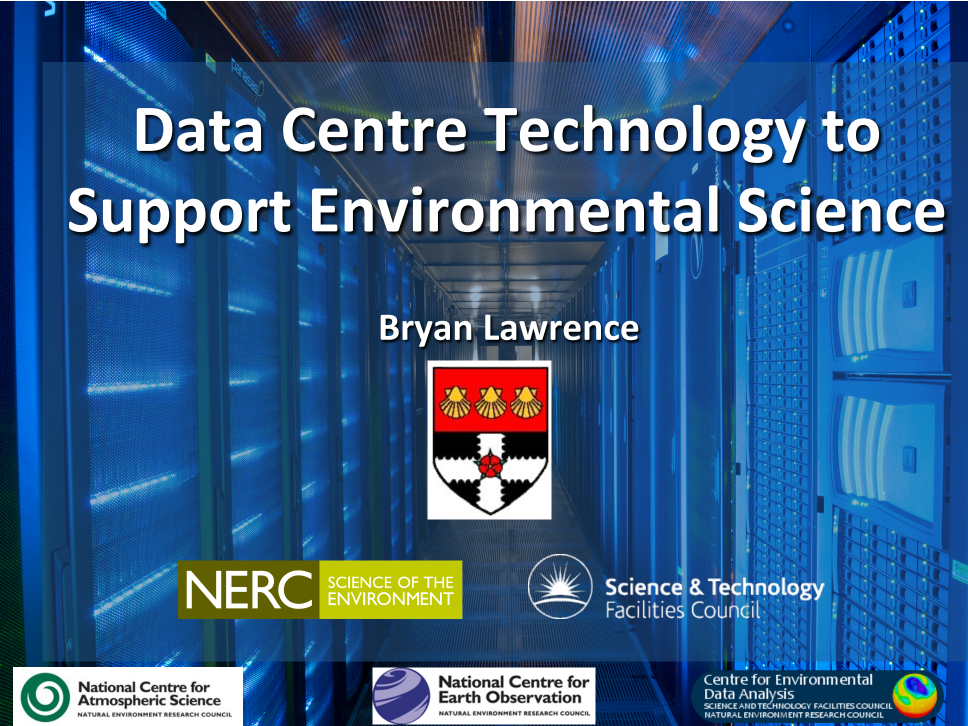


Data Centre Technology to Support Environmental Science

Bryan Lawrence

NERC SCIENCE OF THE ENVIRONMENT

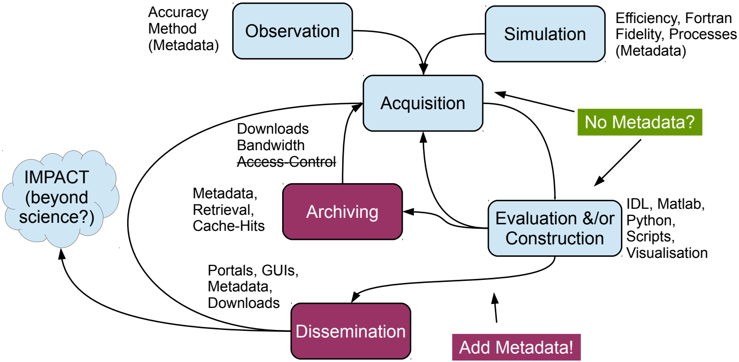
Science & Technology Facilities Council

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

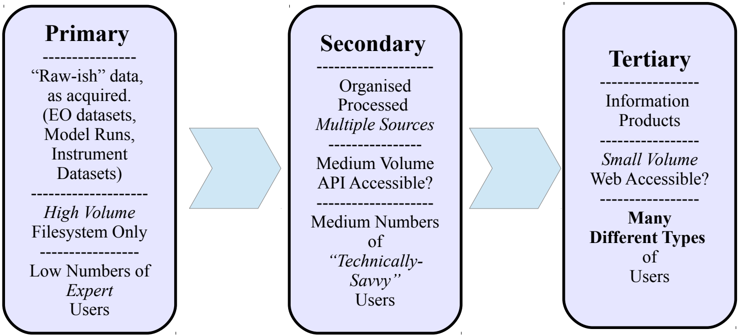
Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

Scientific Data Workflow



. . . with the role of “data centres” primarily in the purple boxes!

