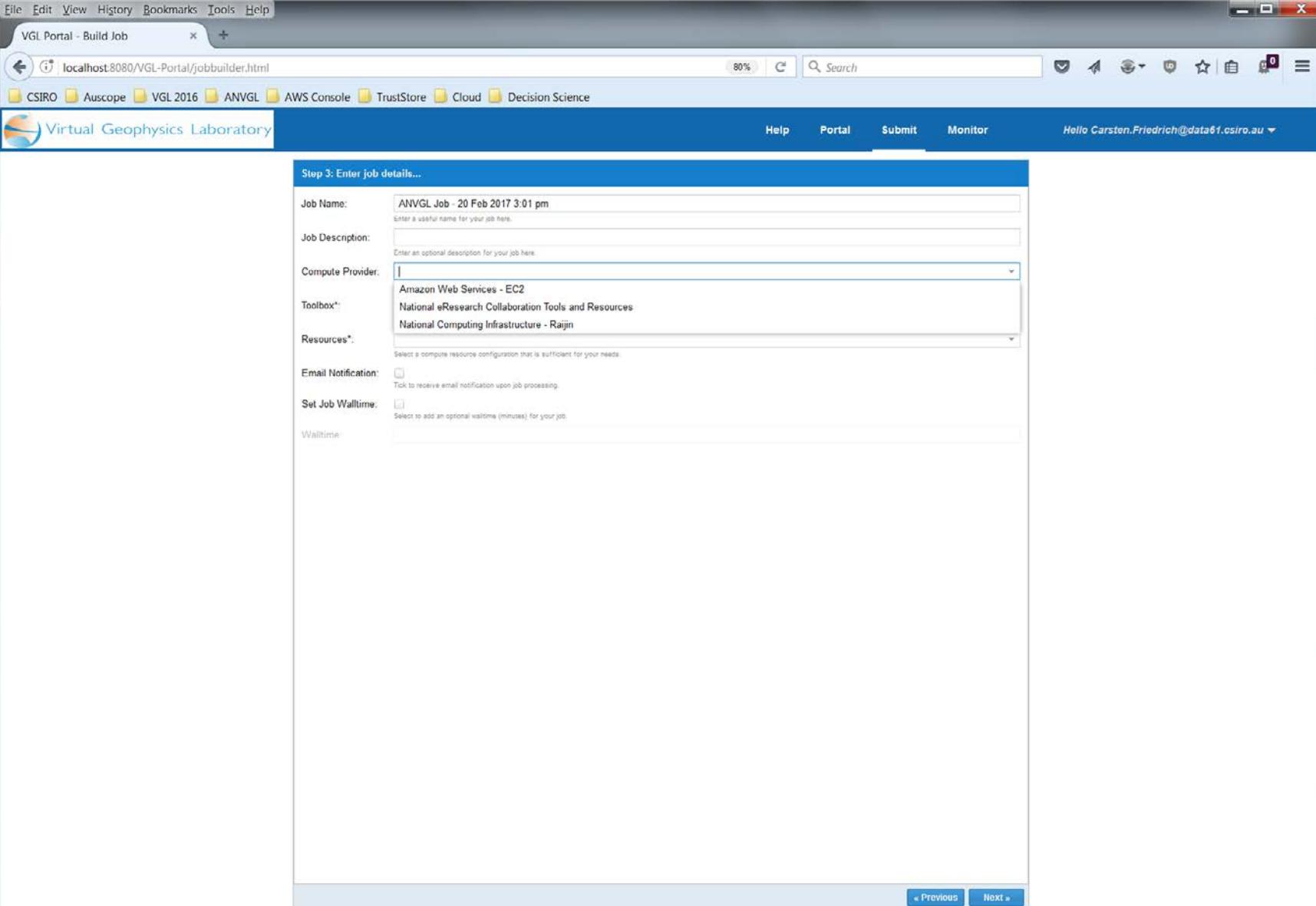


The screenshot shows a web-based interface for building jobs, specifically the "VGL Portal - Build Job" section. The title bar indicates the URL is `localhost:8080/VGL-Portal/jobbuilder.html`. The main content area is titled "Step 2: Define your job script." On the left, there is a sidebar titled "Available Templates" listing various geophysical inversion and processing scripts. The right side contains a large text area for "Script Source" containing a Python script. The script performs several tasks, including extracting a zip file, finding specific files, saving and uploading files to a cloud, reading data and dictionaries, plotting raw data, performing an FFT, and saving the results.

```
#!/usr/bin/python
# Copyright (c) 2017 Carsten Friedrich CSIRO
# http://www.opensource.org/licenses/csl-3.0.php
#
#####
# Import required modules
import matplotlib.pyplot as plt
import numpy as np
import nmrglue as ng
import subprocess
import sys
import os
from zipfile import ZipFile as zip
DATA_DIR='/home/ubuntu/work/nmr'
zip('10.zip').extractall(DATA_DIR)
def find(name, path):
    for root, dirs, files in os.walk(path):
        if name in files:
            return root
def saveAndUpload(fn):
    subprocess.call(["cloud", "upload", fn, fn])
#Pass the parent folder to the fid file
bruks_dir = find('fid', DATA_DIR) # '/home/ubuntu/work/nmr/10' # os.path.join(DATA_DIR, '10')
#Read the data file and dictionaries
dic, data = ng.fileio.bruker.read(bruks_dir)
data = ng.bruker.remove_digital_filter(dic, data)
#Create a plot of raw data - note that imaginary axis is discarded
fig = plt.figure(figsize=(12,4))
ax = fig.add_subplot(1,1,1)
ax.plot(np.arange(0, data.shape[0]), data) # <1>
fig.savefig('Raw.png')
saveAndUpload('Raw.png')
#FFT to show peaks
data_fft = ng.proc_base.fft(data)
#Plot FFT
fig = plt.figure(figsize=(12,4))
ax = fig.add_subplot(1,1,1)
ax.plot(np.arange(0, data_fft.shape[0]), data_fft)
fig.savefig('FFT.png')
saveAndUpload('FFT.png')
```



# Selecting compute platform and resources



The screenshot shows a web browser window titled "VGL Portal - Build Job". The address bar displays "localhost:8080/VGL-Portal/jobbuilder.html". The main content area is titled "Step 3: Enter job details...". The form fields include:

