NCI

NATIONAL COMPUTATIONAL INFRASTRUCTURE

Supporting Data Reproducibility at NCI Using the Provenance Capture System

Jingbo Wang, Nick Car, Wei Si, Edward King, Ben Evans, Lesley Wyborn

Problem statement

More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments. Those are some of the telling figures that emerged from Nature's survey of 1,576 researchers who took a brief online questionnaire on reproducibility in research.

Source: http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

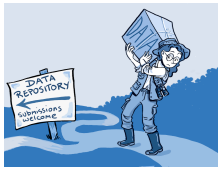
Who are we?

NCI is high performance computer centre, hosting 10PB research data for sharing among research community. - HPC

NCI is building the National Environmental Research Data Interoperability Platform (NERDIP) to enable data intensive science. - HPD

What do we do?

User generate/ transfer data

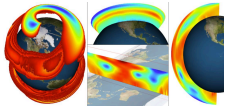


Data Manager fill DMP and create catalogue

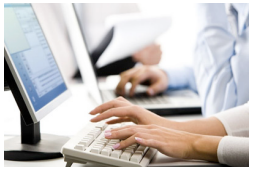
Super computer users

Paper and Data are published

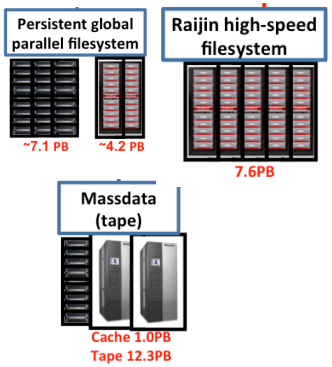
Data visualization



Data share and re-use



NCI provides user with Data as a Service

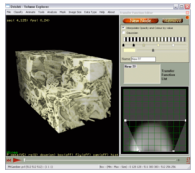


NCI provides fast data storage

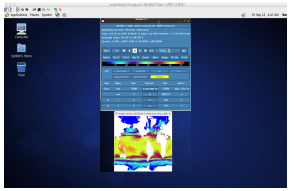
Data Management Portal

HPC

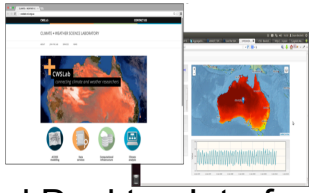
Data Curation, Publish, Citation



NCI Vislab



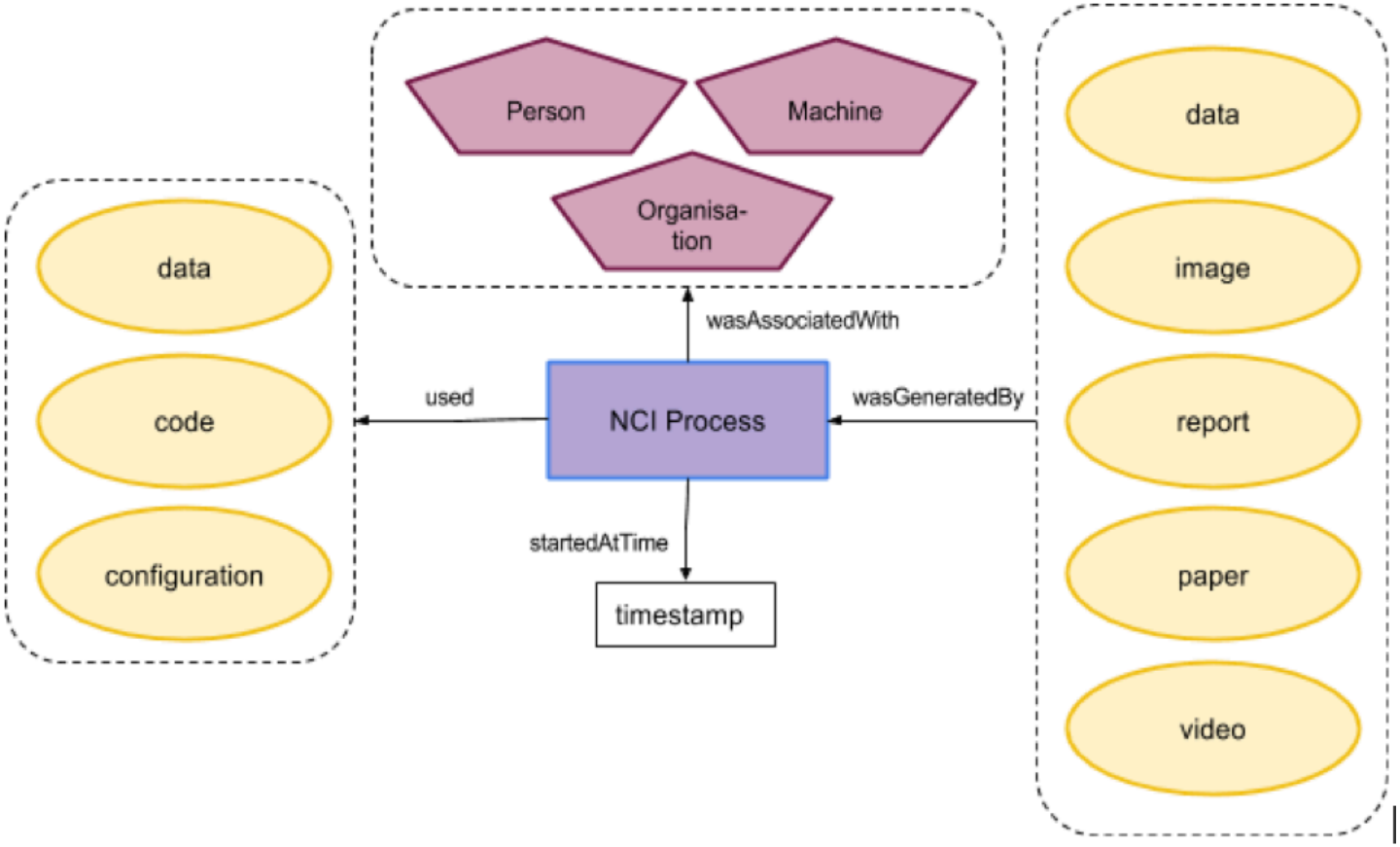
Web-time analytics software



Virtual Desktop Interface, Virtual Laboratory, and other services

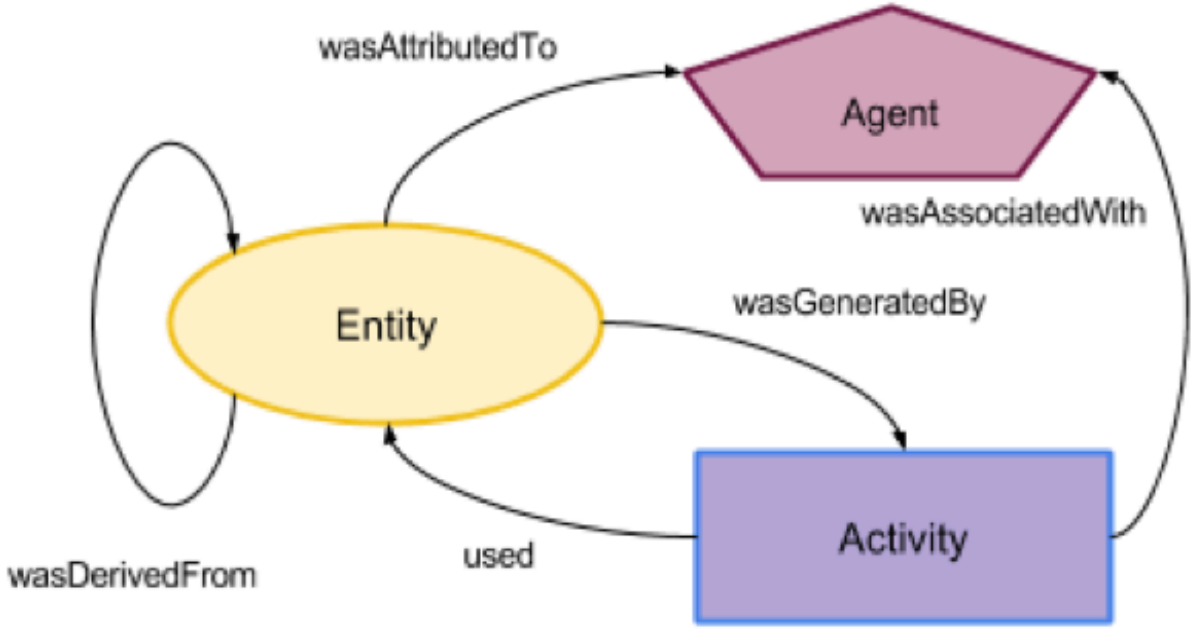
Proposed solution

Provenance Capture System: to support traceable, reproducible, and machine actionable workflow