Transforming data into information



Primary

"Raw-ish" data,
as acquired.
(EO datasets,
Model Runs,
Instrument
Datasets)

High Volume
Filesystem Only

Low Numbers of
Expert
Users

Secondary

Organised
Processed
Multiple Sources

Medium Volume
API Accessible?

Medium Numbers
of
"Technically-Savvy"
Users

Tertiary

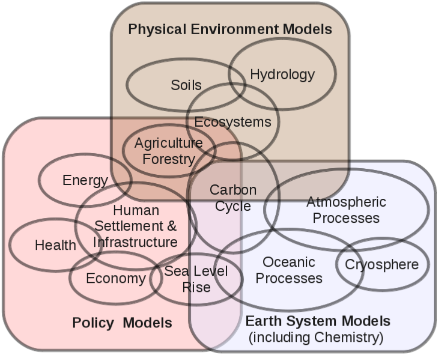
Information
Products

Small Volume
Web Accessible?

Many
Different Types
of
Users

All of these states and activities require "data centre support" for in situ, upstream and downstream users!

Growing range of interacting communities



Many interacting communities, each with their own software, compute environments, observations etc.

Figure adapted from Moss et al, 2010

The Rise of Direct Numerical Simulation

Primarily mathematical representation of a complex system of processes

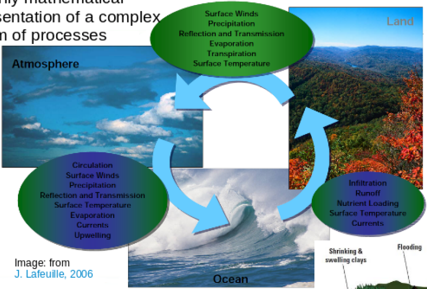
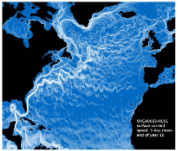
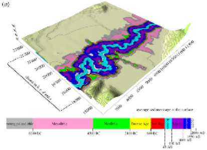
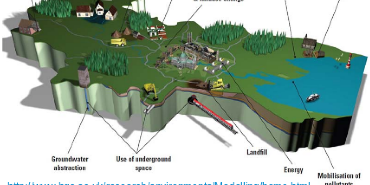


Image: from J. Lafeuille, 2006





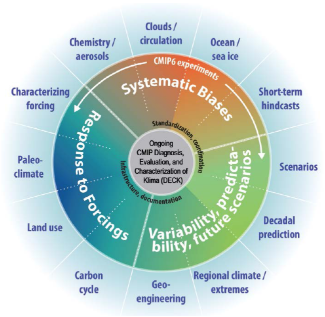
Coulthard and Van De Wiel lDoi: 10.1098/rsta.2011.0597



http://www.bgs.ac.uk/research/environmentalModelling/home.html

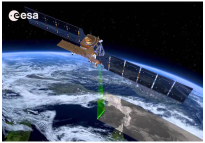
The Organised Data Deluge

WCRP
World Climate Research Programme



CMIP6 data volumes and data rates not yet known, but the European contribution to HiresMIP alone is expected to exceed 2 PB.

Copernicus
Europe's eyes on Earth

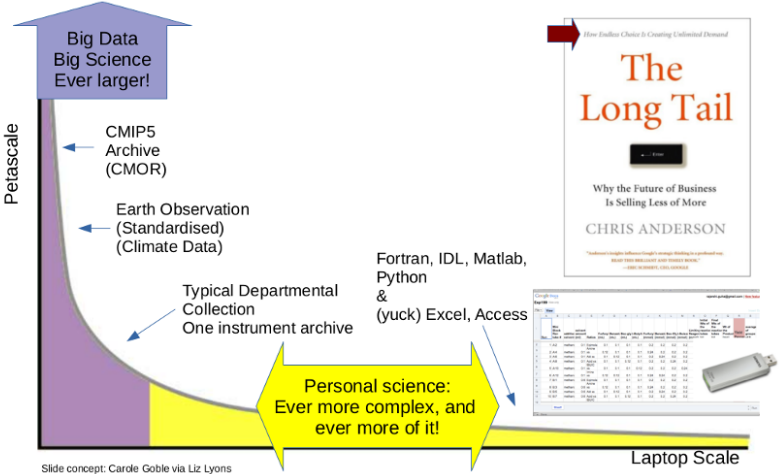


Sentinel 1A (2014), 1B (2016)
Sentinel 2A (2015) 2B (2017?)
Sentinel 3A (2016) 3B (2018?)

