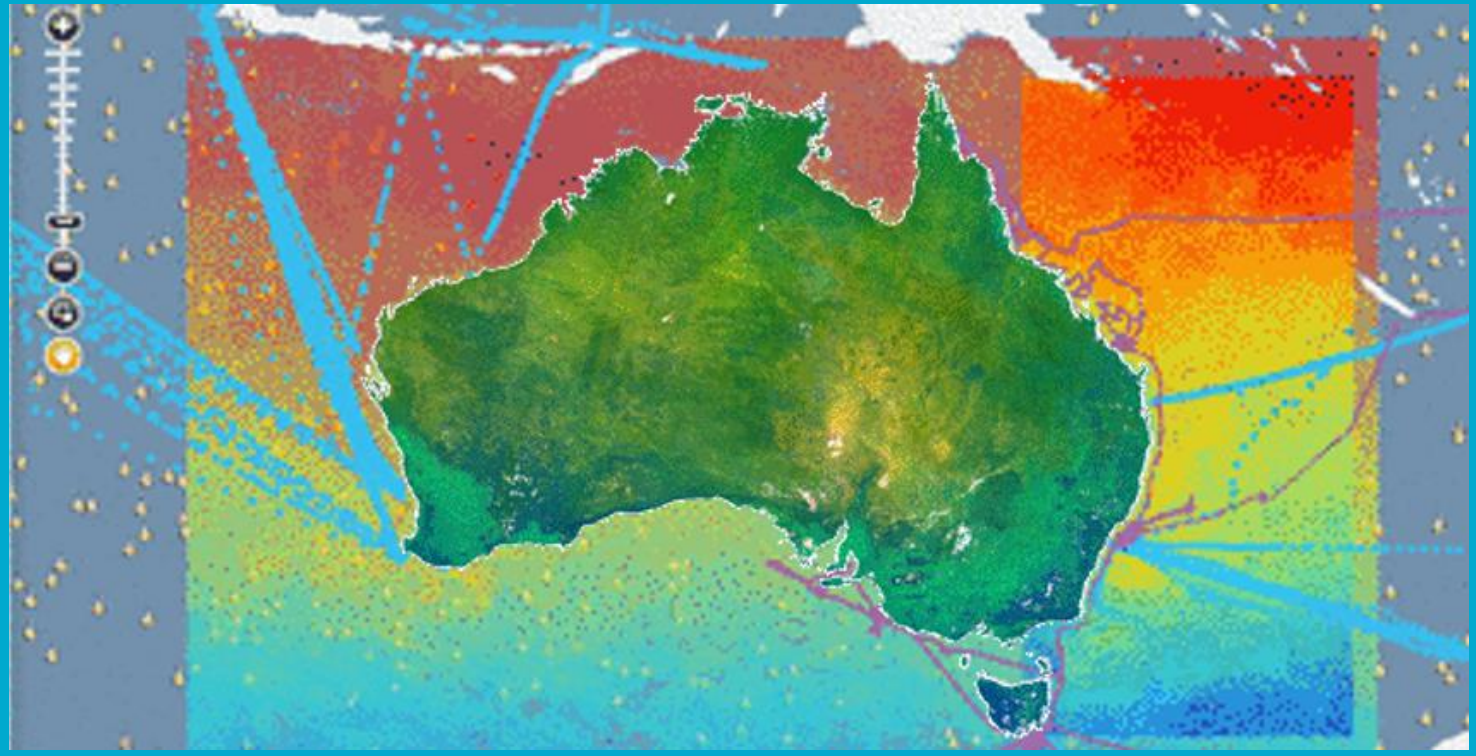


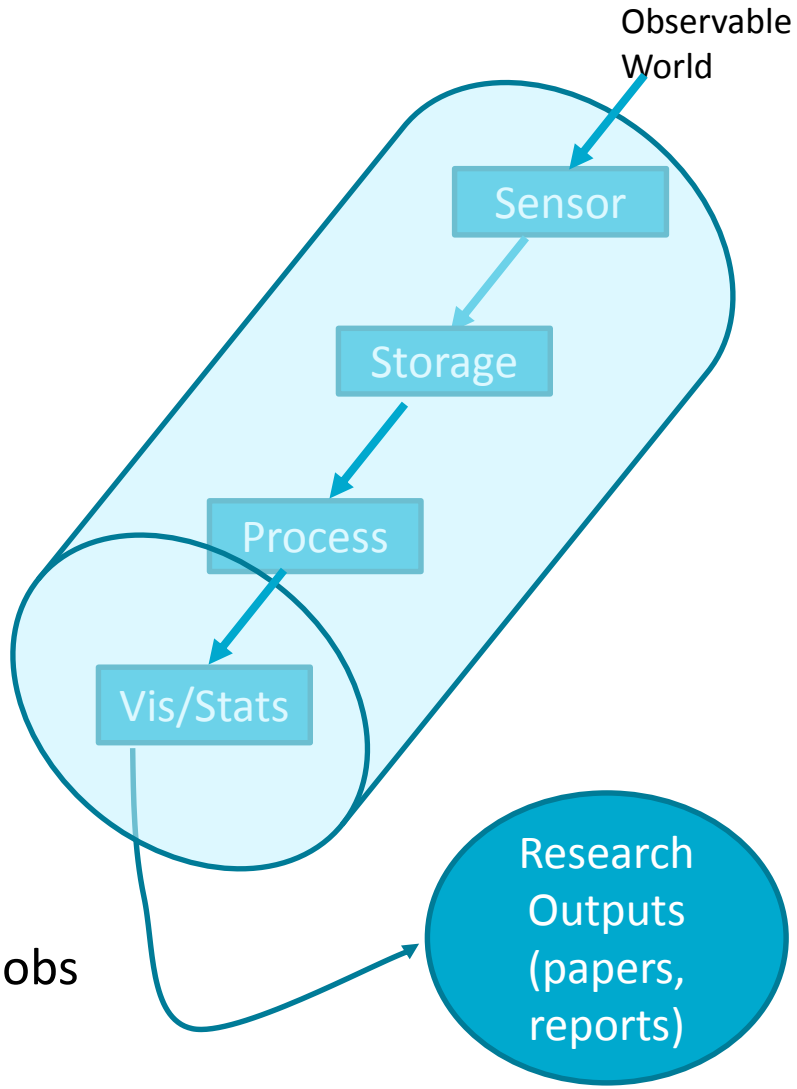
Remotely Sensible Data Infrastructures

Experiences from AWAP, WIRADA, TERN-Auscover, IMOS-SRS, eReefs….

