Earth and Space Science Informatics
2014 Leptoukh Lecture

Bryan Lawrence, NCAS, STFC & University of Reading

Trends in Computing for Climate Research

Dr. Greg Leptoukh

Data management and analysis, large-scale computation and modeling, and hardware and software infrastructure profoundly affect the research capabilities of all AGU disciplines. The Earth and Space Science Informatics Focus Group established the **Leptoukh Lecture** in 2012 to recognize advances in these fields and their contributions to Earth and space science. Named in honor of the late Dr. Greg Leptoukh, a pioneer in satellite data quality and provenance, the Leptoukh Lecture aims to identify and support achievements in the computational and data sciences.

Credits!

University of Reading
NERC SCIENCE OF THE ENVIRONMENT
Science & Technology Facilities Council
Centre for Environmental Data Archival
Science and Technology Facilities Council
Natural Environment Research Council
National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Reporting the work of a cast of thousands . . .

Colleagues at CEDA and NCAS, especially Martin Juckes
CMIP colleagues, especially Karl Taylor and V. Balaji
ESGF colleagues, especially Dean Williams
ENES colleagues, especially Sylvie Joussame, Michael Lautenschlager, and Sebastien Denvil
ESDOC colleagues, especially Eric Guilyardi, Sophie Valcke, Mark Greenslade
CHARMe colleagues, especially Jon Blower, Phil Kershaw

. . . and that of a host of other projects and individuals!

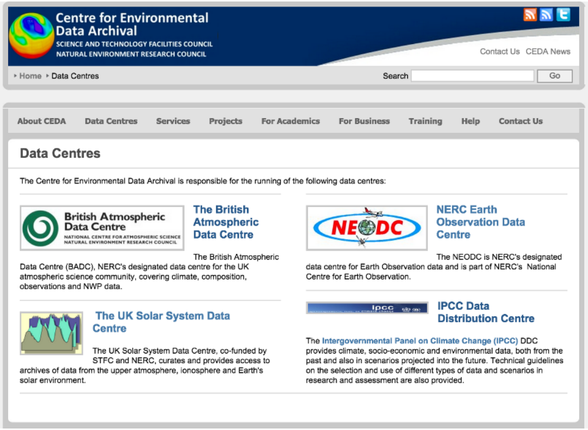
Trends in Computing for Climate Research

The grand challenges of climate science will stress our informatics infrastructure severely in the next decade. Our drive for ever greater simulation resolution/complexity/length/repetition, coupled with new remote and in-situ sensing platforms present us with problems in computation, data handling, and information management, to name but three. These problems are compounded by the background trends: Moore's Law is no longer doing us any favours: computing is getting harder to exploit as we have to bite the parallelism bullet, and Kryder's Law (if it ever existed) isn't going to help us store the data volumes we can see ahead. The variety of data, the rate it arrives, and the complexity of the tools we need and use all strain our ability to cope. The solutions, as ever, will revolve around more and better software, but "more" and "better" will require some attention.

Over Ambitious

In this talk we discuss how these issues have played out in the context of CMIP5, and might be expected to play out in CMIP6 and successors. Although the CMIPs will provide the thread, we will digress into modelling per se, regional climate modelling (CORDEX), observations from space (Obs4MIPs and friends), climate services (as they might play out in Europe), and the dependency of progress on how we manage people in our institutions. It will be seen that most of the issues we discuss apply to the wider environmental sciences, if not science in general. They all have implications for the need for both sustained infrastructure and ongoing research into environmental informatics.

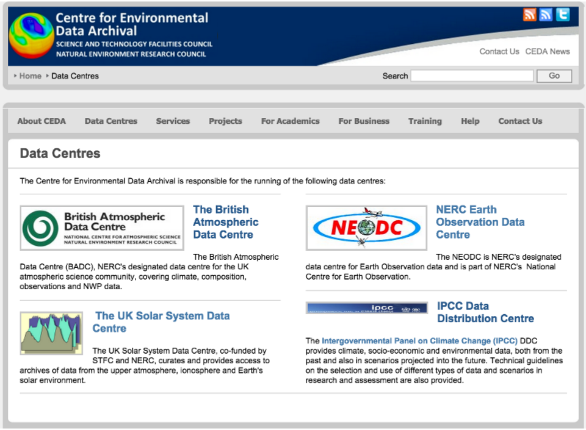
CEDA



Four data centres: http://ceda.ac.uk

Providing Curation (of the archive)

CEDA



Four data centres: http://ceda.ac.uk

Providing Curation (of the archive)

. . . and

Facilitation (of Science)

Using the Archive and the Infrastructure to provide User Compute Services

Centre for Environmental Monitoring from Space
CEMS

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

http://cems.ac.uk

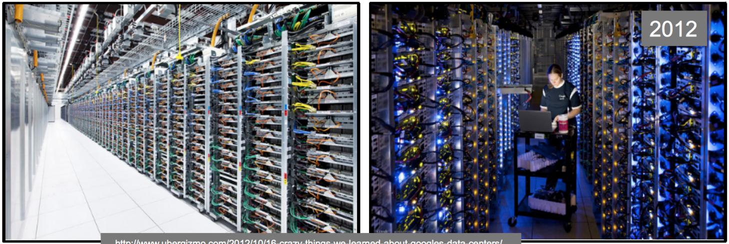
JASMIN
http://jasmin.ac.uk
(User storage on Group Work Spaces: GWS)

CEDA Evolution




Eerily similar to Google

http://infolab.stanford.edu/pub/voy/museum/pictures/display/GoogleBG.jpg Wikipedia



http://www.ubergizmo.com/2012/10/16-crazy-things-we-learned-about-googles-data-centers/,
http://blogs.wsj.com/digits/2012/10/17/google-servers-photos/

Not so subliminal message:

As we move to exascale storage, not everyone will be able to scale from a few machines to one (or more) massive machine rooms.

Not so subliminal message:

As we move to exascale storage, not everyone will be able to scale from a few machines to one (or more) massive machine rooms.

Actual subliminal message:

As well as hardware, one needs an awful lot of software to manage and exploit data at scale. Much of it will be bespoke!

Outline

~~Metadata~~ Information is our knight in the battle against the (big) data dragon

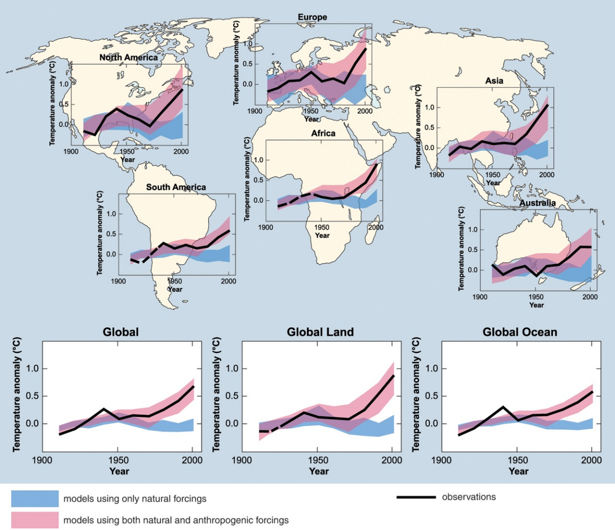
1. Motivation comes from the science challenges, and some of the programmes that result.
2. The background computing trends are faster computing, more parallelism, and increasing storage issues (including relative costs). All these things are leading to a "Tyranny of scale". It's too easy to say million, or petabyte.
3. Dealing with things at scale requires information, aka, metadata, and lots of different types of metadata, all of which enable different parts of the scale problem.
4. Quality is one important sort of metadata ...
5. Compute: From download to community-based analysis computing.
6. The wrap up puts this back in the context of MIPS and real people.

From the Large

Fig 2.5
AR4
Synthesis
Report

Simulations
and
Observations

Big Picture



To the Small

How will climate change affect the global distribution of malaria?

July 2007 Tewkesbury flood: 3B€ loss!

Can we predict risk into the future?

How will climate change affect the incidence of road and rail closures due to landslides?

What would be the impact of leakage from an oil and gas well in UK waters on the national economy, coastal and marine biodiversity and the well-being of the population affected?

To the Small

How will climate change affect the global distribution of malaria?

July 2007 Tewkesbury flood: 3B€ loss!

Can we predict risk into the future?

What would be the impact of leakage from an oil and gas well in UK waters on the national economy, coastal and marine biodiversity and the well-being of the population affected?

How will climate change affect the incidence of road and rail closures due to landslides?

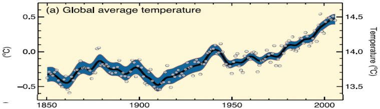
Field Observations

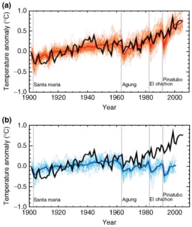
Small Scale Models

User of Big Model Data

The first and second problems of climate change science

Pre AR4: *Are human activities changing climate?*





Post AR4: *What does it mean?*

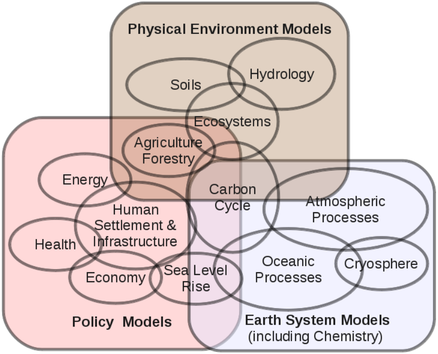
- *What is the signal of anthropogenic climate change on the regional and local scales that really matter to individuals, economies and societies?*
- *What does/will climate change look like where I live? Temperature not the be-all and end-all of answering that question!*

*A far more difficult **grand challenge research** and **development** problem!*

*Greatly **enhanced** national and international **collaboration** and strategy **essential!***

Slide modified from Rowan Sutton's version.