Data rate: o(6) PB/year

aerosol cci
cloud cci
fire cci
ghg cci
glaciers cci
antarctic ice sheet cci
ice sheets greenland cci
land couer cci
ocean colour cci
ozone cci
sea ice cci
sea leuel cci
sst cci
soil moisture cci
cmug cci

The unorganised data deluge



Big Data
Big Science
Ever larger!

Petascale

CMIP5
Archive
(CMOR)

Earth Observation
(Standardised)
(Climate Data)

Typical Departmental
Collection
One instrument archive

Fortran, IDL, Matlab,
Python
&
(yuck) Excel, Access

Personal science:
Ever more complex, and
ever more of it!

Laptop Scale

How Endless Choice Is Creating Unlimited Demand

The Long Tail

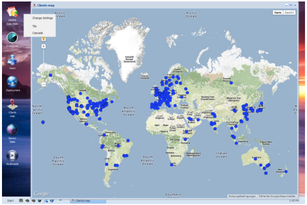
Why the Future of Business
Is Selling Less of More

CHRIS ANDERSON

Slide concept: Carole Goble via Liz Lyons

The consequences of data at scale - download doesn't work!

Earth System Grid Experience:



Slide content courtesy of
Stephan Kindermann, DKRZ
and IS-ENES2

Started with
Individual End Users

- Limited resources (bandwidth, storage)

Moved to
Organised User Groups

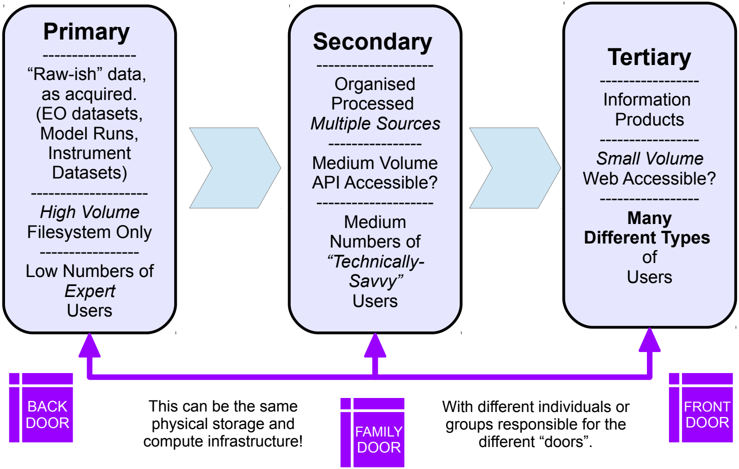
- Organize a local cache of files
- Most of the group don't access ESGF, but access cache.

Then
Data Centre Services

- Provide access to a replica cache
- May also provide compute near to data
- BADC, DKRZ, etc

Trend from download at home, to exploit a cache, to exploit a managed cache with compute!

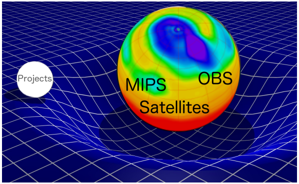
Transforming data into information - Revisited



This can be the same physical storage and compute infrastructure!

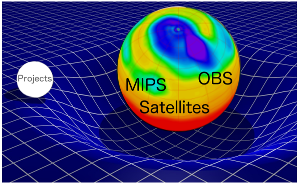
With different individuals or groups responsible for the different "doors".