Edward King| CSIRO Climate Science Centre

5 December 2017

Motivation



- Loss of context, or failure to store (& keep) obs and results, or inadequately documented processes, makes reuse of results, or, even worse, reproduceability - impossible

Eg Solar PV Data

chart.csv - Excel

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW

K1

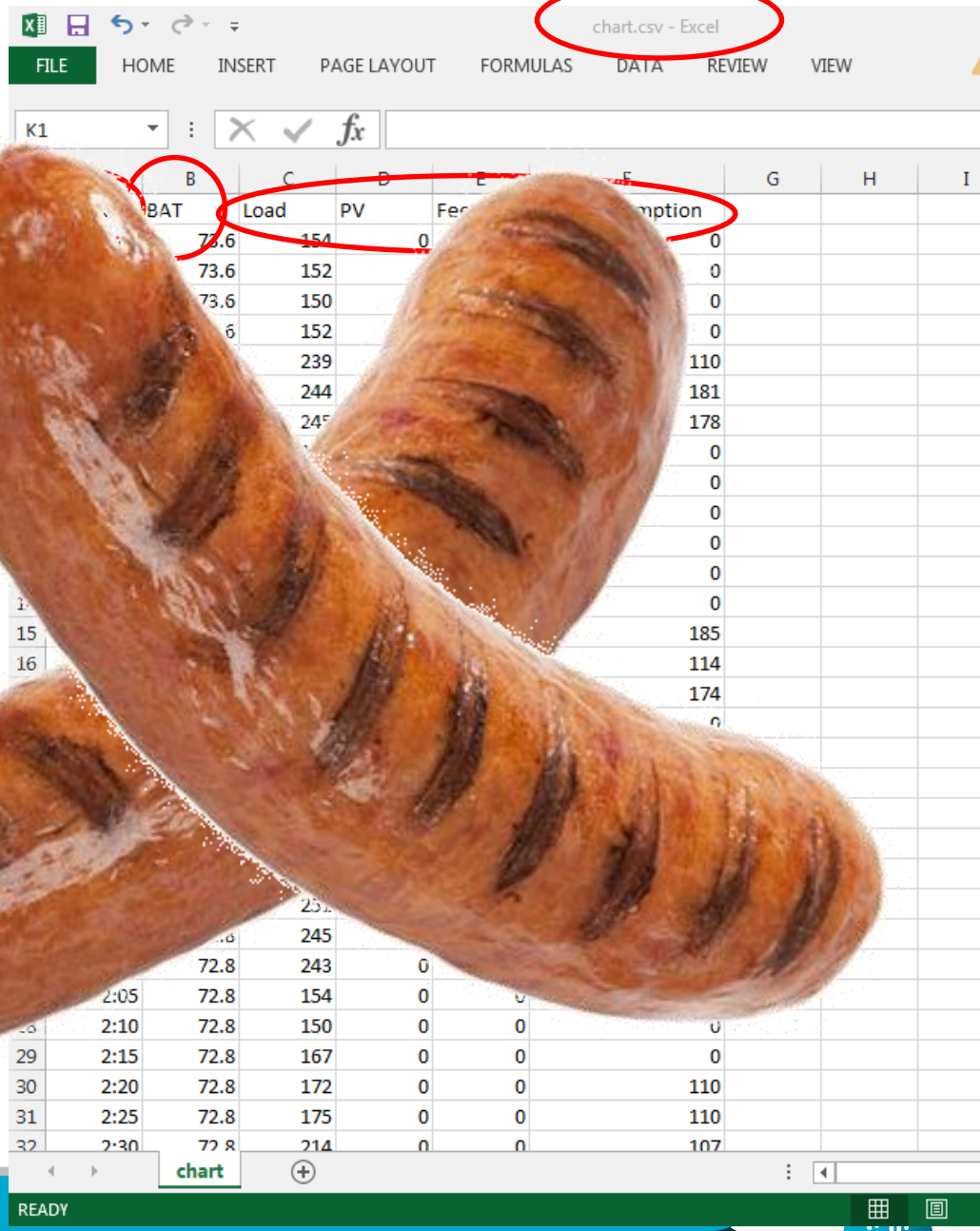
	A	B	C	D	E	F	G	H	I
1	Category	BAT	Load	PV	Feed-In	Grid Consumption			
2	0:00	73.6	154	0	0	0			
3	0:05	73.6	152	0	0	0			
4	0:10	73.6	150	0	0	0			
5	0:15	73.6	152	0	0	0			
6	0:20	73.6	239	0	0	110			
7	0:25	73.2	244	0	0	181			
8	0:30	73.2	245	0	0	178			
9	0:35	73.2	152	0	0	0			
10	0:40	73.2	152	0	0	0			
11	0:45	73.2	153	0	0	0			
12	0:50	73.2	153	0	0	0			
13	0:55	73.2	154	0	0	0			
14	1:00	73.2	153	0	0	0			
15	1:05	73.2	253	0	0	185			
16	1:10	73.2	419	0	0	114			
17	1:15	73.2	244	0	0	174			
18	1:20	73.2	153	0	0	0			
19	1:25	73.2	152	0	0	0			
20	1:30	73.2	193	0	0	0			
21	1:35	73.2	154	0	0	0			
22	1:40	73.2	152	0	0	0			
23	1:45	72.8	154	0	0	0			
24	1:50	72.8	251	0	0	180			
25	1:55	72.8	245	0	0	182			
26	2:00	72.8	243	0	0	173			
27	2:05	72.8	154	0	0	0			
28	2:10	72.8	150	0	0	0			
29	2:15	72.8	167	0	0	0			
30	2:20	72.8	172	0	0	110			
31	2:25	72.8	175	0	0	110			
32	2:30	72.8	214	0	0	107			

chart

READY

Eg Solar PV Data

Wurst Practice



nature International weekly journal of science

Search Go

Advanced search

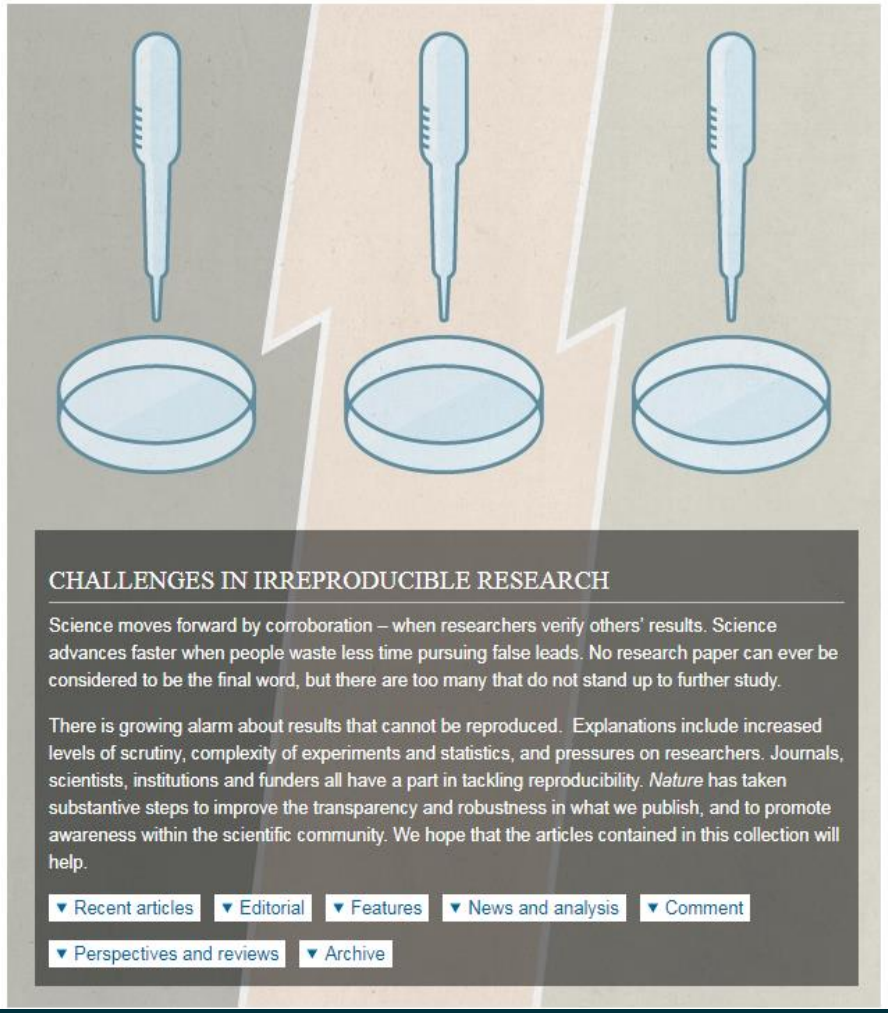
Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Specials and supplements archive > Challenges in irreproducible research