Communities

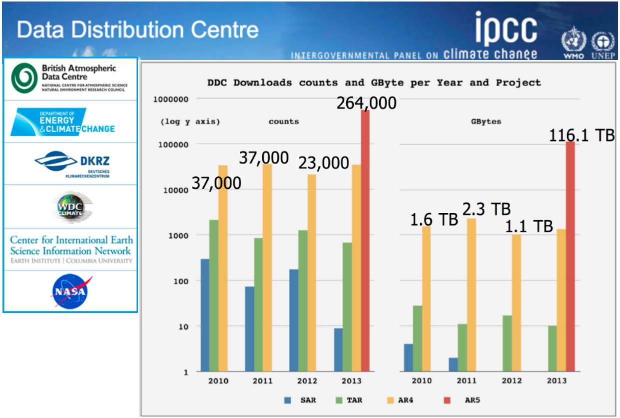


Many interacting communities, each with their own software, compute environments etc.
Often interacting through data.

Figure adapted from Moss et al, 2010

Communities interacting through data: IPCC DDC

GCM data download at DKRZ



The 2013 IPCC-DDC downloads of 161 TB compares with 650 TB delivered via ESGF and 440 TB delivered to DKRZ WDCC users.

Almost certainly different user communities via each method!

(The IPCC-DDC has a snapshot of the CMIP5 archive as it at the time of the AR5 paper submission cut-off … a "reference archive".)

A one slide guide to CMIP5 from a data perspective

Fifth Climate Model Intercomparison Project (CMIP5)

World Climate Research Programme

WCRP- WGCM

Production involved all the major climate modelling *centres*

Original Timing:

- Expect o(2) PB of requested output from 20+ modelling centres finished early 2010!

Actual Timing?

- Years late.

(23/05/13):

101 experiments

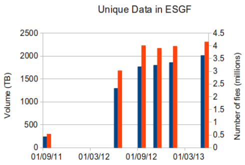
61 model variants

590,000 datasets!

4.5 million files

2 PB in global archive

Unknown PB locally!



Blue: Volume; Red: Files

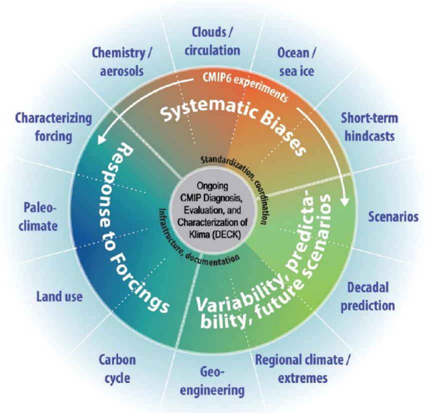
(NB: replicas and versions!)

Data Delivered by:

PCMDI-led, community developed (GO-ESSP) s/w infrastructure for data delivery:

Earth System Grid Federation (ESGF)

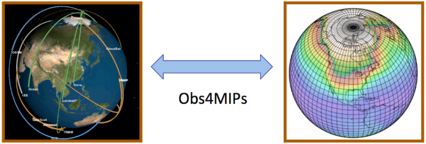
CMIP6



(Meehl et al 2014, EOS)

- Final experiment definition underway (first quarter 2015)
- Data "request" being discussed (2nd quarter 2015)
- Metadata "request" being discussed (3rd quarter 2015)
- A certain amount of "weariness" in the modelling community. It's currently hard to meet the data and metadata "requirements".
- Expecting data to start flowing in 2016.

Obs4MIPs and Friends



Waliser et al, 2012http://goo.gl/2B1YvV

NASA Project:

- Identify key observational datasets relevant to the model evaluation process.
- Establish metadata requirements, document quality.
  - **"Well documented with traceability to track product version changes"**
- Enable the production of quality controlled, model format compliant datasets.
- Disseminate via ESGF.

ESA following the same procedure with the Climate Change Initiative (CCI) project (but with their own portal into ESGF planned).

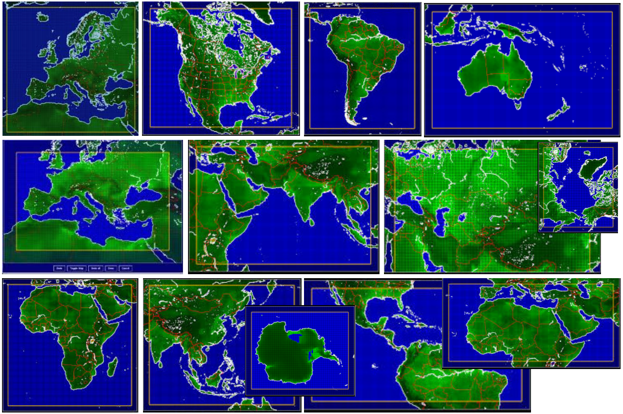
(See also ANA4MIPS, CFMIP-OBS, etc)

(Analysis, cloud feedback observations, etc)

CORDEX: Coordinated Regional Downscaling Experiment

Objectives include:

- To **quality-control datasets** . . . recent historical past and 21st century projections, covering the majority of populated land regions on the globe.
- To build a common set of RCM domains for dynamical downscaling and define a **standard set of variables, frequency and format for output and archival** at a number of CORDEX data centers.
- To develop of Regional Analysis and Evaluation Teams . . .



- To support and inform the climate impact assessment and adaptation groups . . .

Common Trend?

OBS4MIPS and CORDEX “joining” the ESGF community in order to disseminate their data, but also to (respectively)

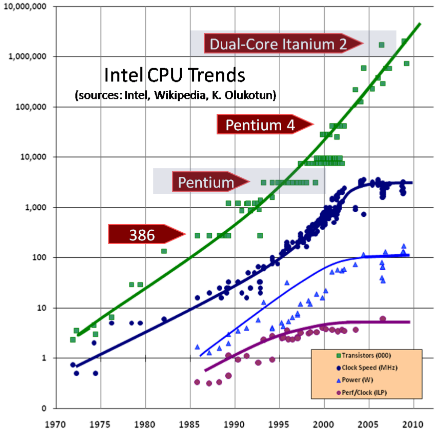
- "ensure data is well documented with traceability to track product version changes" and enable "production of quality controlled format compliant" datasets.
- (carry out) "quality control" of datasets ... define standards for output and archival ... to support and inform impact assessment and adaptation groups ...

The CMIP community has mainly concentrated on the output, where we see (for CMIP5):

- Output requirements: " In order for the output to be easily accessible and useful to researchers and before it can be accepted into the CMIP5 archive, the output must be written conforming with the requirements specified in two documents": the CMIP5 Data Reference Syntax (DRS) and Controlled Vocabulary, and the CMIP5 Model Output Format and Metadata Requirement. (http://cmip-pcmdi.llnl.gov/cmip5/output_req.html)

... but WGCM did attempt to mandate model metadata (of which more later), and some of the archive community at least attempted to address quality control wrt format compliance, and CMIP6 will attempt to look at model quality directly.

Moore's Law



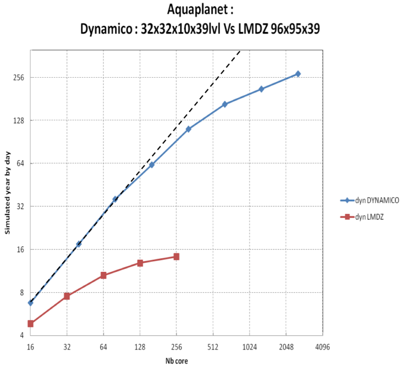
(Herb Sutter, 2004, updated in 2009.)

- Clock speeds not getting any faster (and haven't been for quite a while).
- Transistor density still going up - hence advent of GPUs and accelerators.
- Memory density and bandwidth not keeping up — means it's hard to exploit GPU and accelerators (and going to get harder — fundamental power limits).
- We're kind of used to the problems this means for our simulation codes - massive parallelisation, from MPI to OpenMP to OpenACC . . .
- . . . problems moving data to exploit the parallelisation etc.

Making progress with parallelisation (1)

Much going on with improving simulation codes

both with coarse parallelisation, for example:

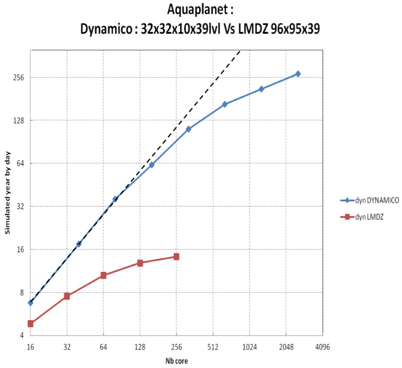


T.Dubos, S.Dubesh, Yann Meurdesoif(LSCE-IPSL)
Results presented at IS-ENES2 workshop, March 2014

Making progress with parallelisation (1)

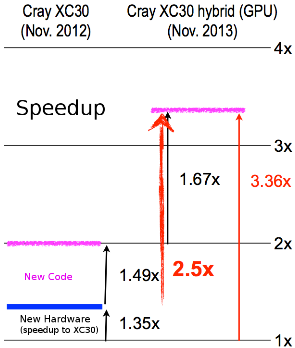
Much going on with improving simulation codes

both with coarse parallelisation, for example:



T.Dubos, S.Dubesh, Yann Meurdesoif(LSCE-IPSL)
Results presented at IS-ENES2 workshop, March 2014

and porting to GPUs, for example:



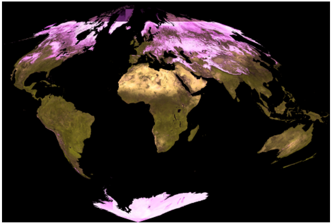
This work also showed the energy to solution falling by an overall factor of nearly 7 (with a factor of 4 from the GPUs!)

T. Schultess (ETH-Zurich) showing results of 3H Meteo Swiss forecast using the COSMO-2 rewrite, presented at IS-ENES2 workshop, March 2014.