- Job Name:** ANVGL Job - 20 Feb 2017 3:01 pm
- Job Description:** (empty text area)
- Compute Provider:** A dropdown menu currently set to "Amazon Web Services - EC2", with other options like "National eResearch Collaboration Tools and Resources" and "National Computing Infrastructure - Raijin" listed.
- Toolbox\*:** A dropdown menu listing "Amazon Web Services - EC2", "National eResearch Collaboration Tools and Resources", and "National Computing Infrastructure - Raijin".
- Resources\*:** A dropdown menu listing "Select a compute resource configuration that is sufficient for your needs."
- Email Notification:** A checkbox labeled "Tick to receive email notification upon job processing." (unchecked)
- Set Job Walltime:** A checkbox labeled "Select to add an optional walltime (minutes) for your job." (unchecked)
- Walltime:** An input field for specifying walltime.

At the bottom of the form are "Previous" and "Next" buttons.



# Monitoring progress and retrieving results

Screenshot of the VGL Portal - Monitor Jobs interface showing job monitoring and results retrieval.

The interface includes:

- File Edit View History Bookmarks Tools Help** menu bar.
- VGL Portal - Monitor Jobs** tab.
- localhost:8080/VGL-Portal/joblist.html** address bar.
- CSIRO Auscope VGL 2016 ANVGL AWS Console TrustStore Cloud Decision Science** navigation links.
- Virtual Geophysics Laboratory** logo and title.
- Your Jobs** table listing jobs with columns: Job Name, Status, and Actions.
- ANVGL Job - 20 Feb 2017 2:11 pm** job details panel:
  - Done** status
  - Job ID: 62 Melbourne/61e8f9eb-b521-4d92-ba97-d098badd0961**
  - Status: Done**
  - Instance ID: Melbourne/639b8b2a-a5a6-4aa2-8592-ca765ee7af63**
  - Submitted: 55 minute(s) ago**
  - Actions** dropdown menu with **Name** and **Details** options.
  - Files in cloud storage (6 Items)** list:
    - 10.zip (532 KB)
    - FFT.png (19.4 KB)
    - Raw.png (24.2 KB)
    - activity.ttl (8.1 KB)
    - vl.sh.log (9.4 KB)
    - vl\_script.py (1.4 KB)
- Help Portal Submit Monitor** navigation tabs.
- Hello Carsten.Friedrich@data61.csiro.au** user information.
- 1e8** scale factor for the plot.
- FFT plot** showing signal amplitude versus frequency or time, with three sharp peaks around 11500, 13000, and 14500 units.
- Add Folder Refresh** buttons at the bottom.