E-alert RSS Facebook Twitter

SPECIAL

See all specials



CHALLENGES IN IRREPRODUCIBLE RESEARCH

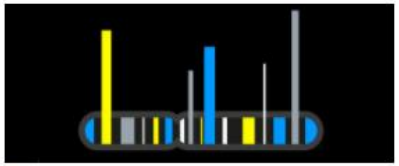
Science moves forward by corroboration – when researchers verify others' results. Science advances faster when people waste less time pursuing false leads. No research paper can ever be considered to be the final word, but there are too many that do not stand up to further study.

There is growing alarm about results that cannot be reproduced. Explanations include increased levels of scrutiny, complexity of experiments and statistics, and pressures on researchers. Journals, scientists, institutions and funders all have a part in tackling reproducibility. *Nature* has taken substantive steps to improve the transparency and robustness in what we publish, and to promote awareness within the scientific community. We hope that the articles contained in this collection will help.

▼ Recent articles ▼ Editorial ▼ Features ▼ News and analysis ▼ Comment

▼ Perspectives and reviews ▼ Archive

Gene count



The most popular genes in the human genome

A tour through the most studied genes in biology reveals some surprises.

Ad closed by Google

Recent | Read | Commented

1. **Gravity signals could speedily warn of big quakes and save lives**
   Nature | 30 November 2017
2. **Huge haul of rare pterosaur eggs excites palaeontologists**
   Nature | 30 November 2017
3. **Health agency reveals scourge of fake drugs in developing world**
   Nature | 29 November 2017

Premise

- If we can capture data and data products in a way that records their meaning and context, and preserves them so that they are accessible in future, then they can be
  - Reused (eg repeating work for verification)
  - Repurposed (eg Other investigations, or as context for other work)
- Ie. More value out of original observations

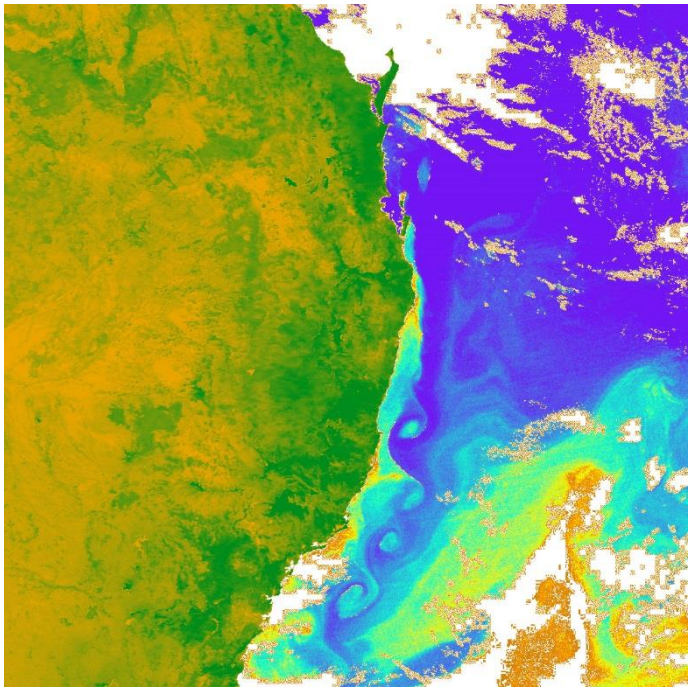
- If we can do it in a consistent (standards-based) way, then we can greatly improve the ease with which these benefits can be realised

- As a fully automated and high volume* digital data acquisition system, satellite remote sensing is a field that illustrates this in spades.

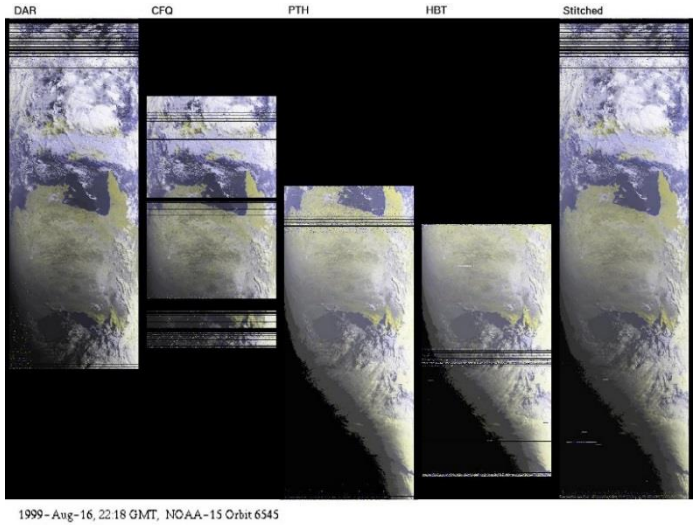
* Data volume is a pervasive issue with RS

Example 1: NOAA/AVHRR

- 1km daily imaging 1981 to now
- From 1992, all Australian imagery merged and stored in consistent format, online.
  - Straightforward to use 25 year archive
- Instrument now superseded by MODIS, VIIRS etc.