Making progress with parallelisation (2)

Rather less going on with analysis parallelisation

At least much of it is embarrassingly parallel, and we can get results from throwing hardware at the problem, for example:



QA4ECV: "Re-processed MODIS Prior in 3 days (on JASMIN-Lotus). 81 times faster than on 8-core blade".

Boersma and Muller (2014)

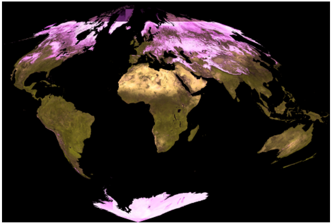
Presentation at http://goo.gl/osEQ6M

From half a year to 3 days!

Making progress with parallelisation (2)

Rather less going on with analysis parallelisation

At least much of it is embarrassingly parallel, and we can get results from throwing hardware at the problem, for example:



QA4ECV: "Re-processed MODIS Prior in 3 days (on JASMIN-Lotus). 81 times faster than on 8-core blade".

Boersma and Muller (2014)

Presentation at http://goo.gl/osEQ6M

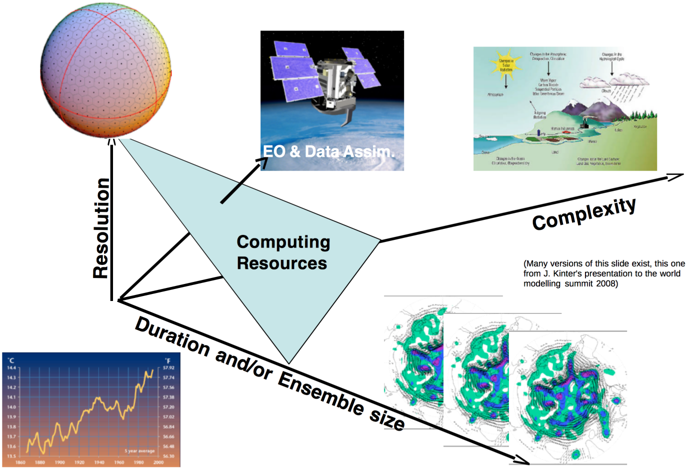
From half a year to 3 days!

But we need to work on the software tools (going beyond exploiting queuing or bespoke MPI).
Here for example are some Python choices :

- Standard: Multiprocessing, PyMPI etc
- The way of the future: ipython-notebook
- Generic Workflow and Map Reduce: Jug
- Extending Numpy:
  - Using more cores: Numexpr
  - Using more processors: DistArray (Enthought), Blaze (Continuum Analytics)
- Atmospheric Science aware:
  - PyReshaper, PyAverager (Mickelson, NCAR)
  - cf-python (Hassel, NCAS) (Exploiting LAMA, extending to MPI under the hood soon.)

(Original list courtesy of Matt Jones, UoR)

Give me more computing? Global Climate Modelling

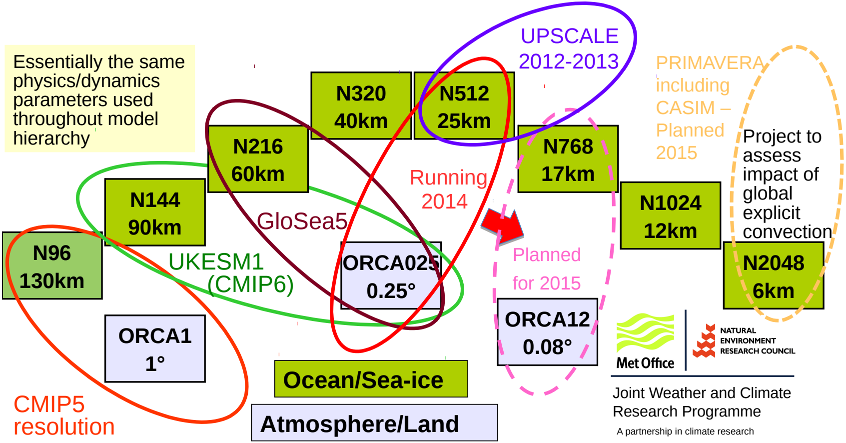


(Many versions of this slide exist, this one from J. Kinter's presentation to the world modelling summit 2008)

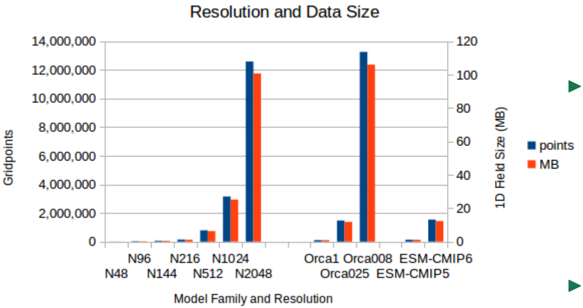
JWCRP Climate Modelling

Earth System Modelling
PI C. Jones (NCAS at the Met Office)

High Resolution Climate Modelling
Joint PIs: P-L. Vidale (NCAS), M. Roberts (Met Office)



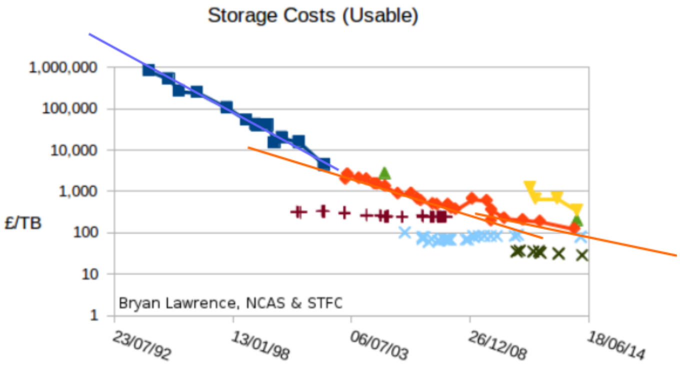
Resolution and Data!



Consequences:

- 1 MB output per 2D field with 10 ensemble members and 100 output variables and 100 levels for $100 * 12 \approx 1000$ time steps $= 10^8$ MB $=$ 100 TB!
- If the UK runs the same number of years for CMIP6 as CMIP5, looks like about o(10) times more data for CMIP6, but could be much worse — more "physics" experiments, means more high resolution experiments, and likely to use bigger ensembles.
- My own experience? Running high resolution gravity wave experiments, 2 years of N512L180 writing data hourly $\approx$ 100 TB. Now!

Kryder's Law - Two decades of storage costs at BADC



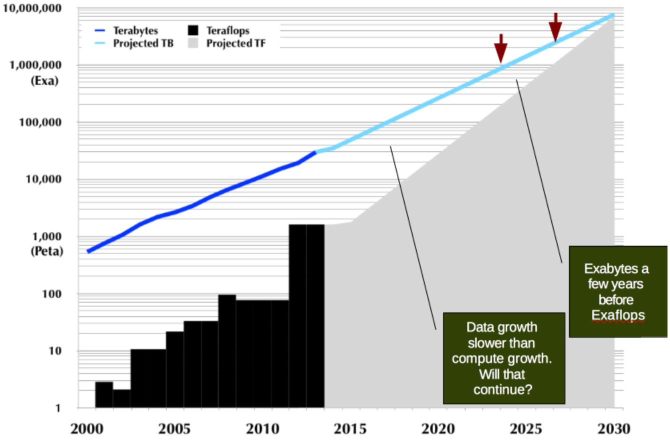
Solid objects: colours are different generations of disk. Crosses: different generations of tape.

Kryder's Law definitely slowing down!
Every new generation of disk storage slightly shallower downward gradient.

Plenty of mileage still in tape though!

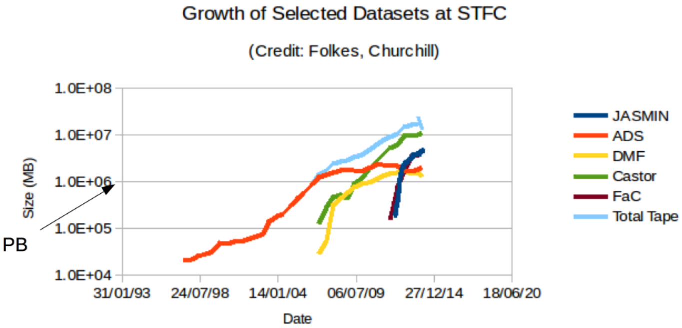
Institutional - NCAR

Storage, and power for storage, will dominate NCAR's compute budget within a few years! (Rich Loft, 2014).



NCAR Compute and Data (courtesy Gary Strand)

Institutional - STFC and CEDA

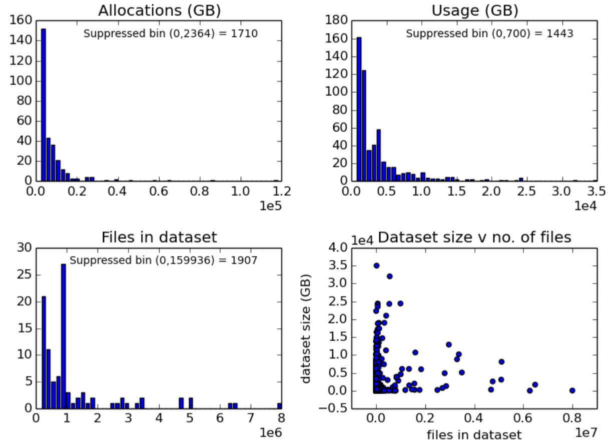


JASMIN storage is all on spinning disk, all other curves are tape!

Predictions for JASMIN in 2020? 30 — 85 PB of unique data[1]!
But we think we will only get 30 PB disk in the physical space available[2]!

([1] Not including CMIP6, which might be anything from 30 PB up. [2] Unless we can throw out the CERC Tier1 centre with whom we share!)

It's too easy to say "petabyte" !



Snapshot data 01/12/2014 via Sam Pepler.