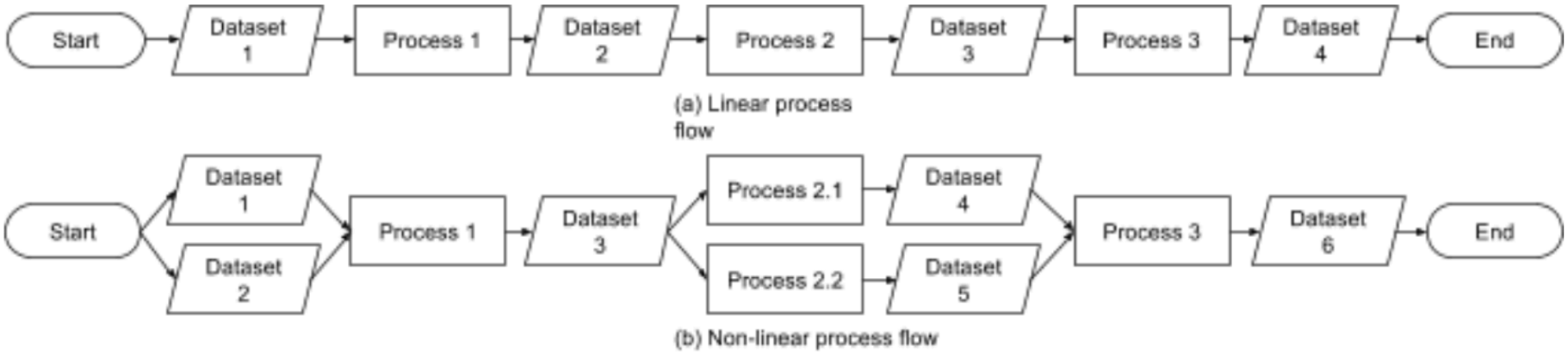
Central system: PROMS

The basic classes and relationships of PROV-DM, drawn in accordance with the PROV Ontology.