- BUT: When you are looking for slow long-term trends, what is the most useful data?
- 1981-1991 (although patchy <1986)



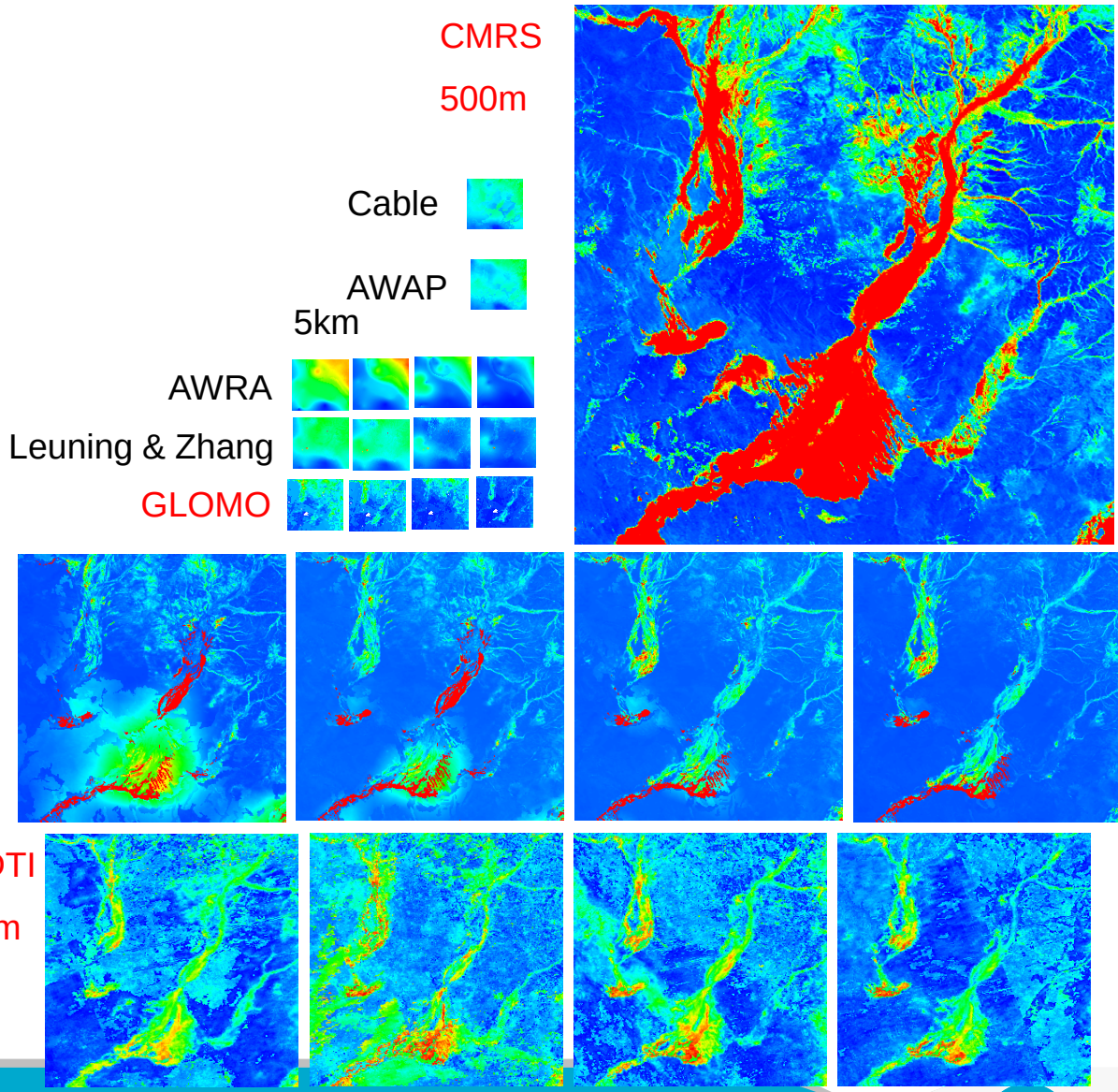
1999-Aug-16, 22:18 GMT, NOAA-15 Orbit 6545

- We have most of it on tape, but we got most of it off
- Formats unknown.
- Creators have all retired.
- Will the earliest data ever be recovered?

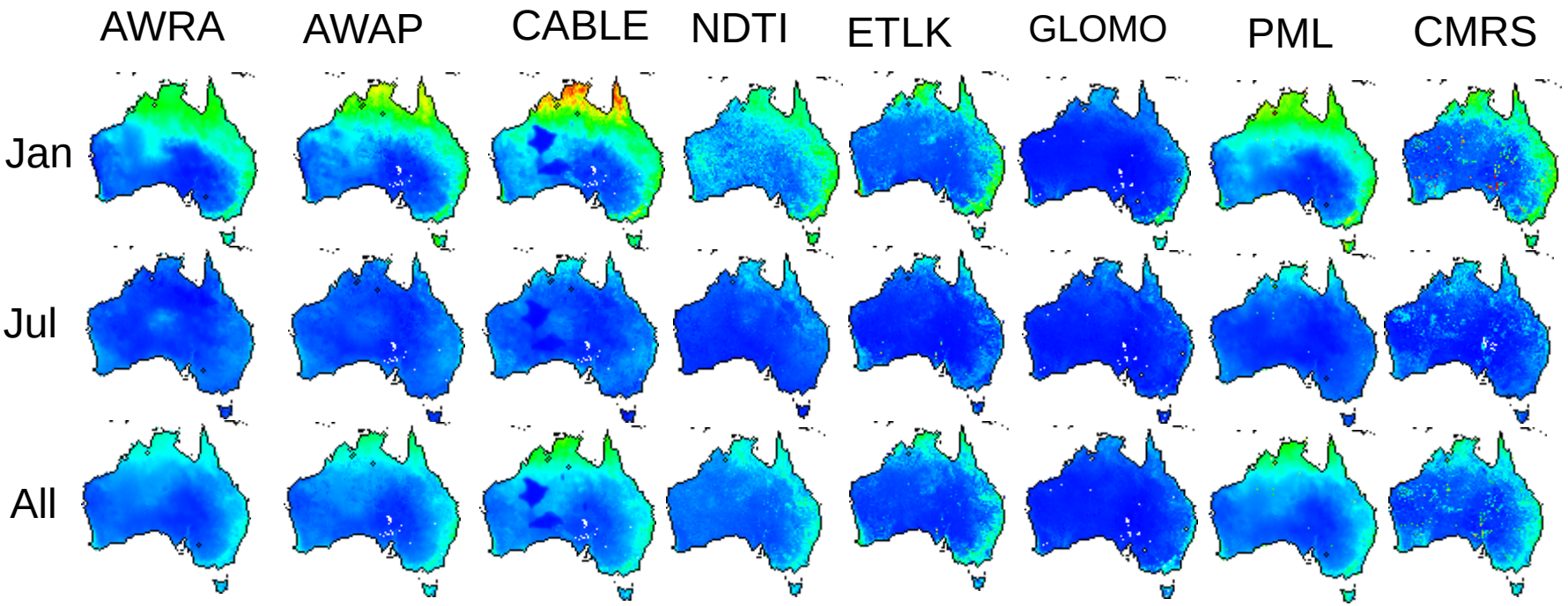
Example 2: WIRADA Actual ET Inter-comparison

- 12 different AET products to be compared
- Different
  - Formats
  - Frequency
  - Units
  - Resolution

- Shown: 8 day period for one region

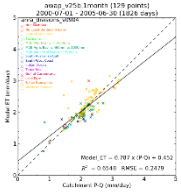


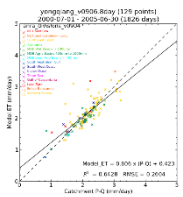
Convert to common format, representation and metadata…

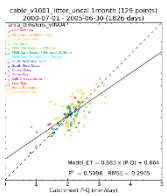


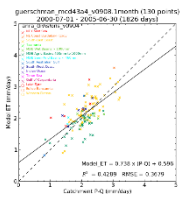
Makes it easier to perform analysis across each