CEDA Archive Snapshot, Complexity and Volume!

- 3.0 PB of allocated archive, 2.3 PB used in 2,176 "filesets" totalling 152M files (of which our subset of CMIP5 is 1.2 PB in 1,174 filesets totalling 3.2M files).
- 1 copy on disk, at least one on tape near line, and one offsite!
- Long tail in both dataset size and number of files.
- Volume and number of files not correlated, although the high volume datasets tend not to have the most files.

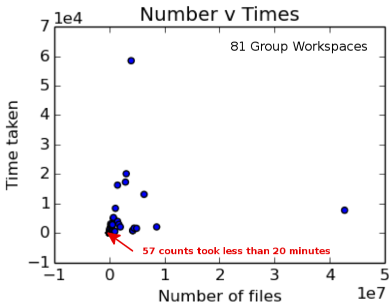
And too easy to say "thousands" and "millions" !

Humans struggle to do things with thousands of tasks!

- A person working full time for a year has about 1500 hours per year to "get things done" (or 90,000 minutes).
- Maximum of 1,500 things can be done per year if they take any intellectual interaction at all (and probably a lot less). Even if the interaction is relatively mechanical, we're still talking o(10,000) things a year (and a job no-one I know would want to do).
- If you have thousands (maybe only hundreds) of things which differ in any significant way, you have to automate handling them!

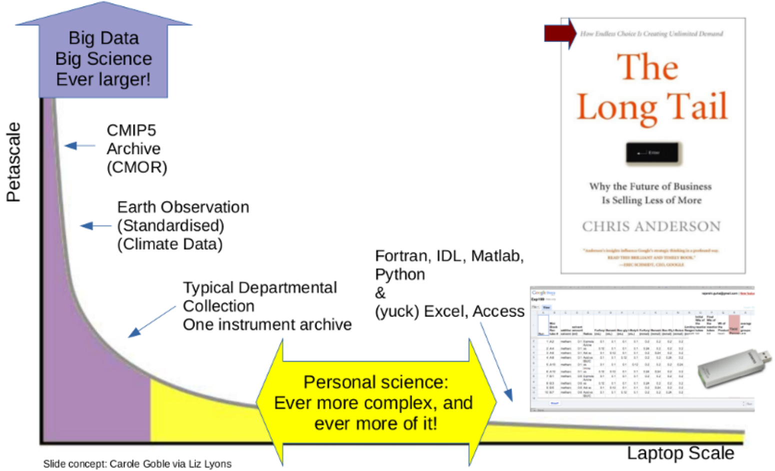
(In the UK 220 working days a year is about standard. Let's remove about 20 days for courses, staff meetings etc ... so that leaves about 200 days or, for a working day of 7.5 hours, a working year of about 1500 hours.)

But even automation takes time! When we get to millions, even computers take time . . . Consider the simple task of counting files in our "group workspaces"?



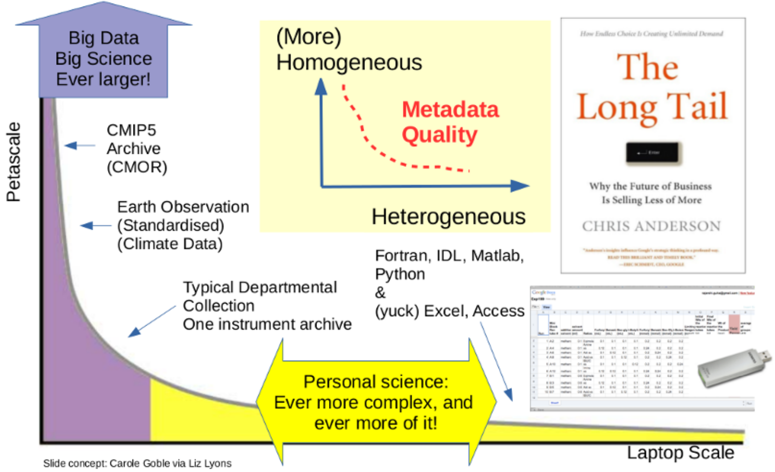
. . . so just asking the question "how many files are there in our group workspaces?" required parallelisation. Deduplication? Asking what is "in" them requires metadata (to avoid humans) and parallelisation (to get it done).

The Long Tail Reality

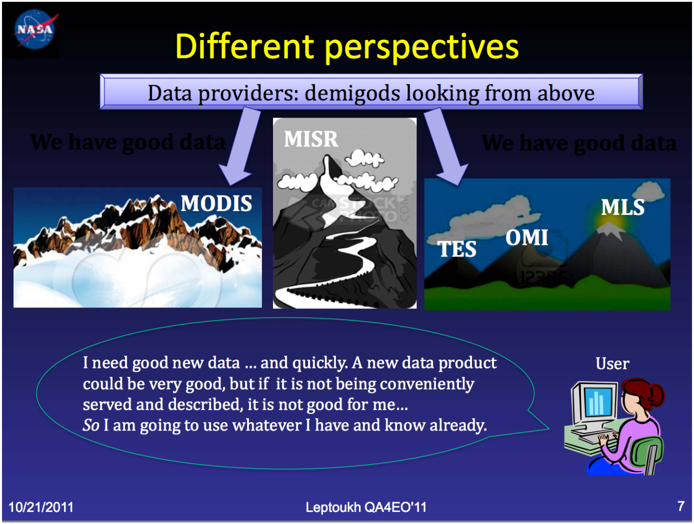


Slide concept: Carole Goble via Liz Lyons

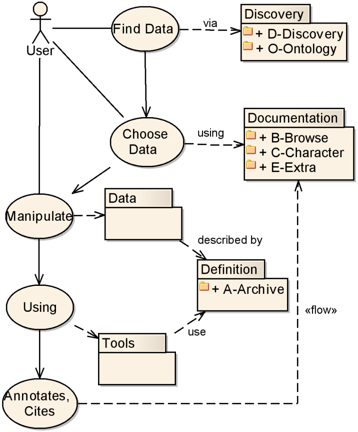
The Long Tail Reality



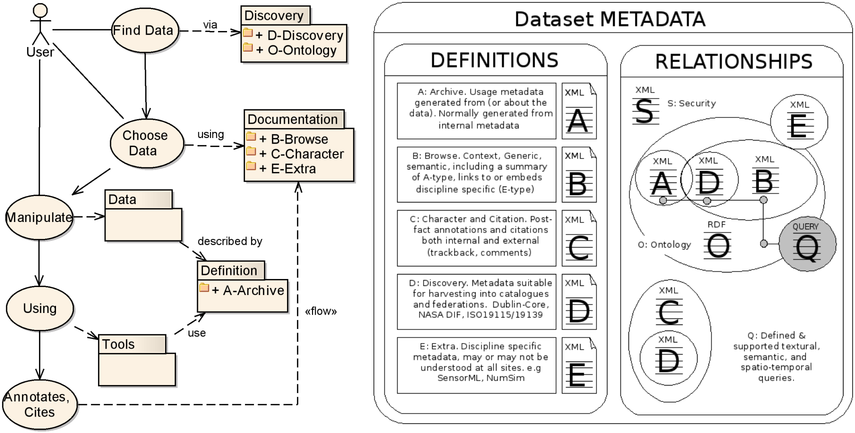
I know what I know



Information and Workflow



Information and Workflow



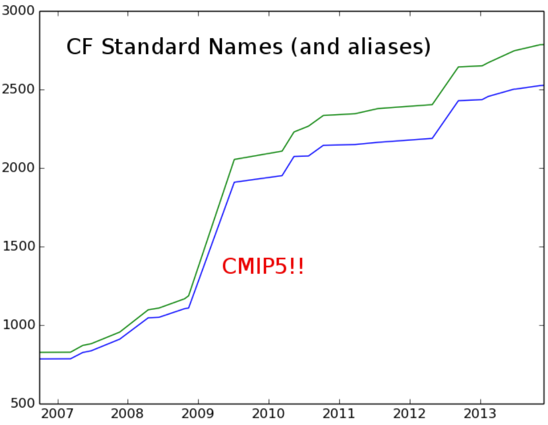
Lawrence, Lowry, Miller, Snaith and Woolf: Information in environmental data grids, Phil. Trans. R. Soc. A (2009) doi:10.1098/rsta.2008.0237

My A,B,C,D ...

Some examples:

- A for Archive: CF NetCDF metadata. Provides the information necessary to manipulate data. NCAS is directly supporting CF, Jonathan Gregory, David Hassel and Alison Pamment ...
- B for Browse: NCAS has developed "MOLES": Metadata Objects for Linking Environmental Sciences.
- C for Character (or Citation or Commentary): We are working on improving what can be done here - the CHARMe project.
- D for Discovery. We continue to provide ISO discovery records into national and international catalogues, but . . . Google?

A is for Archive: CF-NetCDF



The perils of scale!

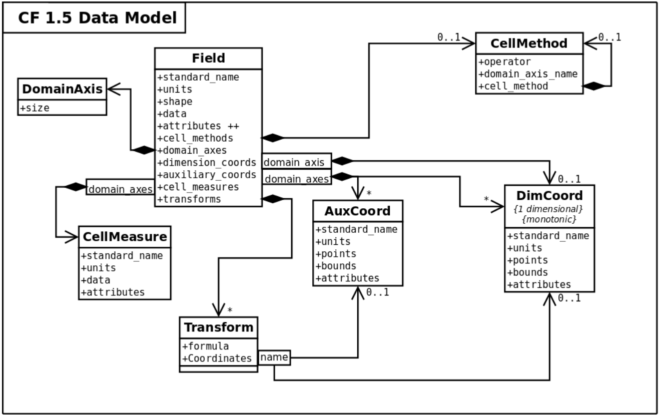
Peaking at 1000 standard names in about a year?

That's about as good as it gets . . . 1500 hours a year? 1.5 hours per standard name? Only happened because a lot of people engaged and we did them in big bunches (atmospheric chemistry, ocean biogeochemistry, and ice sheet dynamics!)