JASMIN — The Data Commons

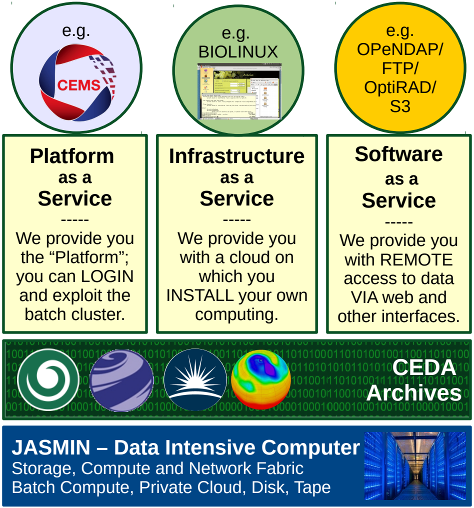


- Provide a state-of-the art storage and computational environment
- Provide and populate a managed data environment with key datasets (the "archive").
- Encourage and facilitate the bringing of data and/or computation alongside/to the archive!
- Provide FLEXIBLE methods of exploiting the computational environment.

JASMIN — The Data Commons



- Provide a state-of-the art storage and computational environment
- Provide and populate a managed data environment with key datasets (the “archive”).
- Encourage and facilitate the bringing of data and/or computation alongside/to the archive!
- Provide FLEXIBLE methods of exploiting the computational environment.



Platform as a Service

We provide you the “Platform”; you can LOGIN and exploit the batch cluster.

Infrastructure as a Service

We provide you with a cloud on which you INSTALL your own computing.

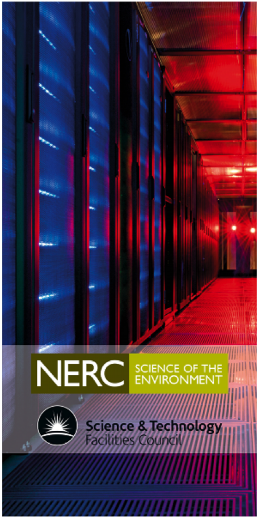
Software as a Service

We provide you with REMOTE access to data VIA web and other interfaces.

CEDA Archives

JASMIN – Data Intensive Computer
Storage, Compute and Network Fabric
Batch Compute, Private Cloud, Disk, Tape

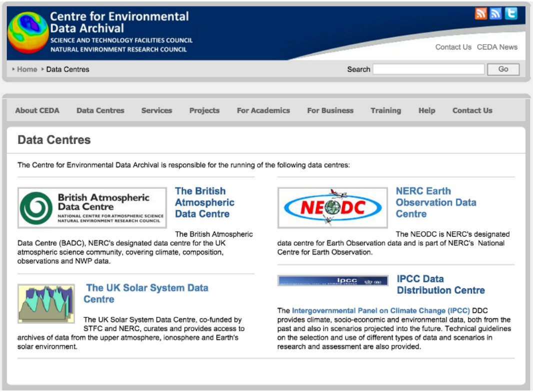
JASMIN


NERC SCIENCE OF THE ENVIRONMENT
Science & Technology Facilities Council

- 16 PB of fast disk; 0.5 PB of bulk disk (for virtual compute); >30 PB of tape.
- 5000 compute cores (cluster and hypervisors); dedicated high memory and transfer machines.

- The Archive - curated data directly available to local compute.
- Group Work Spaces — fast storage with tape accessible via the “Elastic Tape” service.
- Generic Platform Compute — machines configured for generic scientific analysis and data transfer.
- Hosted Platform Compute — bespoke machines deployed in the “Managed Cloud”.
- Infrastructure Compute — private cloud portal and customised compute in the “Un-Managed Cloud”.
- Lotus Batch Cluster — managed cluster with a range of node configurations (processor and memory).

CEDA


Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL
Contact Us
CEDA News
Home
Data Centres
Search
Go
About CEDA
Data Centres
Services
Projects
For Academics
For Business
Training
Help
Contact Us
British Atmospheric Data Centre
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL
NEODC
IPCC

Data Centres

The Centre for Environmental Data Archival is responsible for the running of the following data centres:

The British Atmospheric Data Centre

The British Atmospheric Data Centre (BADC), NERC's designated data centre for the UK atmospheric science community, covering climate, composition, observations and NWP data.

NERC Earth Observation Data Centre

The NEODC is NERC's designated data centre for Earth Observation data and is part of NERC's National Centre for Earth Observation.