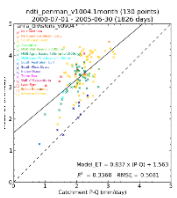
- ET = P - Q + ΔS and assume ΔS=0 averaged over multiple years

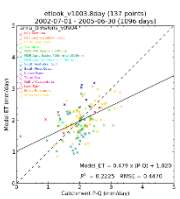

AWRA

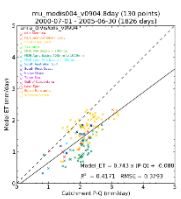

AWAP


PML


CABLE

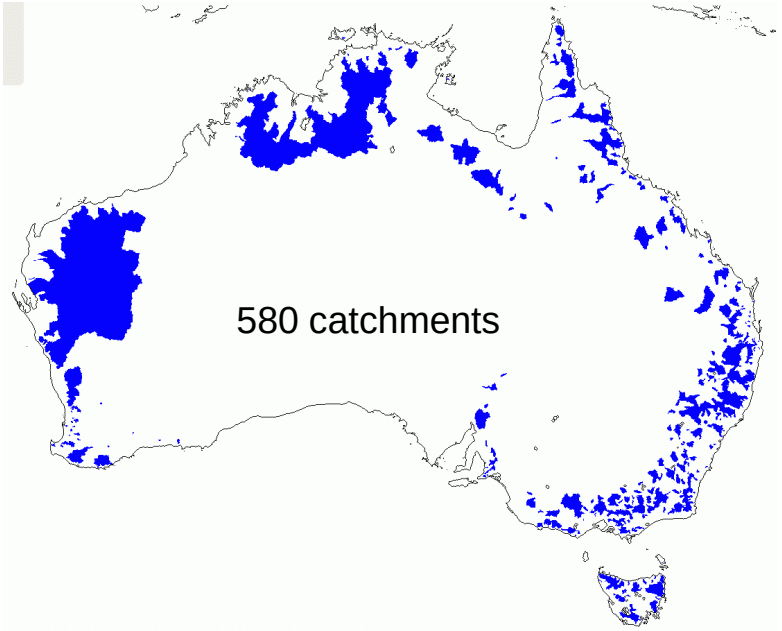

CMRS


NDTI


GLOMO

Better

Poorer



- Diverse data sets
- Well-defined interface
- 1 set of processing code
- Easy to ensure consistency of evaluation
  - And re-evaluation

NCRIS – National Collaborative Research Infrastructure Strategy

- TERN and IMOS both part of NCRIS, Auscover and SRS are the RS components respectively
- Mandate to collect and make available environmental observations in support of research on a sustained basis

- Gil et al (2016) Best practices for documenting and sharing data
- Identified 7 key elements:
  1. Data accessibility
  2. Data documentation
  3. Software accessibility
  4. Software documentation
  5. Provenance documentation
  6. Methods documentation
  7. Author identification

Issues

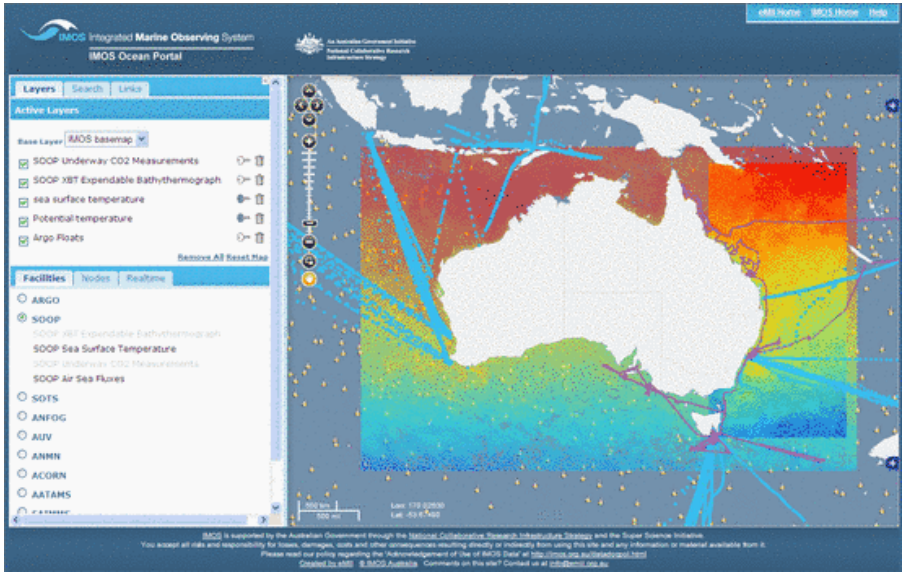
- At the start (2007) we only really appreciated 1 &2 (and 7)
- There were lots of (mostly incompatible) ways to do each of these things
- Few obvious choices
- Approach was one of “resource-limited trial and error” constrained by community practices (and data volumes)

NCRIS – TERN/Auscover and IMOS-SRS

- Starting point was multiple historical and contemporary data sets
- Either:
  - Received in Australia
  - Downloaded from overseas (physical media)
- With a range of:
  - Formats
  - Metadata
  - Custodians
  - Locations
  - Completeness
  - Processing level
  - We set out to introduce some consistency and consolidate these

IMOS

- Mandated for all facilities:
  - formats (netCDF with CF conventions)
  - metadata standard (ISO 19115 MCP)
- Uniform ISO metadata enables a searchable catalogue
- netCDF created a common data interface
- Enabled use of spatial data www protocols
- Facilitates WWW portal for ALL IMOS data



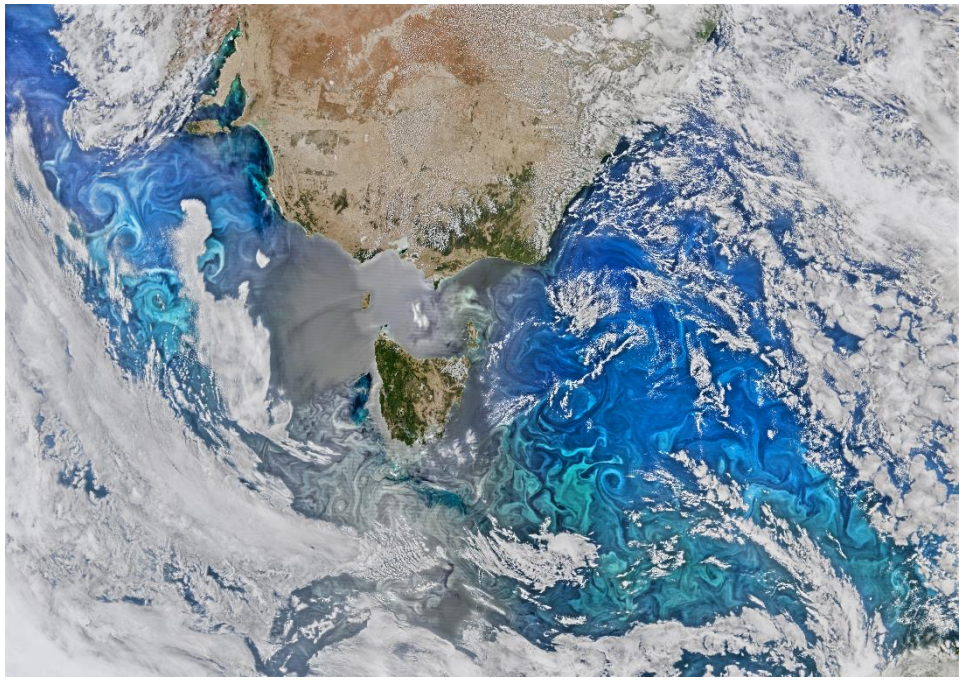
These choices also

- Annoyed everyone who didn't already use netCDF
- Forced people to think about metadata