Current effort is on air quality and better support for satellite based remote sensing.

Namecheck:Alison Pamment and Jonathan Gregory!

CF Data Model



Still a work in progress
http://cf-trac.llnl.gov/trac/wiki/DataModel1.5Draft
(David Hassell, Jonathan Gregory, Bryan Lawrence, NCAS; Mark Hedley, UKMO)

Exploiting a data model

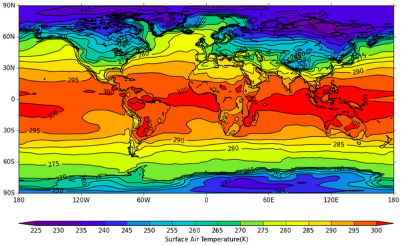
```
>>> f = cf.read('file.nc')[0]
>>> type(f)
<class 'cf.field.Field'>
>>> f
<CF Field: air_temperature(latitude(4), longitude(5)) K>
>>> print f
air_temperature field summary
-----------------------------
Data            : air_temperature(latitude(4), longitude(5)) K
Cell methods    : time: mean
Dimensions      : latitude(4) = [-2.5, ..., 5.0] degrees_north
                : longitude(5) = [0.0, ..., 15.0] degrees_east
                : time(1) = [2000-01-16 00:00:00] 360_day calendar
                : height(1) = [2.0] m
```

Exploiting a data model

```
>>> f = cf.read('file.nc')[0]
>>> type(f)
<class 'cf.field.Field'>
>>> f
<CF Field: air_temperature(latitude(4), longitude(5)) K>
>>> print f
air_temperature field summary
-----------------------------
Data
Cell methods
Dimensions
```

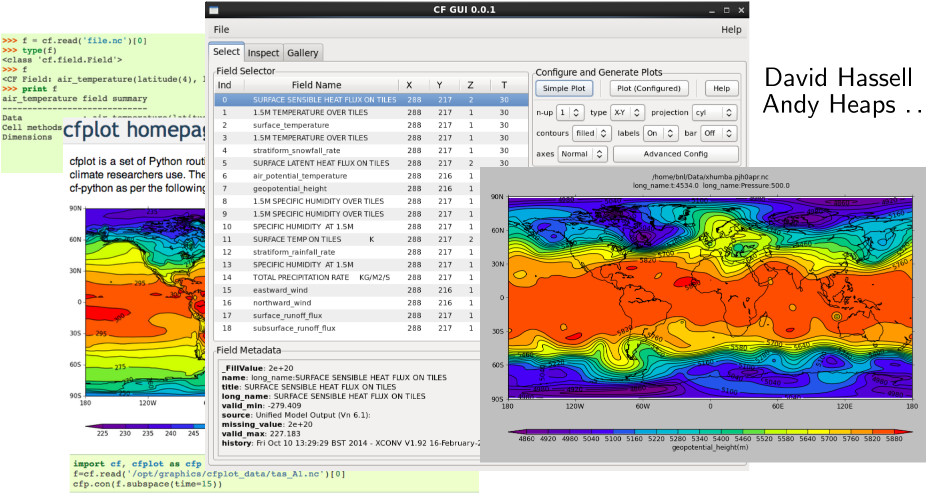
cfplot homepage

cfplot is a set of Python routines for making the common contour and vector plots that climate researchers use. The data to make a contour plot can be passed to cfplot using cf-python as per the following example.




```
import cf, cfplot as cfp
f=cf.read('/opt/graphics/cfplot_data/tas_A1.nc')[0]
cfp.con(f.subspace(time=15))
```

Exploiting a data model



David Hassell
Andy Heaps . . .

cf-python, cf-plot, and cf-gui — all built on the cf data model!

A is for Archive: but it's not only what's IN the files

DRS: Directory Reference Syntax, Taylor et al (2012)

http://cmip-pcmdi.llnl.gov/cmip5/docs/cmip5_data_reference_syntax.pdf

Why? Recall CMIP5 = 4.5 million files!

Use Cases

1. Provides search facets and pointers to collections of files.
2. Support identifying what should be **replicated**.
3. Consistent logical layout.
4. Download scripts can be modified according to logical patterns which are predictable.

A is for Archive: but it's not only what's IN the files

DRS: Directory Reference Syntax, Taylor et al (2012)

http://cmip-pcmdi.llnl.gov/cmip5/docs/cmip5_data_reference_syntax.pdf

Why? Recall CMIP5 = 4.5 million files!

Use Cases

1. Provides search facets and pointers to collections of files.
2. Support identifying what should be **replicated**.
3. Consistent logical layout.
4. Download scripts can be modified according to logical patterns which are predictable.

Atomic dataset definition: a subset of the output saved from a single model run which is uniquely characterized by a single set of facets.

Publication-level dataset definition: The collection of atomic datasets which share a single combination of all DRS component values . . . an intersection of several atomic datasets.

Filename convention: to support key DRS concepts in the file as well, to aid downloading and data management.

Controlled vocabularies (for the DRS terms).

A is for Archive: but it's not only what's IN the files

DRS: Directory Reference Syntax, Taylor et al (2012)

http://cmip-pcmdi.llnl.gov/cmip5/docs/cmip5_data_reference_syntax.pdf

Why? Recall CMIP5 = 4.5 million files!

Use Cases

1. Provides search facets and pointers to collections of files.
2. Support identifying what should be **replicated**.
3. Consistent logical layout.
4. Download scripts can be modified according to logical patterns which are predictable.

Atomic dataset definition: a subset of the output saved from a single model run which is uniquely characterized by a single set of facets.

Publication-level dataset definition: The collection of atomic datasets which share a single combination of all DRS component values . . . an intersection of several atomic datasets.

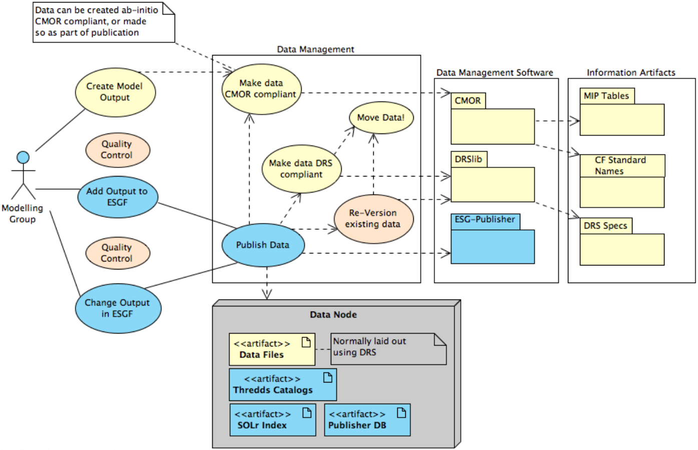
Filename convention: to support key DRS concepts in the file as well, to aid downloading and data management.

Controlled vocabularies (for the DRS terms).

CMIP5: activity, product, institute, model, experiment, data sampling frequency, modeling realm, variable name, MIP table, ensemble member, version number, subset, extended

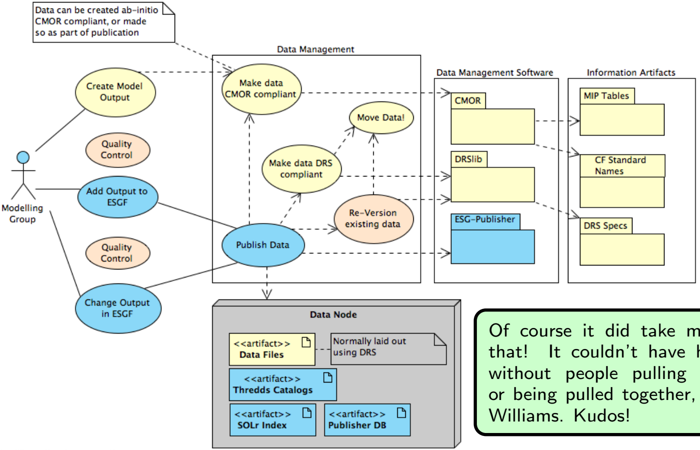
Most of the ESGF depends on A metadata!

Arguably most the success of ESGF is the metadata, the esg-publisher, plus a few simple interfaces:



Most of the ESGF depends on A metadata!

Arguably most the success of ESGF is the metadata, the esg-publisher, plus a few simple interfaces:



Of course it did take more than that! It couldn't have happened without people pulling together, or being pulled together, by Dean Williams. Kudos!

Metadata Objects for Linking Environmental Sciences

How do we find the “right” dataset?
Navigate to the vicinity (teleport) and browse (orienteer)?
Typically though, the latter seems always to get to “You are in a maze of twisty little passages, all alike”.

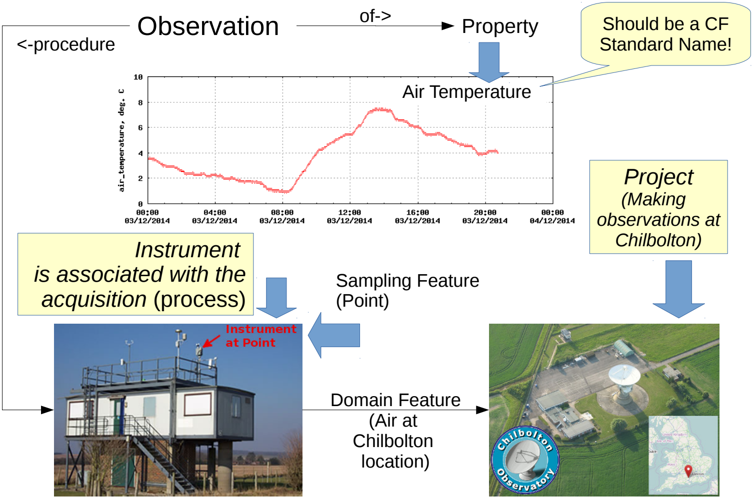
Metadata Objects for Linking Environmental Sciences

How do we find the “right” dataset?
Navigate to the vicinity (teleport) and browse (orienteer)?
Typically though, the latter seems always to get to “You are in a maze of twisty little passages, all alike”.