The UK Solar System Data Centre

The UK Solar System Data Centre, co-funded by STFC and NERC, curates and provides access to archives of data from the upper atmosphere, ionosphere and Earth's solar environment.

IPCC Data Distribution Centre

The Intergovernmental Panel on Climate Change (IPCC) DDC provides climate, socio-economic and environmental data, both from the past and also in scenarios projected into the future. Technical guidelines on the selection and use of different types of data and scenarios in research and assessment are also provided.

Four internal data centres: http://ceda.ac.uk
Acquiring and Curating Data Archives

- Provides the initial mass for the “gravity well”, by feeding in both NERC and third party data products, available through the “back door”.
- An example of a tenant organisation in its own right, delivering services through the “front door”.
- Supports groups delivering customised services through “family doors”.

Other data centres could be tenants and contribute to the data commons in the same way.

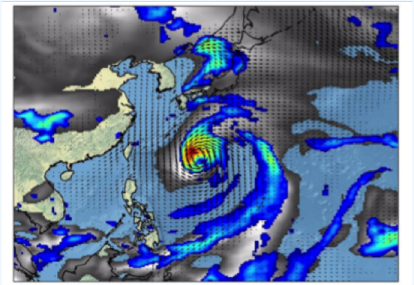
HRCM — simulating the building blocks of climate

The High Resolution Climate Modelling (HRCM) programme is a collaboration between the Hadley Centre (UK Met Office) and the NCAS Climate Directorate.

The programme produces and uses hundreds of terabytes of data, with data stored on a JASMIN Group Work Space and Elastic Tape.

The use of the JASMIN LOTUS batch cluster has

- enabled routine tracking of tropical cyclones from model simulations (50 years of N512 data can now be processed in one day with just 50 jobs).
- vastly sped up key analyses: e.g. calculation of eddy vectors has been reduced from 3 months to 24 hours with 1600 batch jobs.

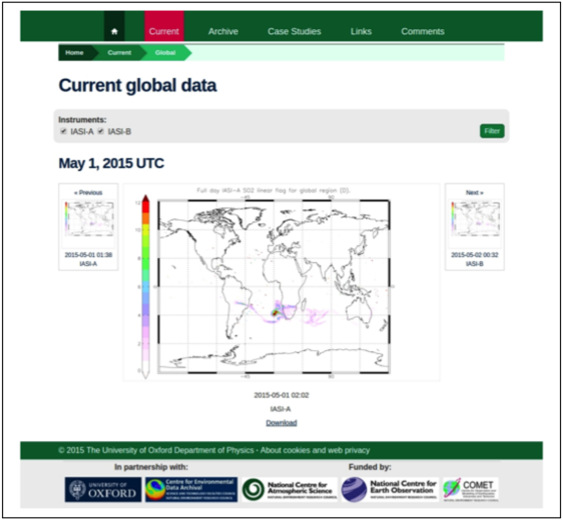


For more details contact: Prof P.L Vidale (NCAS, University of Reading) or visit https://hrcm.ceda.ac.uk/research/

Near-Real Time volcanic plumes on JASMIN

- Real-Time (NRT) observations of atmospheric disturbances such volcanic plumes of ash and SO2 are increasingly important, especially with respect to air travel.
- A volcanic SO2 monitoring website has been launched displaying near real time (NRT) data from both IASI instruments within 3 hours of measurement.
- The unique relationship available on JASMIN between data archive and data processing facilities is invaluable for this work.

More details: Elisa Carboni (University of Oxford) or visit http://www.nrt-atmos.cems.rl.ac.uk/


Current
Archive
Case Studies
Links
Comments
Home
Current
Global

Current global data

Instruments:
☑ IASI-A ☑ IASI-B

May 1, 2015 UTC

« Previous
2015-05-01 01:38
IASI-A
Full day IASI-A SO2 linear flag for global region (D).
-45
90
12
10
8
6
4
2
0
Next »
2015-05-02 00:32
IASI-B

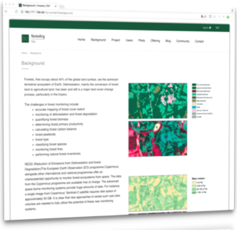
2015-05-01 02:02
IASI-A
Download

© 2015 The University of Oxford Department of Physics - About cookies and web privacy