BUT – All IMOS observations are now in a comprehensively self-describing portable collection. They can be discovered and interpreted. netCDF may not be the right choice of format, but translation to the appropriate format in the future is an automatable task. The collection (not just RS) is effectively future-proofed (and some people don't use it).
Similar (but not so comprehensive) process in TERN, particularly AusCover.

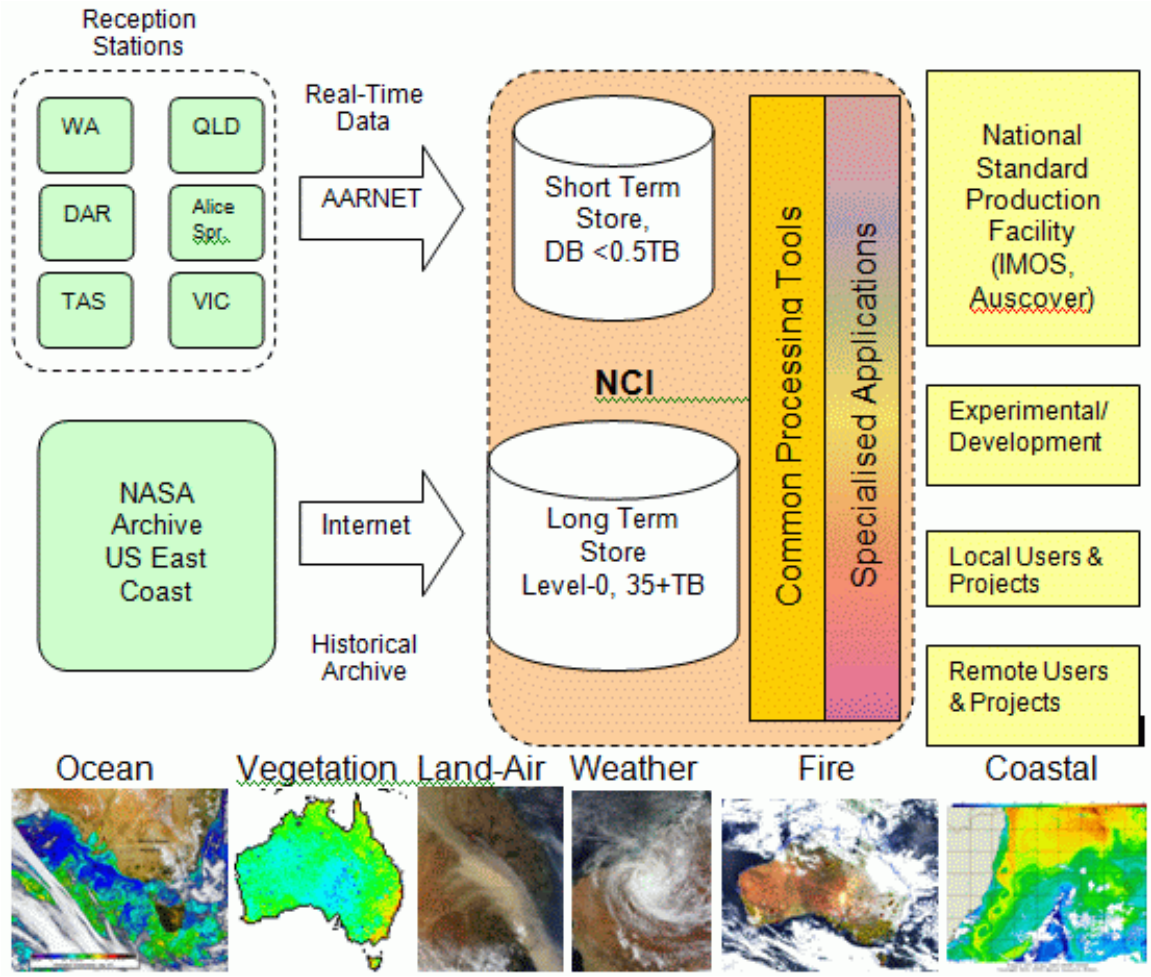
MODIS (Aqua & Terra)

- Key pair of sensors over period 2000-2017+
- High resolution, frequency, and spectral capability
- Land, sea, atmosphere
- 10’s of TB/year + products



- By 2007 there were something like 6 separate and incompatible archives in different agencies, institutions
- Independently curated, formatted etc
- Processed to different levels in different ways (excludes other uses)
- Not easy to access (behind firewalls)
- Not easy to re-use (hence multiple instances)