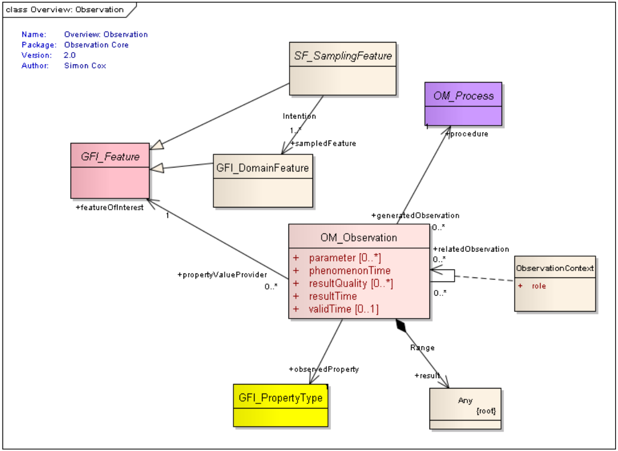
MOLES: Metadata Objects for Linking Environmental Sciences

- Provides a logical structure for a catalogue describing key aspects of datasets, and making navigable links between datasets which share those aspects.

Observation Concepts



MOLES: Observations and Measurements under the hood

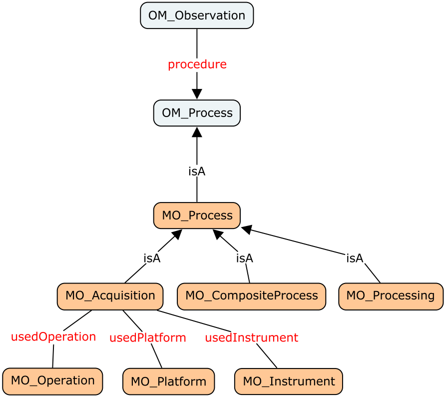


ISO19156

An observation is an action whose result is an estimate of the value of some property of the feature-of-interest, obtained using a specified procedure.

(In practice, an “observation” could be a measurement or a simulation!)

Moles Acquisition, an O&M Process!



The point of all this "domain modelling" is to decide what things are important to describe, and understand their relationships. Then we know:

1. These are the important things to know (have in our metadata),
2. That we can build tools which manipulate the information (including the links).

MOLES in practice

Consider Chilbolton Observatory:

Home to many instruments, and the base for many observation campaigns.

MOLES in practice

Consider Chilbolton Observatory:

Home to many instruments, and the base for many observation campaigns.

MOLES has five primary entities: Dataset Collections, Datasets, Projects, Instruments, and Platforms. We can search against any or all. Today (3/12/14):

- There are **10** projects which are related to Chilbolton, including "ARSF flights in the Chilbolton area", "Cirrus and Anvils: European Satellite and Airborne Radiation", "Convective Storm Initiation Project" . . .
- There are **28** related instruments, including a "GBS satellite receiver". If I navigate from the Chilbolton project page to the GBS instrument page, I discover three related acquisitions, two in and around Chilbolton, and one in Dundee . . .
- There are **189** datasets . . . at which point our MOLES interface isn't as functional as we would like . . .

. . . we're back in a maze of twisty passages. There is more to do. But we now have the information infrastructure!

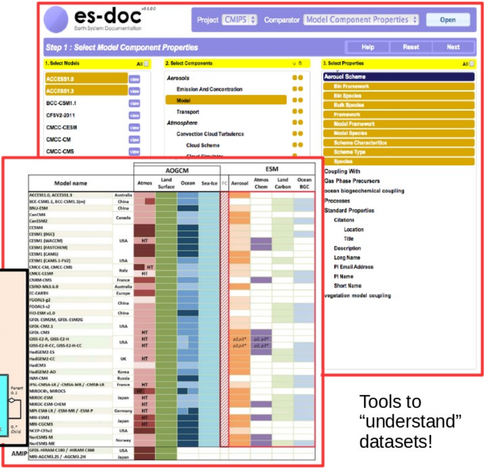
MOLES

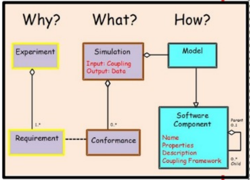
- Version 2 was deployed operationally at CEDA for about five years, and replaced with version 3 in October 2014 after a multi-year multi-person project to develop and then upgrade the catalogue (which involved migrating over 5000 records, many of which needed new content — the tyranny of scale again!).
- Now in version 3, with over a decade of development. Version 3 is a specialisation of ISO19156 Observations and Measurements.
- We already know lots of things wrong with MOLES3, not least the user interface!
- We have already started thinking about MOLES4, which will be some years away, but that's what "curation" means!

esdoc - documenting earth system models

CMIP5 (23/05/13):

- 101 experiments
- 61 model variants
- 590,000 datasets!
- 4.5 million files
- 2 PB in global archive
- Unknown PB locally!

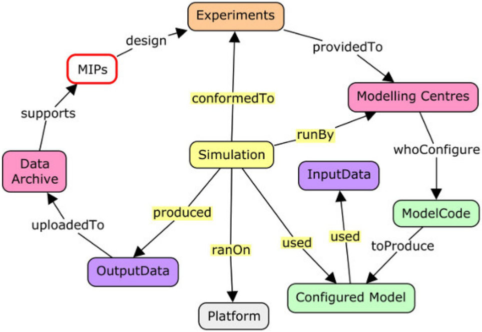




Tools to "understand" datasets!

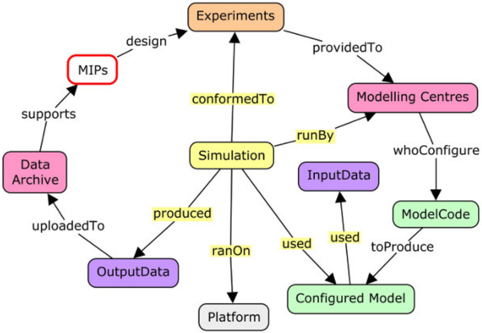
(Major global initiative - with the nuclei in NOAA, NCAS and IPSL!)

esdoc - the information cycle



MIPs define experiments, which are then used by modelling centres, who configure models to exploit (particular) inputs to run conforming simulations (on platforms) which produce data for archival.

esdoc - the information cycle



MIPs define experiments, which are then used by modelling centres, who configure models to exploit (particular) inputs to run conforming simulations (on platforms) which produce data for archival.

It's fair to say that, despite much effort, we didn't do a great job of providing methods for collecting the information for CMIP5, but it turned out to be a hard problem!

We're working hard on doing better for CMIP6, starting by trying to do a better job of formally defining the experiments and the data request. Meanwhile,you can

- Search CMIP5 documentation at `http://search.es-doc.org`,

and

- Compare models at `http://compare.es-doc.org/`

A digression on quality (of data and metadata)

The internet says quality is

- An essential and distinguishing attribute; A degree or grade of excellence or worth
- The degree to which a man-made object or system is free from bugs and flaws . . .
- A perceptual, conditional and somewhat subjective attribute and may be understood differently by different people.
- The suitability of procedures, processes and systems in relation to the strategic objectives.

A digression on quality (of data and metadata)

The internet says quality is

- An essential and distinguishing attribute; A degree or grade of excellence or worth
- The degree to which a man-made object or system is free from bugs and flaws . . .
- A perceptual, conditional and somewhat subjective attribute and may be understood differently by different people.
- The suitability of procedures, processes and systems in relation to the strategic objectives.

Philosophically: Distinction between primary and secondary qualities (John Locke):

- Primary qualities are intrinsic to an object — a thing or a person — whereas
- Secondary qualities are dependent on the interpretation and context. Extrinisic!

Scientifically (I would assert):

- We often confuse quality with accuracy or our knowledge of uncertainty.
- Data Producers worry about intrinsic quality.
- Data Consumers worry about extrinsic quality.

Greg Leptoukh again:

Data provider vs. User perspective

- Algorithm developers and Data providers: solid science + validation
- Users: fitness for purpose
  - **Measuring Climate Change**:
    - *Model validation*: **gridded contiguous data with uncertainties**
    - *Long-term time series*: **bias assessment** is the must , especially sensor degradation, orbit and spatial sampling change
  - **Studying phenomena using multi-sensor data:**
    - **Cross-sensor bias** is needed
  - **Realizing Societal Benefits through Applications:**
    - *Near-Real Time for transport/event monitoring* - in some cases, **coverage and timeliness** might be more important that accuracy
  - **Educational** (generally not well-versed in the intricacies of quality) – **only the best products**

10/21/2011 Leptoukh QA4EO'11 8

Commentary Metadata

What is “C” Metadata?

- Post-fact annotations, e.g. citations, ad-hoc comments and notes;
- Results of assessments, e.g. validation campaigns, intercomparisons with models or other observations, reanalysis;
- Provenance, e.g. dependencies on other datasets, processing algorithms and chain, data source;
- Information about external events that may affect the data, e.g. volcanic eruptions, El-Nino index, satellite or instrument failure, operational changes to the orbit calculations.

Differs from “B-Browse” in that it originates from users or external entities, not original data providers, although it may migrate to become “B” with time.