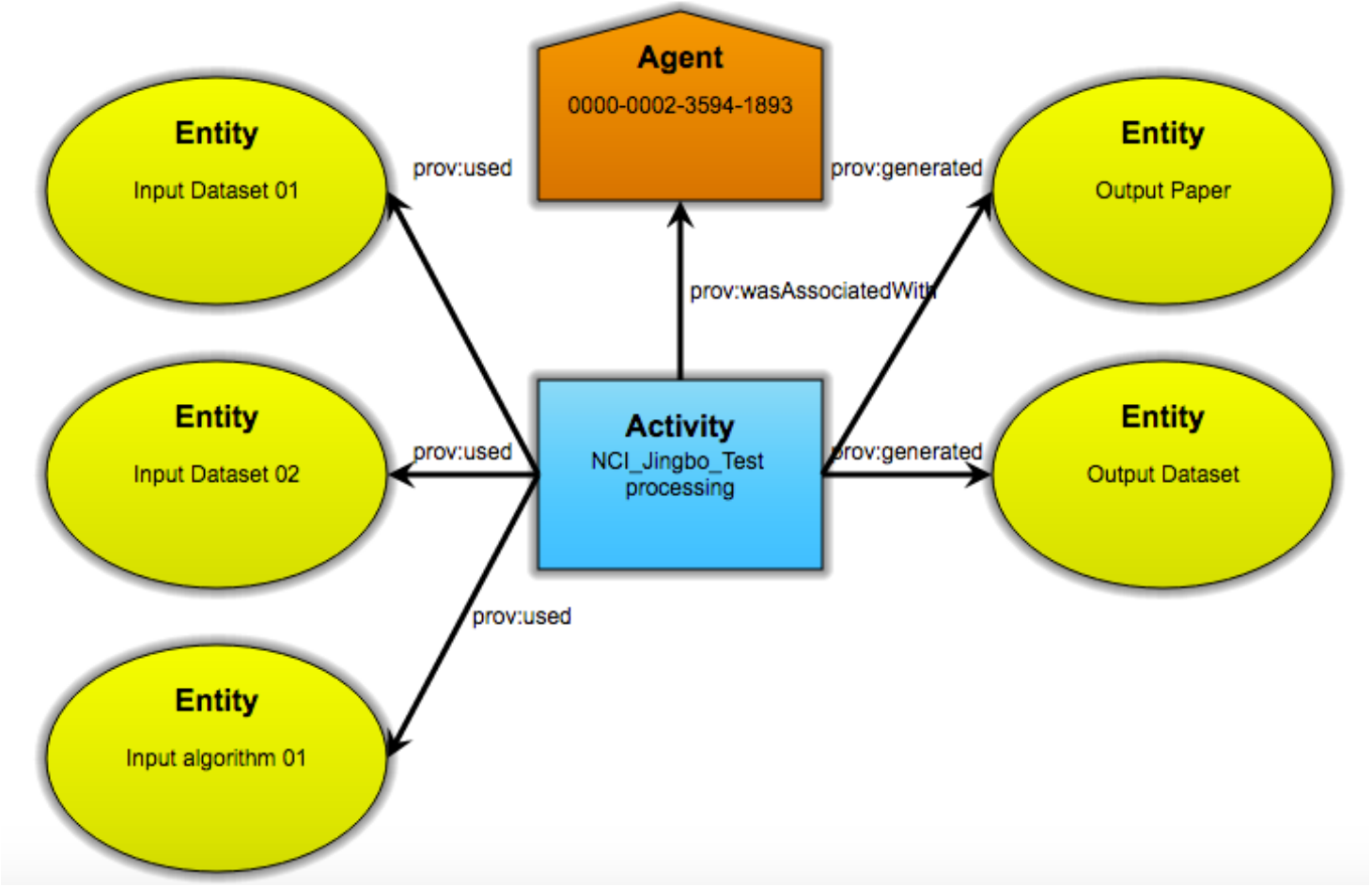


The information is stored as RDF documents in an RDF graph database, which is then published via a HTTP-based API as a combination of web pages and data.