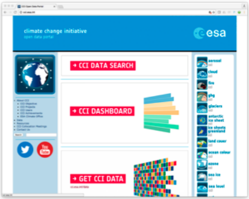
In partnership with:
Funded by:

UNIVERSITY OF OXFORD
Centre for Environmental Data Archival
National Centre for Atmospheric Science
National Centre for Earth Observation
COMET

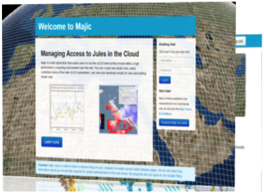
Virtual Research Environments on JASMIN hosted cloud



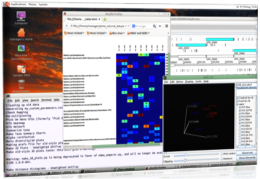
Thematic Exploitation Platforms for ESA



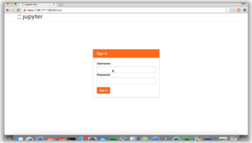
CCI Open Data Portal for ESA



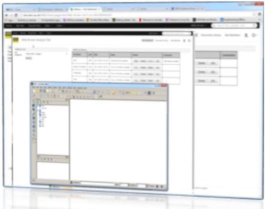
MAJIC interface to JULES model



EOS Cloud — Desktop-as-a-Service for Environmental Genomics

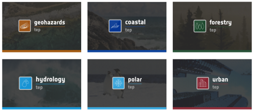


Hosted Ipython Notebooks



NERC Environmental Workbench

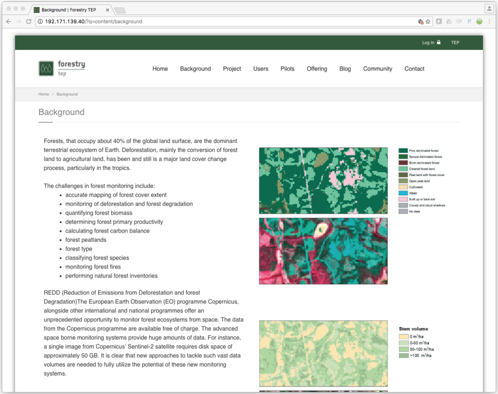
Thematic Exploitation Platforms for ESA



Forestry TEP

- A one-stop shop for forestry remote sensing services for the academic and commercial sectors.
- Offers access to pre-processed satellite and ancillary data, computing power, and software access and hosting.

. . . built by VTT Technical Research Centre & Arbonaut (FIN), CGI IT & STFC (UK), and Spacebel (BEL).



Background

Forests, that occupy about 40% of the global land surface, are the dominant terrestrial ecosystem of Earth. Deforestation, mainly the conversion of forest land to agricultural land, has been and still is a major land cover change process, particularly in the tropics.

The challenges in forest monitoring include:

- accurate mapping of forest cover extent
- monitoring of deforestation and forest degradation
- quantifying forest biomass
- determining forest primary productivity
- calculating forest carbon balance
- forest peatlands
- forest type
- classifying forest species
- monitoring forest fires
- performing natural forest inventories

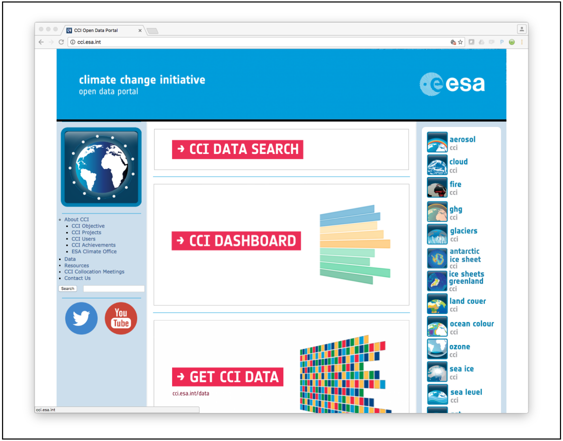
REDD (Reduction of Emissions from Deforestation and forest Degradation)The European Earth Observation (EO) programme Copernicus, alongside other international and national programmes offer an unprecedented opportunity to monitor forest ecosystems from space. The data from the Copernicus programme are available free of charge. The advanced space borne monitoring systems provide huge amounts of data. For instance, a single image from Copernicus' Sentinel-2 satellite requires disk space of approximately 50 GB. It is clear that new approaches to tackle such vast data volumes are needed to fully utilize the potential of these new monitoring systems.