CHARMe

CHARMe:

Sharing climate knowledge through commentary metadata and Linked Data

(Jan 2013 – Dec 2014)

CHARMe enables a range of annotations - Simple

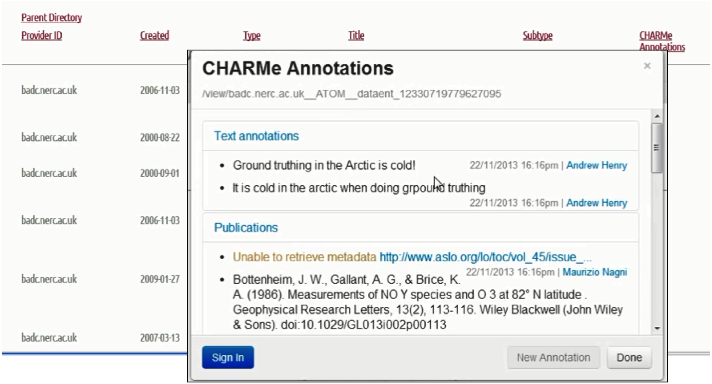
CHARMe slide content via Jon Blower

Parent Directory Provider ID	Created	Type	Title	Subtype	CHARMe Annotations
badc.nerc.ac.uk	2006-11-03	Data Entity	North Atlantic Marine Boundary Layer EXperiment (NAMBLEX) Data	Measurement	
badc.nerc.ac.uk	2000-08-22	Data Entity	Climatology Interdisciplinary Data Collection (CIDC)	Analysis	
badc.nerc.ac.uk	2000-09-01	Data Entity	Data from the Limb Infrared Monitor of the Stratosphere (LIMS) Instrument	Measurement	
badc.nerc.ac.uk	2006-11-03	Data Entity	Sea Surface Temperatures from the Along Track Scanning Radiometer (ATSR-1) - 1991-1995	Measurement	
badc.nerc.ac.uk	2009-01-27	Data Entity	COBRA (impact of COmbined iodine and Bromine Release on the Arctic atmosphere) Project Data	Measurement	
badc.nerc.ac.uk	2007-03-13	Data Entity	Upper Troposphere-Lower Stratosphere (UTLS)	Measurement	

Data provider hosts a "plugin" that allows annotations.

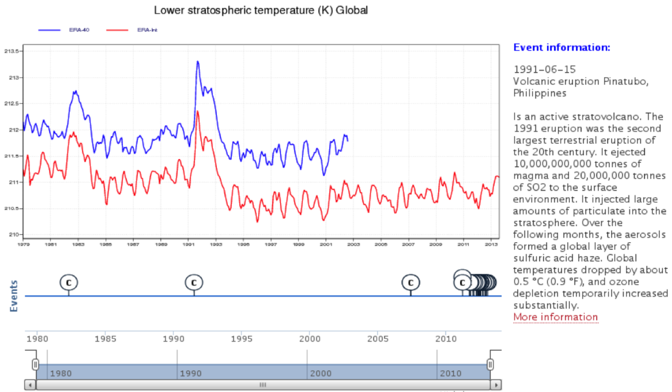
CHARMe enables a range of annotations - Simple

CHARMe slide content via Jon Blower



Annotations are posted to a "central annotation server" where they are viewable by anyone.

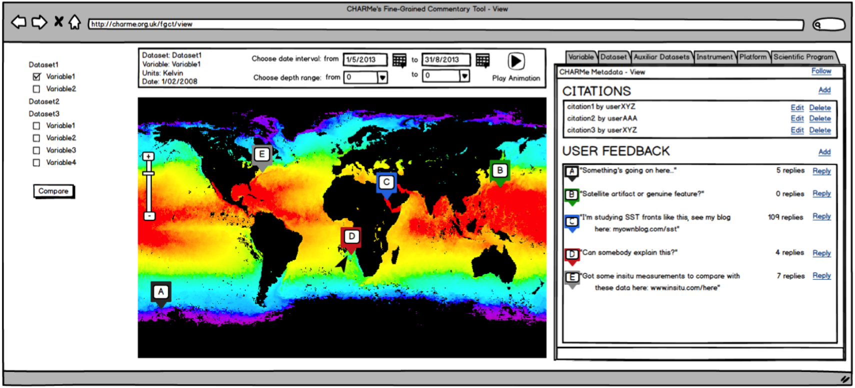
CHARMe enables a range of annotations - Events



Tool to match data with "significant events", e.g. algorithm changes, instrument failures, new data sources . . .

Will allow user annotations on the data to record events discovered before, during, and after data collection and/or production.

CHARMe enables a range of annotations - Maps



Tool will allow creation and discovery of commentary about specific parts of datasets, e.g. variables, geographic locations, time ranges.

The problem with extrinsic quality

THE PROBLEM WITH AVERAGING STAR RATINGS

The problem with extrinsic quality

THE PROBLEM WITH AVERAGING STAR RATINGS

first two cartoons from
xkcd.com
via Jon Blower

The problem with extrinsic quality

first two cartoons from
xkcd.com
via Jon Blower

IF cartoon all over the internet, original source unknown

The problem with extrinsic quality

THE PROBLEM WITH AVERAGING STAR RATINGS

first two cartoons from
xkcd.com
via Jon Blower

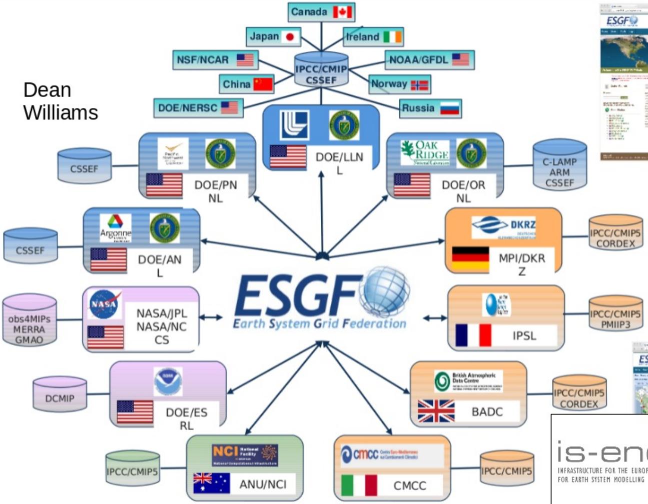
"If your IF is above 10, then you enter here. If it's lower, well... "

IF cartoon all over the internet, original source unknown

. . . but how do you book a hotel?

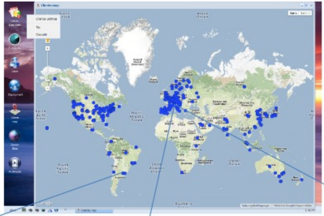
ESGF

Dean Williams



is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

The trend



Slide courtesy of Stefan Kindermann, DKRZ and IS-ENES2

Individual End Users
- Limited resources (bandwidth, storage,..)

Organized User Groups
- Organize a local cache of required files
- Most of group don't access ESGF, use cache instead!

Data Centre Service Group
- Provides access to ESGF replica cache
- May also provide access to data near compute resources
- (BADC, DKRZ, IPSL, KNMI, UC)

Trend

Needed: Replacement for „*Download and Process at Home*“ Approach

U.S. National Academy

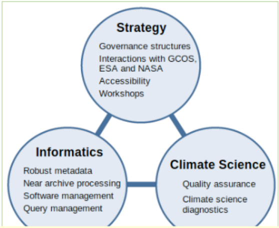
"Without substantial research effort into new methods of storage, data dissemination, data semantics, and visualization, all aimed at bringing analysis and computation to the data, rather than trying to download the data and perform analysis locally, it is likely that the data might become frustratingly inaccessible to users"

A National Strategy for Advancing Climate Modeling, 2012

Semantic Analysis: "substantial research effort" "new methods" "computation to data" "rather than trying to download" "frustratingly inaccessible" (to whom?)

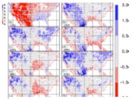
The ExArch Project - Taking compute to the data!

ExArch: Climate analytics on distributed exascale data archives (Juckes PI, G8 funded)

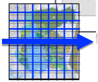


Martin Juckes, V. Balaji, B.N. Lawrence,
M. Lautenschlager, S. Denvil, G. Aloisio, P. Kushner, D. Waliser,
S. Pascoe, A. Stephens, P. Kershaw, F. Laliberte, J. Kim, S. Fiore

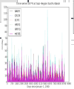
Regional Climate Model Evaluation System (RCMES)



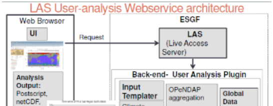
Observation/Model rainfall



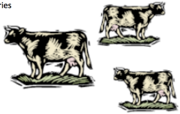
Map over a basin using an area-matching method



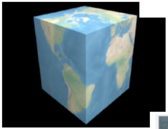
Get basin-mean time series



NOAA – PMEL "Live Access Server"



CEDA OGC Web Services



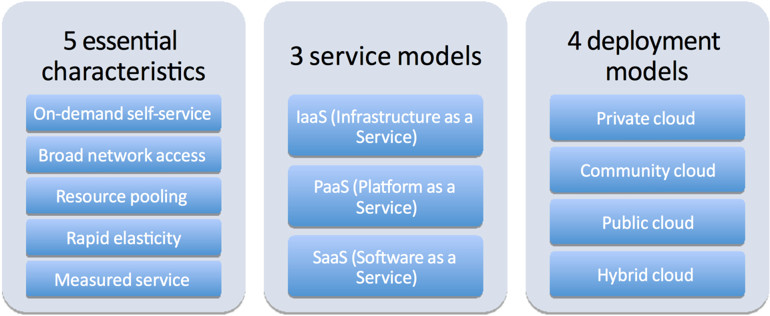
CMCC parallel data analytics framework

http://climate4impact.eu/

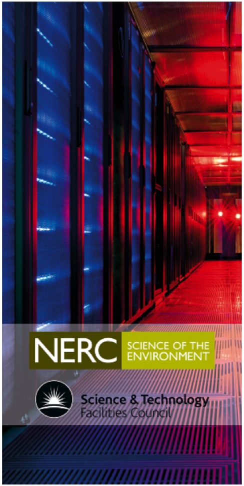
An introduction to the cloud

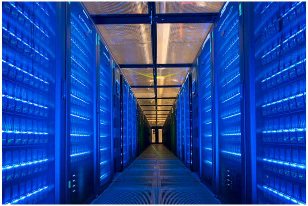
Why cloud? Remember all this communities, with their own software environments?