Flowchart of different types of workflows captured by PROMS

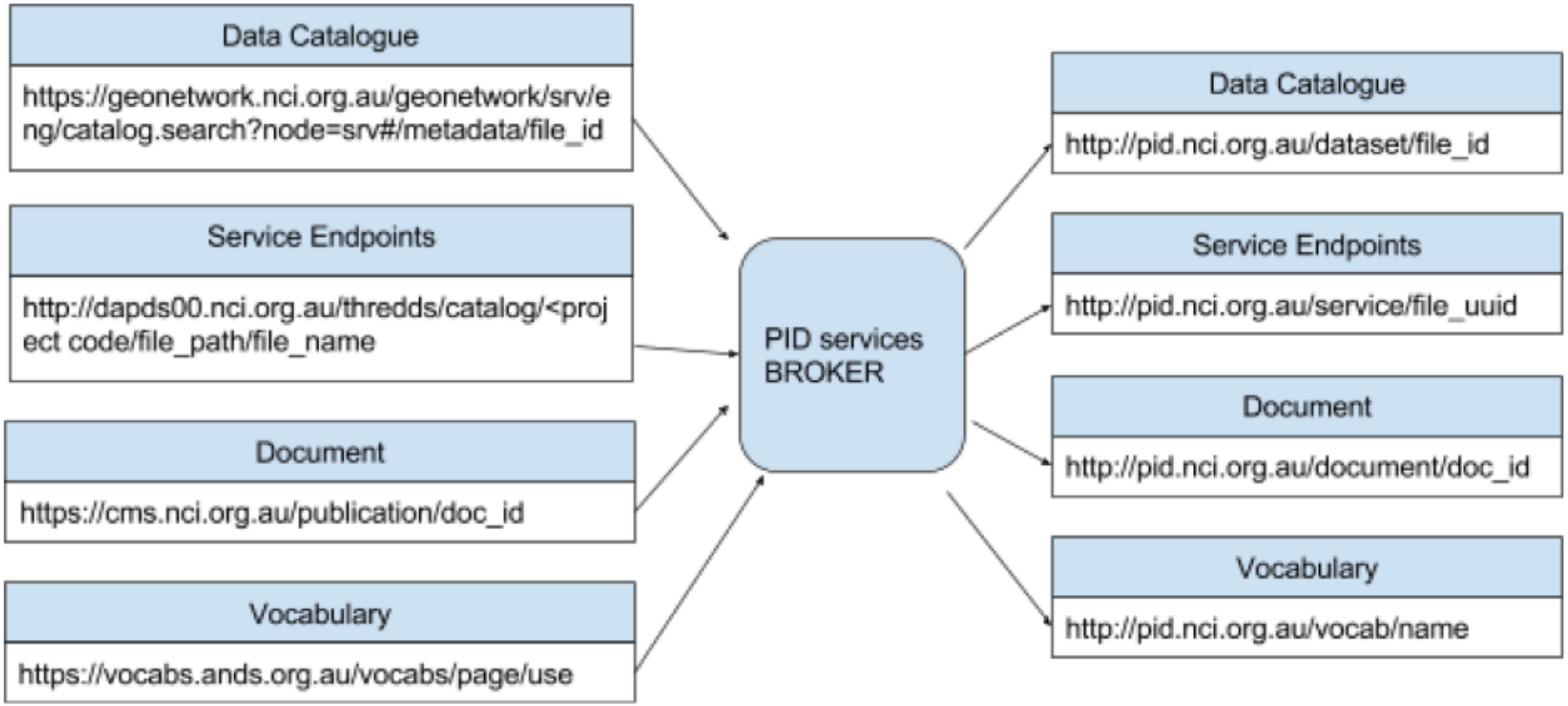


Proof of concept workflow



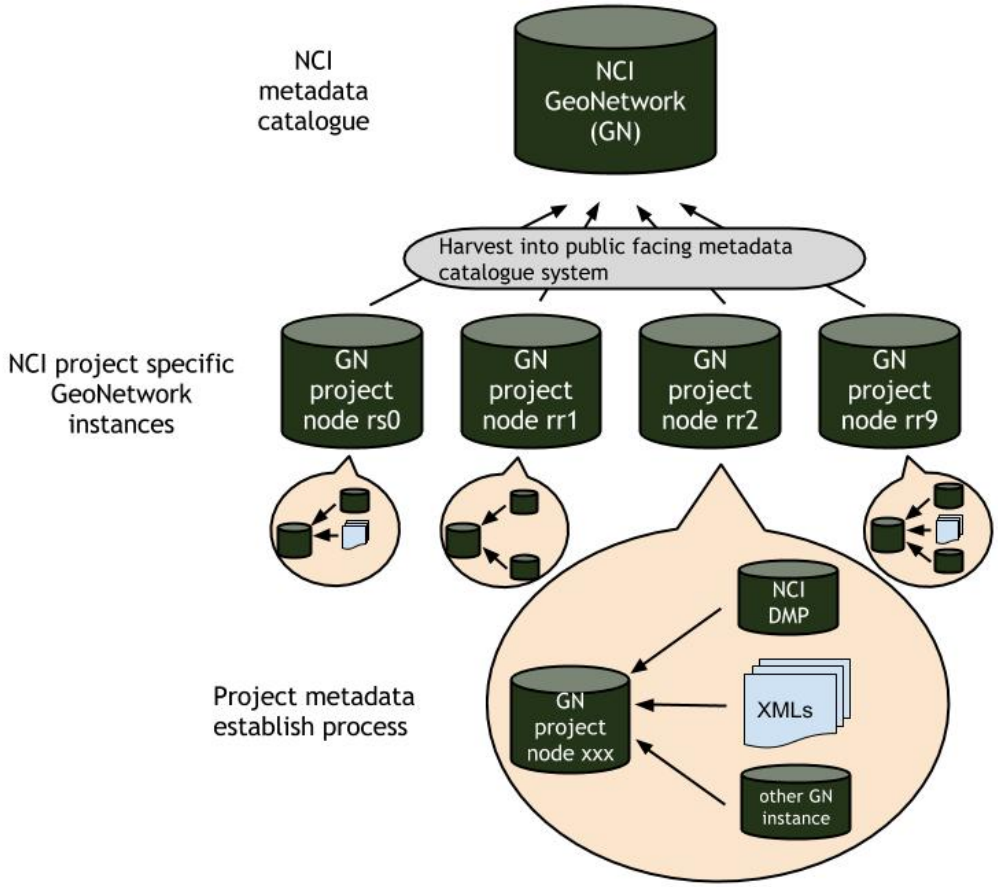