# Task Results

VGL Portal - Monitor Jobs    +

localhost:8080/ANVGL-Portal/joblist.html

CSIRO Auscope VGL 2016 ANVGL AWS Console TrustStore Cloud Decision Science

 Virtual Geophysics Laboratory

Your Jobs

Job Name	Status
ANVGL Job - 08 Feb 2017 10...	Provisioning
ANVGL Job - 07 Feb 2017 12...	Done
ANVGL Job - 07 Feb 2017 1...	Saved
ANVGL Job - 03 Feb 2017 1...	Done
ANVGL Job - 03 Feb 2017 1...	Done

ANVGL Job - 07 Feb 2017 12:55 pm

**Done 45 2250966.r-man2 ncpus=64&jobfs=16gb&mem=128gb 21 hour(s) ago**

Status Job ID Instance ID Instance Type Submitted

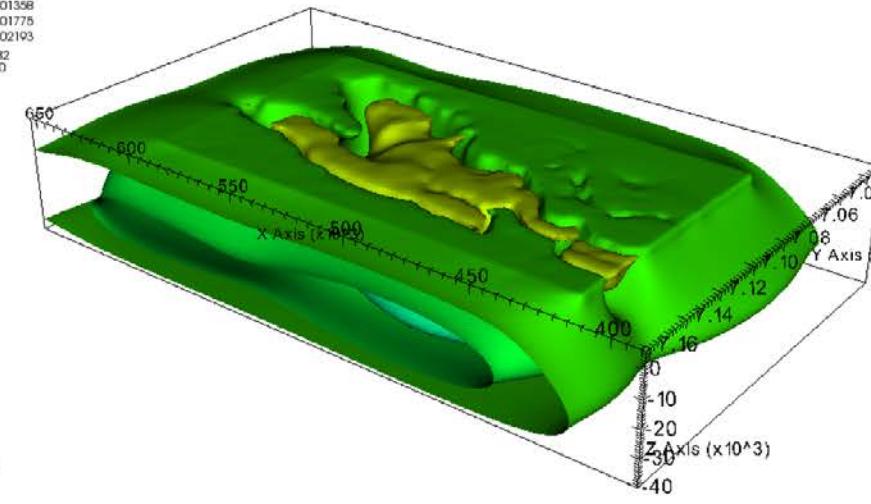
**Actions**

Name	Details
grav_complete_bou...	Service call to dap...
Files in cloud storage (8 items)	
ncl-run.job	2.8 KB
ncl-download.job	2.1 KB
ncl-util.sh	524 bytes
result.silo	82.5 MB
activity.ttl	1.1 KB
vl_script.py	3.2 KB
<b>result-visit.png</b>	147.3 KB
vl.sh.log	265.8 KB