> *"Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction."* — *NIST SP800-145*



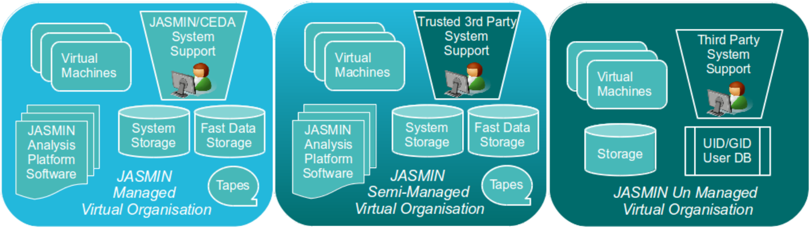
So we have built a Data-Intensive HPC cloud: JASMIN





- 12 PB Fast Storage
- 1 PB Bulk Storage
- Elastic Tape
- 4000 cores: half deployed as hypervisors, half as the "Lotus" batch cluster.

Virtual Organisations

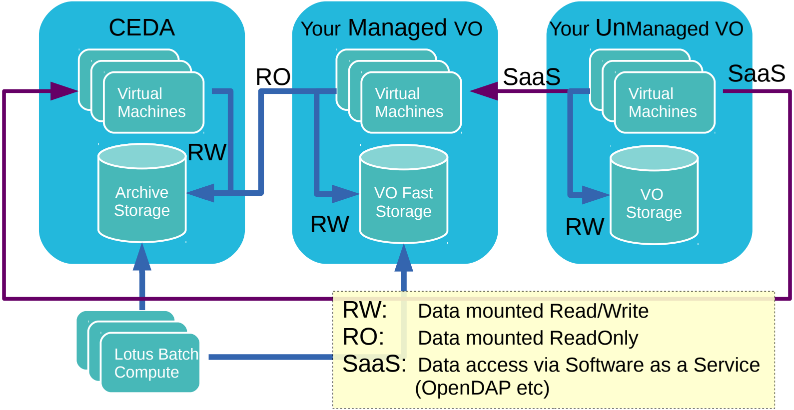


Platform as a Service $\longrightarrow$ Infrastructure as a Service

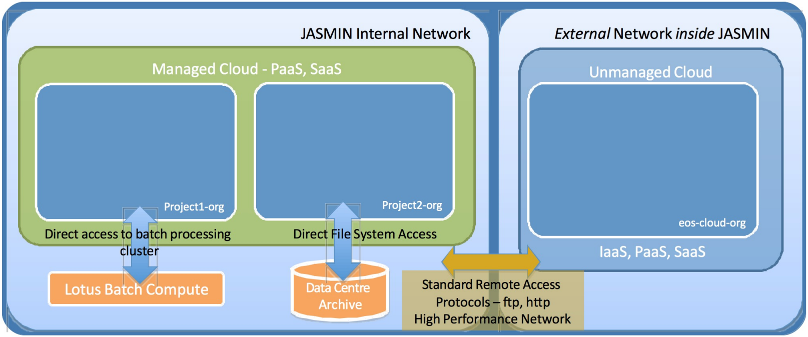
NCAS itself will run a semi-managed virtual organisation (with multiple group work spaces), but large groups within NCAS can themselves also run virtual organisations.

High performance, curation + facilitation

Objective is to provide an environment with high performance access to curated data archive **and** a high performance data analysis environment!

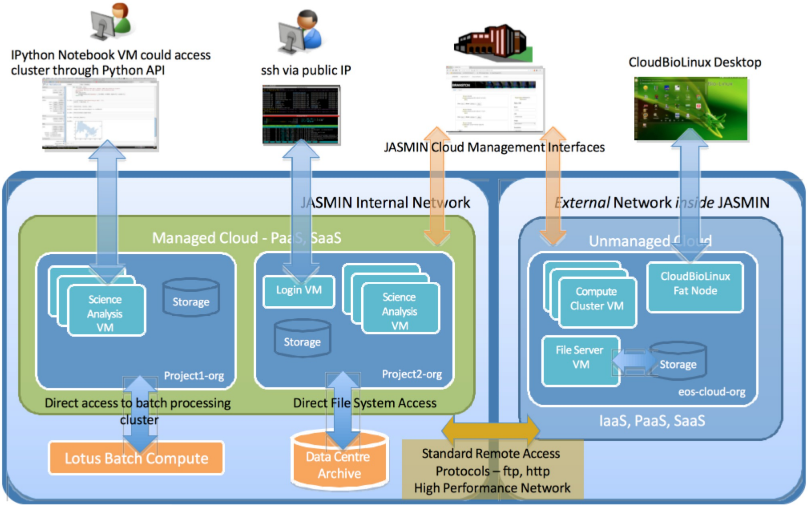


Integrated Cloud Provisioning

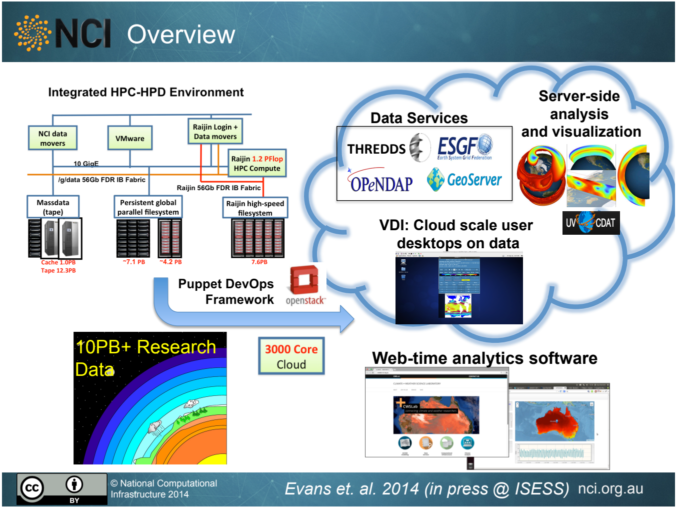


Currently o(100) "Group Work Spaces" in the managed cloud serving o(100) "virtual organisations" and o(500) users (there is some overlap). Unmanaged cloud is currently in testing with a few brave souls.

Integrated Cloud Provisioning



We are not the only ones

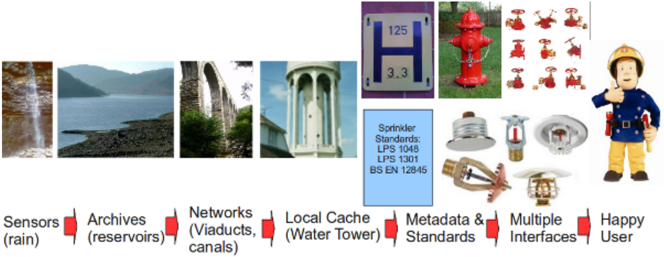


The Water Analogy

(A version of a cartoon I first saw in a Kevin Trenberth presentation, origin unknown.)

The Water Analogy

(A version of a cartoon I first saw in a Kevin Trenberth presentation, origin unknown.)



On Model Intercomparison Projects

- Model Intercomparison Projects are a response to two factors:
  1. The need to evaluate our assumptions and our models. Evaluation requires comparison, between simulations and reality, and differing ways of doing simulation.
  2. The need to share information between groups. Deciding what to do, sharing requirements, and sharing output, at scale, is hard, and needs organising.
- All of this is bigger than any one group can do alone. In the UK, the response has been the UKESM programme, bringing many more communities into the ESM frame.
- As move to more societally relevant science (the "grand challenge"), we will have more and more communities in play, and more evaluation to do.

Scale, Information, Collaboration and MIPs

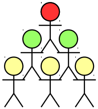
- Both the computing trends and the human trends are towards scale!
- At scale, we need to work harder on formally codifying information.
  - Dealing with the volume, variety, and velocity of data, being shared by multiple communities, requires more than just un-structured documents to capture information and requirements. Need metadata, targeted at specific parts of workflow.
  - There is a necessary inertia in developing, populating, and exploiting information systems at scale. Agility, at scale, is hard!

Scale, Information, Collaboration and MIPs

- Both the computing trends and the human trends are towards scale!
- At scale, we need to work harder on formally codifying information.
  - Dealing with the volume, variety, and velocity of data, being shared by multiple communities, requires more than just un-structured documents to capture information and requirements. Need metadata, targeted at specific parts of workflow.
  - There is a necessary inertia in developing, populating, and exploiting information systems at scale. Agility, at scale, is hard!
- We're all going to have to get used to "conforming" if we want to collaborate, and future MIPS are going to require even more information constraints, from definitions, to data outputs.

How do we work?

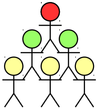
How we worked



PI stands on the shoulders of
her postdocs and students
(and as Newton would have
said, the giants.)

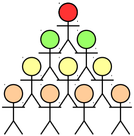
How do we work?

How we worked



PI stands on the shoulders of her postdocs and students (and as Newton would have said, the giants.)

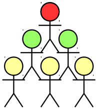
How we work



PI stands on the shoulders of her postdocs, students, software engineers and data scientists.
(Are the giants down with the turtles?).

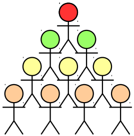
How do we work?

How we worked



PI stands on the shoulders of her postdocs and students (and as Newton would have said, the giants.)

How we work



PI stands on the shoulders of her postdocs, students, software engineers and data scientists.
(Are the giants down with the turtles?).

- It's fair to say that our institutions have not really caught onto the necessity to have careers for everyone in that stack.
- From the people managing vocabularies and manually entering metadata, to the software engineers and data scientists, we have new careers appearing, and we're not really ready for it.
- Mercifully we're not alone, bioinformatics is blazing a similar trail, but we have much to do.