Supporting infrastructure: PID

Persistent Identifier Services



Supporting infrastructure: catalogue

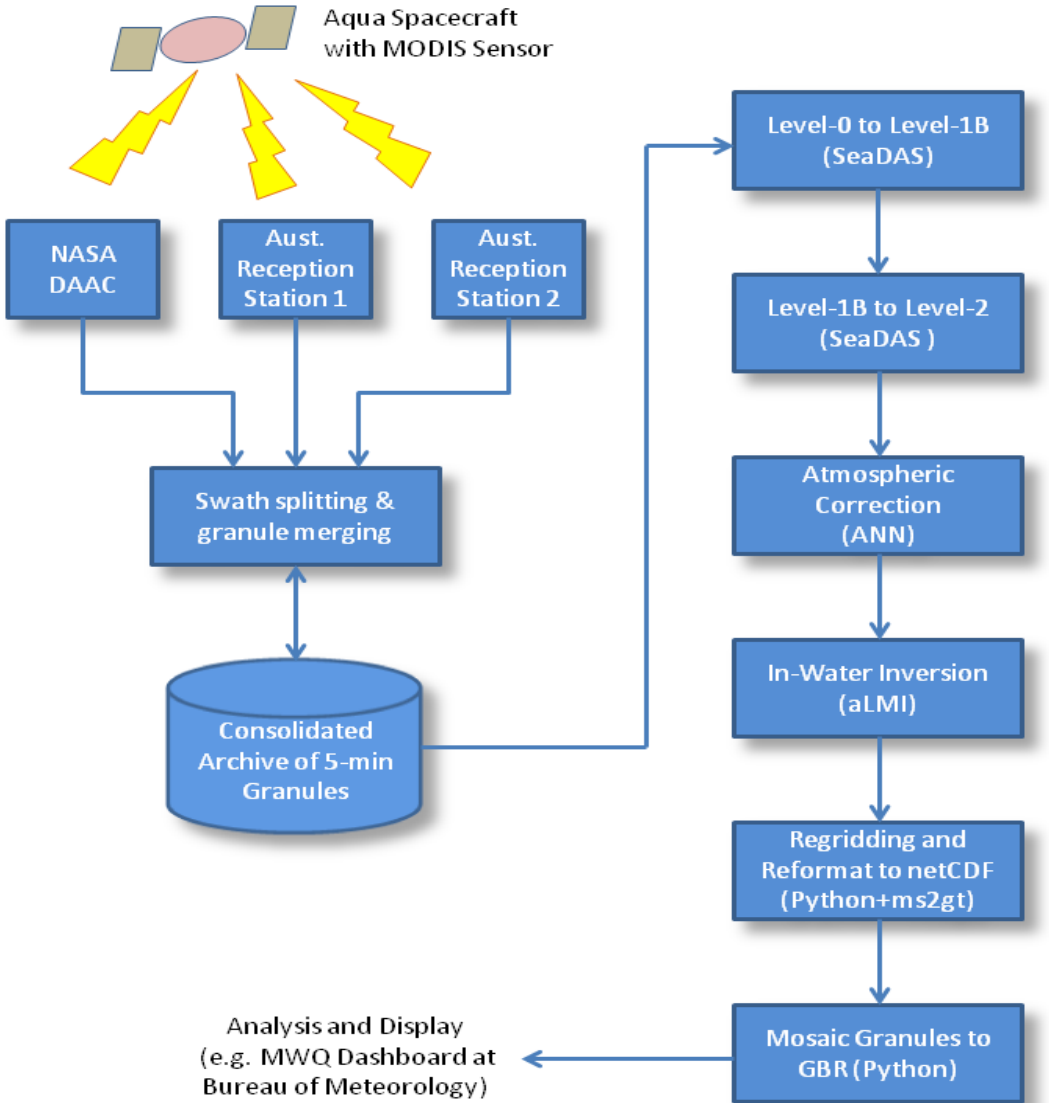
NCI GeoNetwork architecture https://geonetwork.nci.org.au