DB: result.silo  
Cycle: 0 Time:0

Contour Var: susceptibility

Max: 0.01982 Min: -0.02610

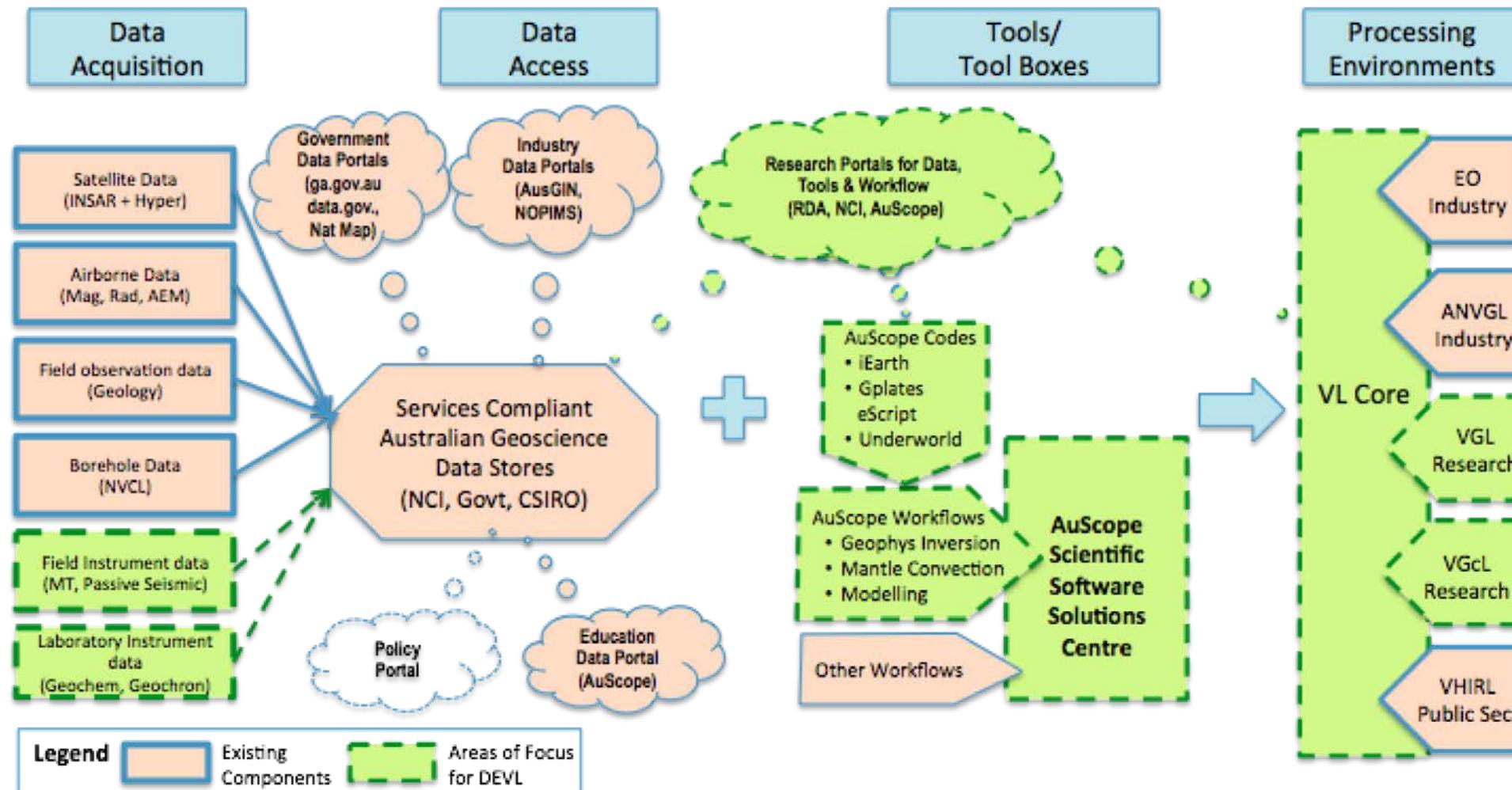


X Axis (X) Y Axis (Y) Z Axis (x10<sup>3</sup>)