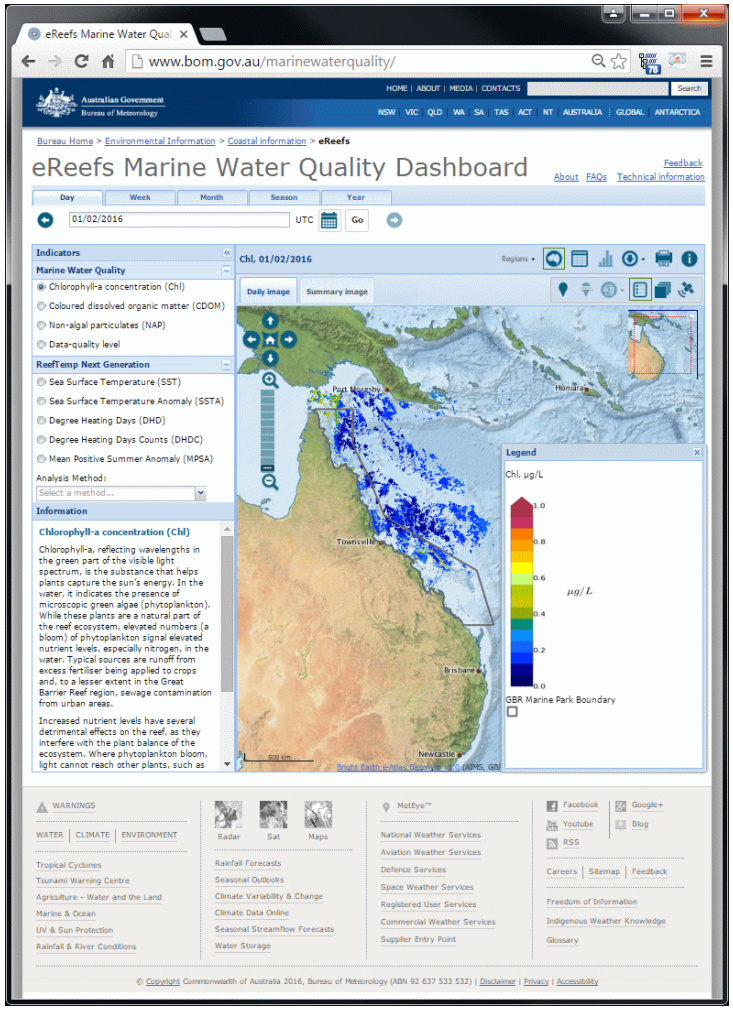
MODIS TERN/AusCover+IMOS/SRS+ NCI

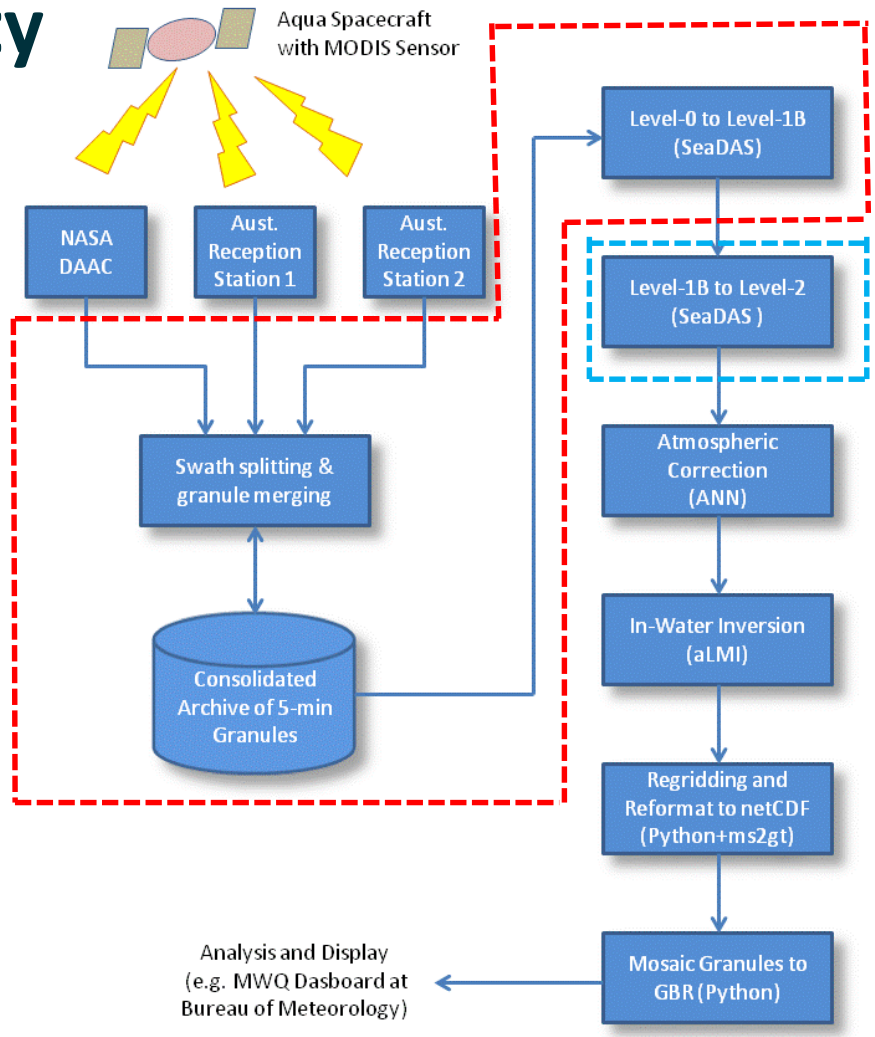


- Open platform (NCI)
- Large scale platform (NCI)
- Comprehensive (all base products)
- Consistent across life of mission and domains
- Both IMOS+TERN built their own products off it
- Eliminated several of the existing archives
- Served as a source for Australian sub-archives