Supporting infrastructure: doc repository

NCI document repository https://cms.nci.org.au

Ocean Colour Data Processing Flow



1.Assembling the data (Level-0 granules);
2.Calibration and geolocation (Level-0 to Level-1b);
3.Generation of ancillary fields (Level-1b to Level-2);
4.Data selection and masking;
5.Atmospheric correction;
6.In-water inversion (deriving concentrations of optically active substances);
7.Regridding (conversion of satellite grids to map projection);
8.Mosaicing (joining multiple satellite images to provide more complete coverage);
9.Data distribution and access.

Provenance report of Ocean Colour Case

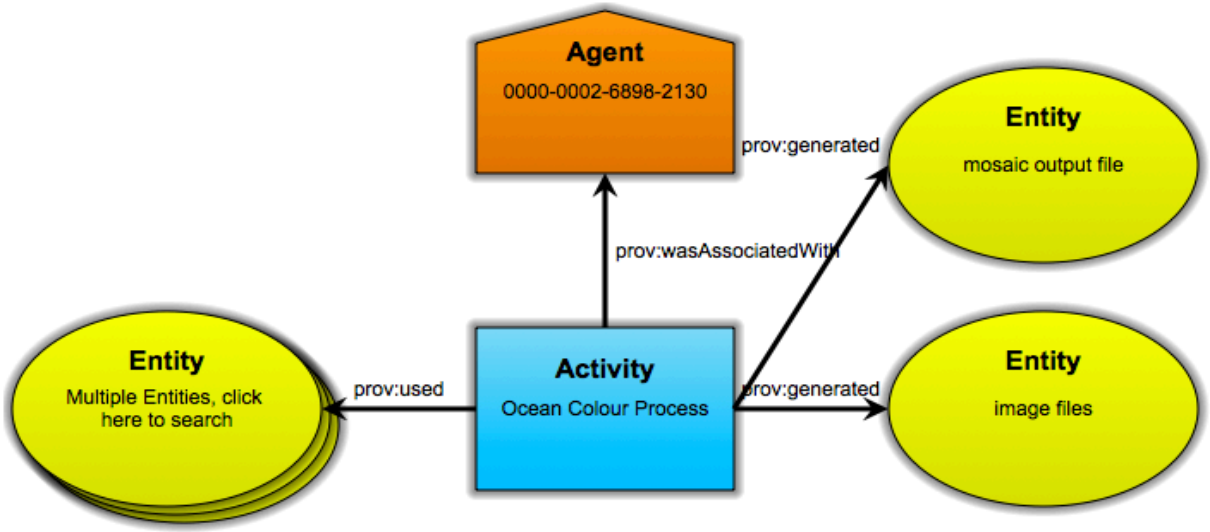
An Activity

URI: file:///opt/proms/raijin/g/data/u83/code/ereefs/legacy/proms/promReport_OC/

html | rdf/turtle

Title:	Ocean Colour Process
Started at:	2016-07-01T11:20:56
Ended at:	2016-07-01T11:20:56
Was associated with:	http://orcid.org/0000-0002-6898-2130
Report:	Ocean Colour External Report - 2016-07-01 11:20:56.261893

Neighbours view



Staff Resource Effort

The staff resources needed to establish the Ocean Colour reporting were:

•Part-time work over several months to implement the central information services;

•Several days work from three people (an OC product owner and modellers) to model the Ocean Colour process as per PROV-O;

•A week to implement logging within the OC workflow that could be used to establish PROV-O reports from each workflow run;

•storage location URIs generated by dataset registry and PID services

•A day or two to schedule reporting to work continuously.

Thank you

Questions?