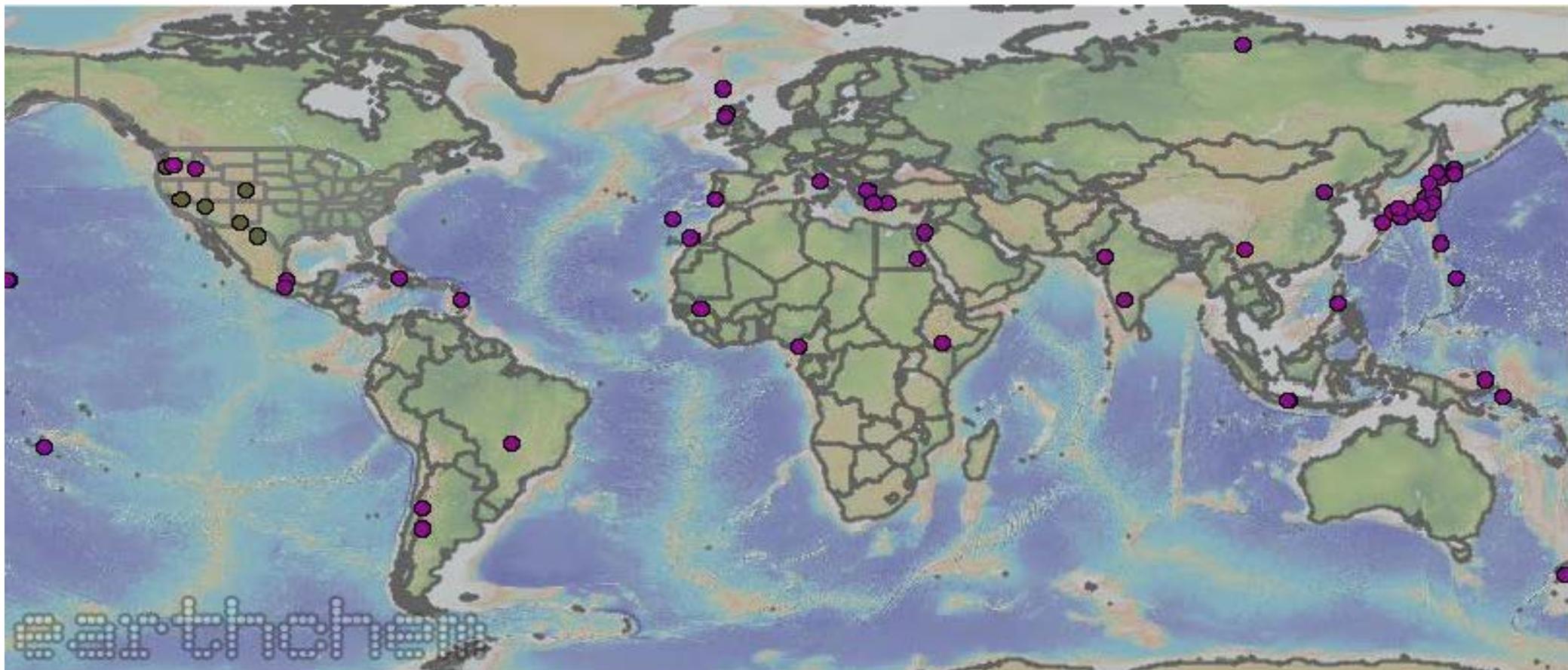
user: cxf599  
Tue Feb 7 14:06:4

Add Folder Refresh

## AuScope (data enhanced) Virtual Research Environment (AVRE)



## What is the IGSN?



“M-1” is obviously not a unique name.

- A registry of identifiers that are guaranteed to be **globally unique** and **persistent**.
- It facilitates **online discovery and access** to physical samples
  - Web application and programmatic access to sample **metadata catalogues**
  - **Network** with sample repositories and data centres
- Ensures **preservation and access** of sample metadata (but not samples)
- Aids in the identification of samples in the literature
- Currently has **6 million IGSN's** registered from agents in 8 countries

Try it out: <http://doi.org/10273/AU288>

# What IGSN can be used for

- Geological samples and other materials (rocks, water, biological materials, ...)
- Collections (groupings of samples)
- Sampling features (boreholes, outcrops, ...)
- Samples can be linked to each other through the “related identifier” metadata element.





**Australian Government**  
**Geoscience Australia**



- The three current IGSN members in Australia:
  - CSIRO – Government Research Organisation
  - Geoscience Australia – Government Geological Surveys
  - Australian National Data Service – Research Sector
- Involvement with Australian National Data Service
  - Help provide infrastructure for the research community
  - Currently provides a catalogue for data, software
  - Mints DOI's
  - Assisted with community building (e.g. workshops) to introduce sample identifiers to other institutions and other science disciplines
  - In 2018 will be working with Curtin, Macquarie and Melbourne Universities

- Chair of the Academy in Science, Data in Science Committee
- On the AGU Data Management Advisory Board
- On the Steering Committee of the ICSU/CODATA Commission on Scientific Standards for Integration of Research Data
- On the Steering Committee for the AGU/Research Data Alliance Project for FAIR and Open Data in Research Publications
- On the AuScope Strategic Reference Panel