- Demonstrated a pathway for forthcoming sensors
- Eg Copernicus Hub

eReefs – GBR Water Quality



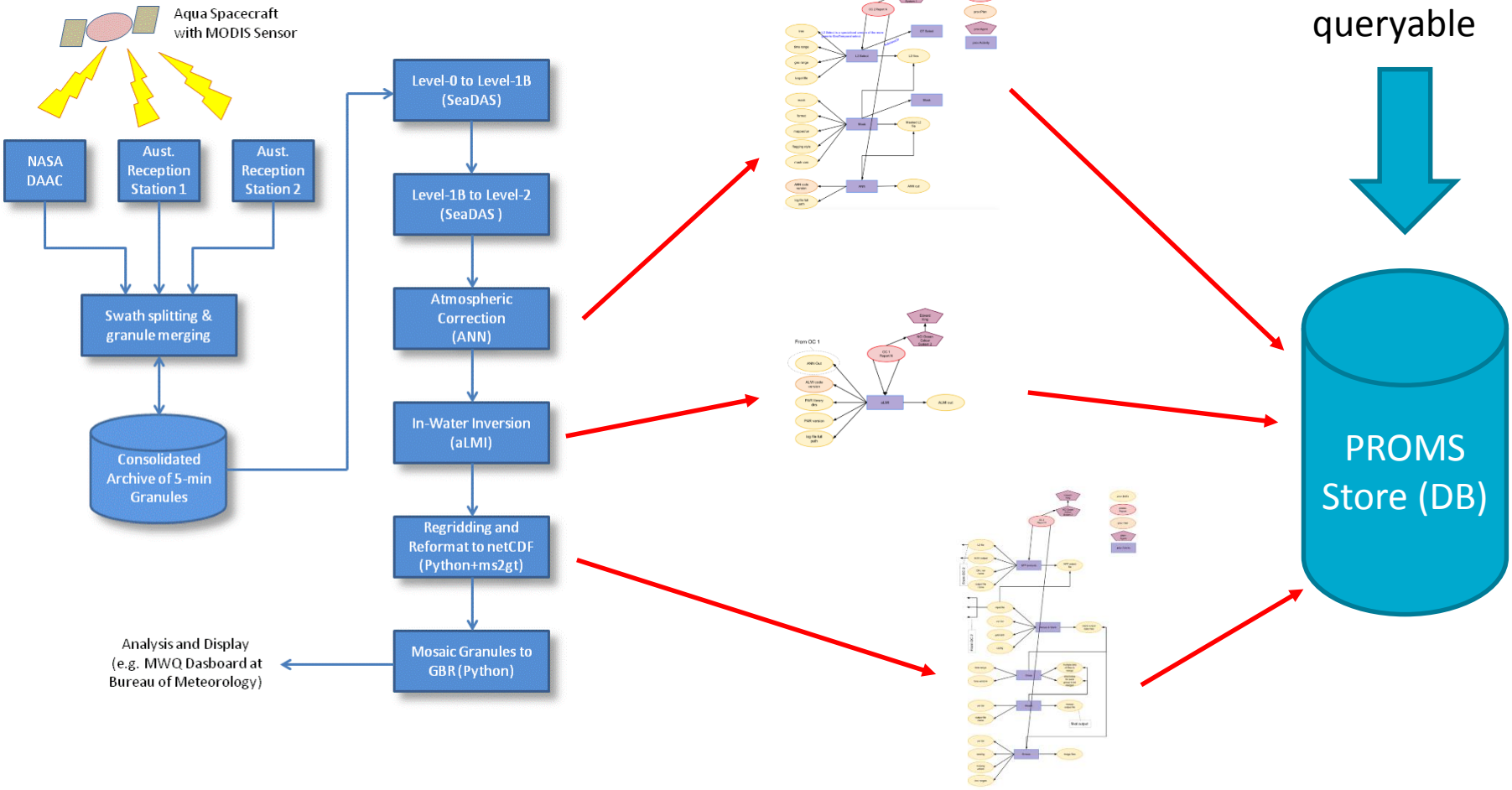


Process metadata - provenance

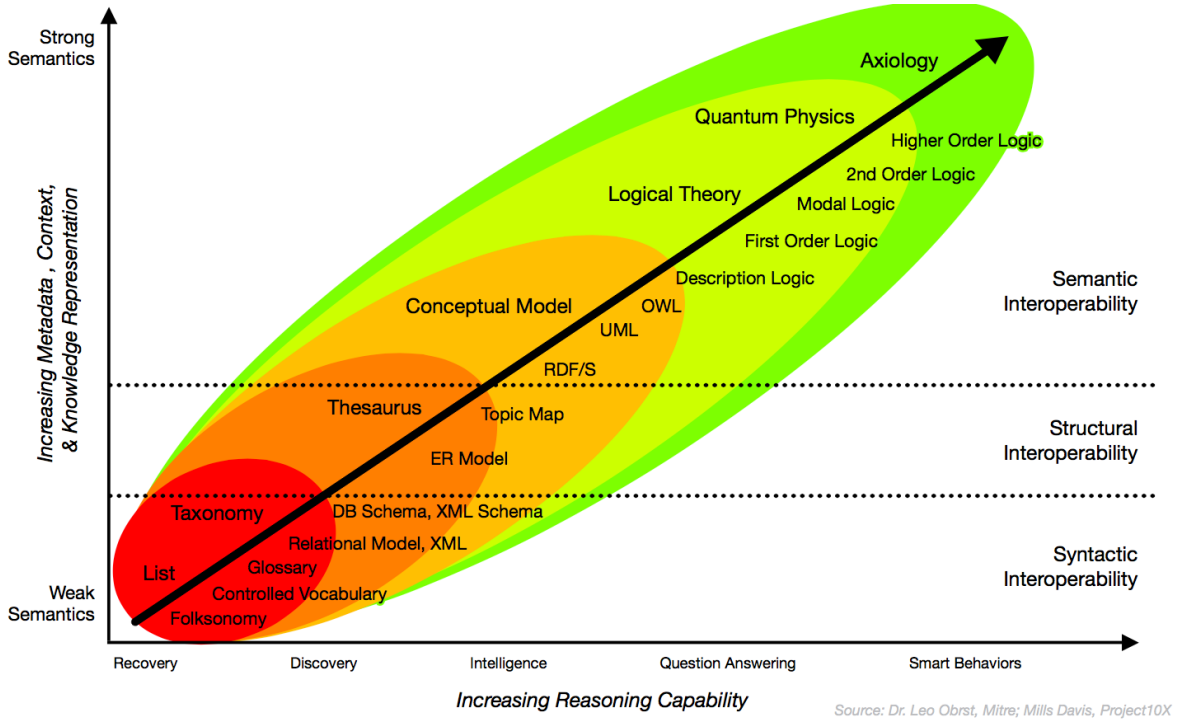
- How did my data get like this?
- Many possible approaches
  - Processing code adds metadata to the file
  - Code save metadata separately in another file (or DB)
- Challenges
  - Code may already exist – how to retrofit/wrap?
  - Code version?
  - What is the right level of granularity/detail?
  - What format should it be in?


```
{
  "l1bFile": "A20170204_0315.20170226224234.L1B_1KM.hdf",
  "l1bCalFiles": [
    "MYD02_REFLECTIVE_LUTS.V6.1.35.28_OC2.HDF",
    "MYD02_EMISSIVE_LUTS.V6.1.35.28_OC2.HDF",
    "MYD02_QA_LUTS.V6.1.35.28_OC2.HDF"
  ],
  "gringPointLongitude": [
    174.241910948841,
    147.823232999521,
    145.525897973034,
    167.90419663897
  ],
  "gringPointLatitude": [
    -34.8664379020068,
    -38.654226022052,
    -19.933939392599,
    -16.9030076452908
  ],
  "orbitNumber": "78490",
  "startDate": "2017-02-04",
  "startTime": "03:15:00.904953",
  "stopDate": "2017-02-04",
  "stopTime": "03:20:09.942920",
  "platform": "AQUA",
  "qaPercent": "0",
  "east": "174.237956369907",
  "west": "145.549182901782",
  "north": "-16.9740295046892",
  "south": "-38.6321574997786"
}
```

PROMS demonstrator



Infinity and beyond! The semantic web



Source: Dr. Leo Obrst, Mitre; Mills Davis, Project10X

2007, 2008 Copyright MILLS•DAVIS. All rights reserved

Back to reality…

Australia's Regional Copernicus Data Hub

All Sentinel Products For

Supporting

- esa
- European Commission
- European Environment Agency
- EUMETSAT

Consortium

- Queensland Government
- CSIRO
- Australian Government Geoscience Australia
- Government of Western Australia
- Landgate
- NSW Government
- Office of Environment & Heritage

Partners

- NCI National Computational Infrastructure
- aarnet Australia's Academic and Research Network

Collaborators

- Australian Government Bureau of Meteorology
- TERN
- RDS
- IMOS Integrated Marine Observing System
- crc·si

- 10 years after MODIS merging
- National approach
- Consistency within satellites (formats, metadata)
- One step further down the road… (currently 0.5 PB and growing)

Reflections, Risks and Priorities

- There is an awful lot happening
- The tools and practices are a moving target, and can get complicated fast
  - Risk of either adopting too soon, or never
  - Some are maturing (file formats, metadata stds, software revision control)
  - Maturing does not mean converging (eg GIS vs HDF/netCDF)
- One thing that can be done is to make tools easier to use
  - Libraries that hide the complexity of formats, metadata etc make it easier to write conforming code
  - Ubiquitous server infrastructure for catalogues, provenance
  - Tools that easily translate
  - Digital literacy amongst scientists (eg Software Carpentry)
- Consistent and stable infrastructure can make the transition from research into operations easier
- NCRIS facilities are “operational” research infrastructure
- Soft infrastructure is an enabler of trans-domain work, because it provides a common language by which translation can occur
- There are many choices to be made – what are the “no regrets” ones?

Thank you

Oceans and Atmosphere
Edward King
Oceans Group, Climate Science Centre

t +61 3 6232 5334
e edward.king@csiro.au
w www.csiro.au/lorem