CEDA is supporting the Forestry and Polar TEPS on the JASMIN un-managed cloud.

CCI Open Data Portal for ESA

The Climate Change Initiative

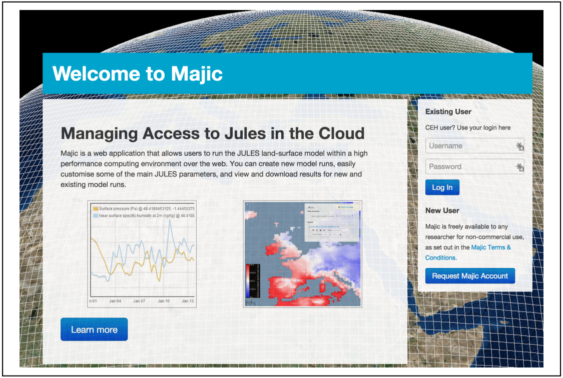
- Exploiting Europe's EO space assets to generate robust long-term global records of essential climate variables such as greenhouse-gas concentrations, sea-ice extent and thickness, and sea-surface temperature and salinity.
- The CCI Open Data Portal is hosted on JASMIN and exploits a near complete copy of the CCI datasets held in the CEDA archive.



MAJIC: Managing Access to JULES in the cloud

JULES
Joint UK Land Environment Simulator

- JULES is a community land surface model incorporating processes such as surface energy balance, the hydrological cycle, carbon cycle, dynamic vegetation etc.
- MAJIC provides a web portal running in the un-managed cloud which allows users to configure JULES to run on the JASMIN/LOTUS batch cluster and return results.



Welcome to Majic

Managing Access to Jules in the Cloud

Majic is a web application that allows users to run the JULES land-surface model within a high performance computing environment over the web. You can create new model runs, easily customise some of the main JULES parameters, and view and download results for new and existing model runs.

Surface pressure (Pa) @ 48.4189453125, -1.44450378
Near surface specific humidity at 2m (kg/kg) @ 48.4189
n 01
Jan 04
Jan 07
Jan 10
Jan 13
Learn more

Existing User

CEH user? Use your login here

New User

Majic is freely available to any researcher for non-commercial use, as set out in the Majic Terms & Conditions.

Summary

- Key role of data centres in the scientific workflow, dealing with the range of data from primary data to tertiary data, from expert users to consumer.
- Underlying trends: more data (volume and variety), more communities, and (more complexity of workflow).

Summary

- Key role of data centres in the scientific workflow, dealing with the range of data from primary data to tertiary data, from expert users to consumer.
- Underlying trends: more data (volume and variety), more communities, and (more complexity of workflow).
- Data gravity is “a thing”! Users value having “other” data with “their” data — provided there is adequate compute and storage available.
- Data gravity leads to “data lakes”. With a data lake, it’s possible to have a range of entrances[1], from a front door for consumers to back doors for data experts.

[1] Yes, I know a lake with doors is approaching an oxymoron!

Summary

- Key role of data centres in the scientific workflow, dealing with the range of data from primary data to tertiary data, from expert users to consumer.
- Underlying trends: more data (volume and variety), more communities, and (more complexity of workflow).
- Data gravity is "a thing"! Users value having "other" data with "their" data — provided there is adequate compute and storage available.
- Data gravity leads to "data lakes". With a data lake, it's possible to have a range of entrances[1], from a front door for consumers to back doors for data experts.
- JASMIN provides a suitable environment for a "data commons", already supporting a range of data centres and users exploiting a range of "doors": from bespoke portals to batch cluster based data analysis.
- There is a strong argument that NERC should aggregate more of its data into the common environment (but perhaps not all, e.g. JASMIN won't offer commercial levels of service that some applications such as BGS commercial might need).

[1] Yes, I know a lake with doors is approaching an oxymoron!