Data Interoperability and Integration: A climate modelling perspective

Bryan Lawrence

NERC SCIENCE OF THE ENVIRONMENT

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

On standards and history in meteorology and climate

Weather and Climate have been at the standards game for a long time:

- 1873 First International Meteorological Conference — (attendees recognised) "if there is any branch of science in which it is especially advantageous to work according to a uniform system, then that branch is the study of the laws of weather."
  - (…but even then a tension between government/agencies and research.)
- It is always one thing to set a standard and quite another to implement it. Inertia, resistance, ignorance, competing standards, lack of resources, legal barriers, and dozens of other problems must be overcome.
- But gathering the numbers is only the beginning. One must trust them too. Methodological skepticism is the foundation of science, so creating trust is not an easy task, nor should it be.

Paul Edwards. *"A vast machine: computer models, climate data, and the politics of global warming".* MIT press. 2010.

The Rise of Direct Numerical Simulation

Primarily mathematical representation of a complex system of processes

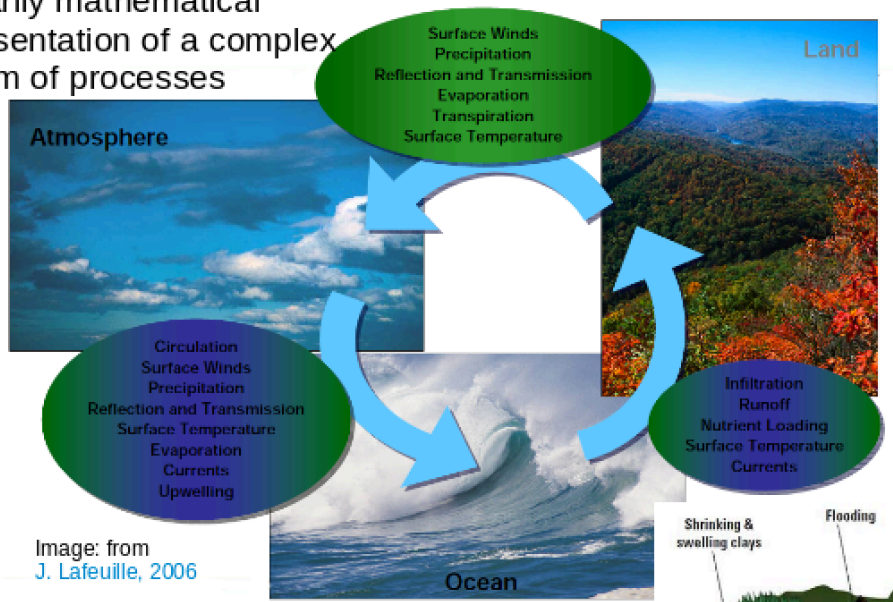
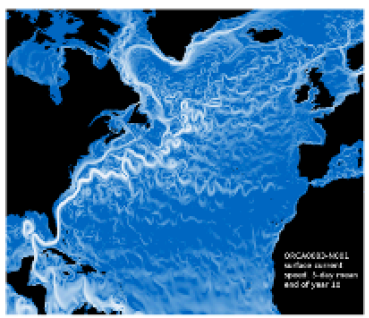
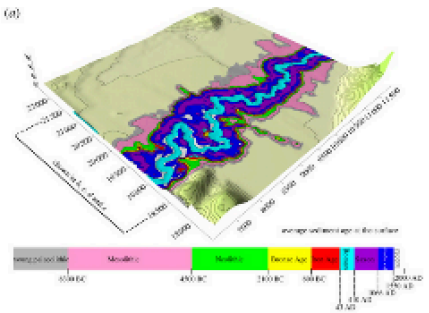
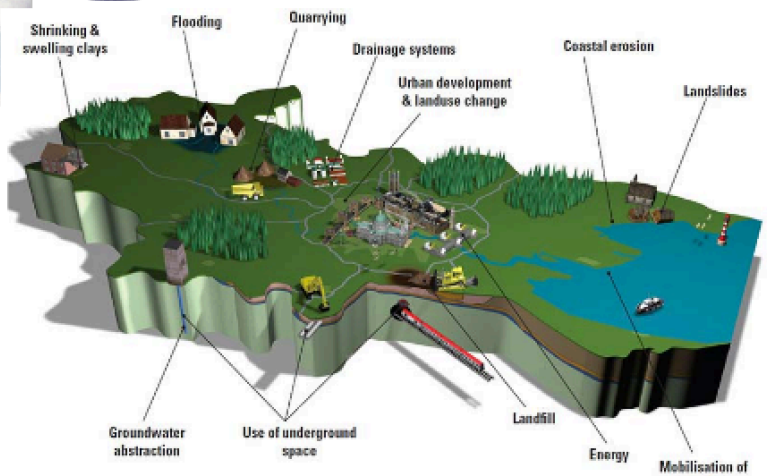


Image: from J. Lafeuille, 2006



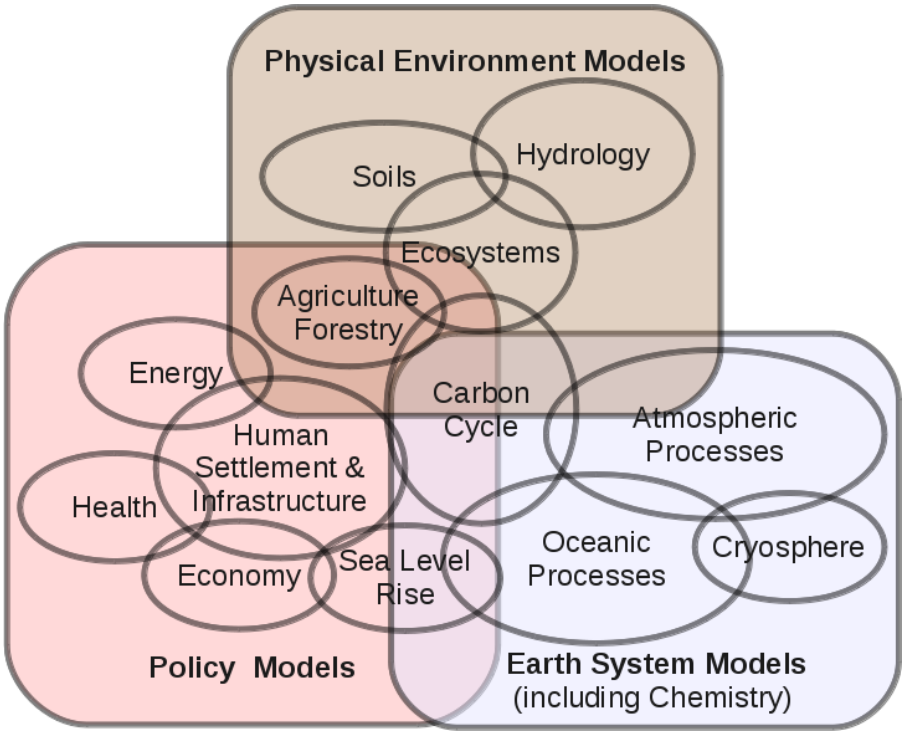


Coulthard and Van De Wiel IDoi: 10.1098/rsta.2011.0597



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

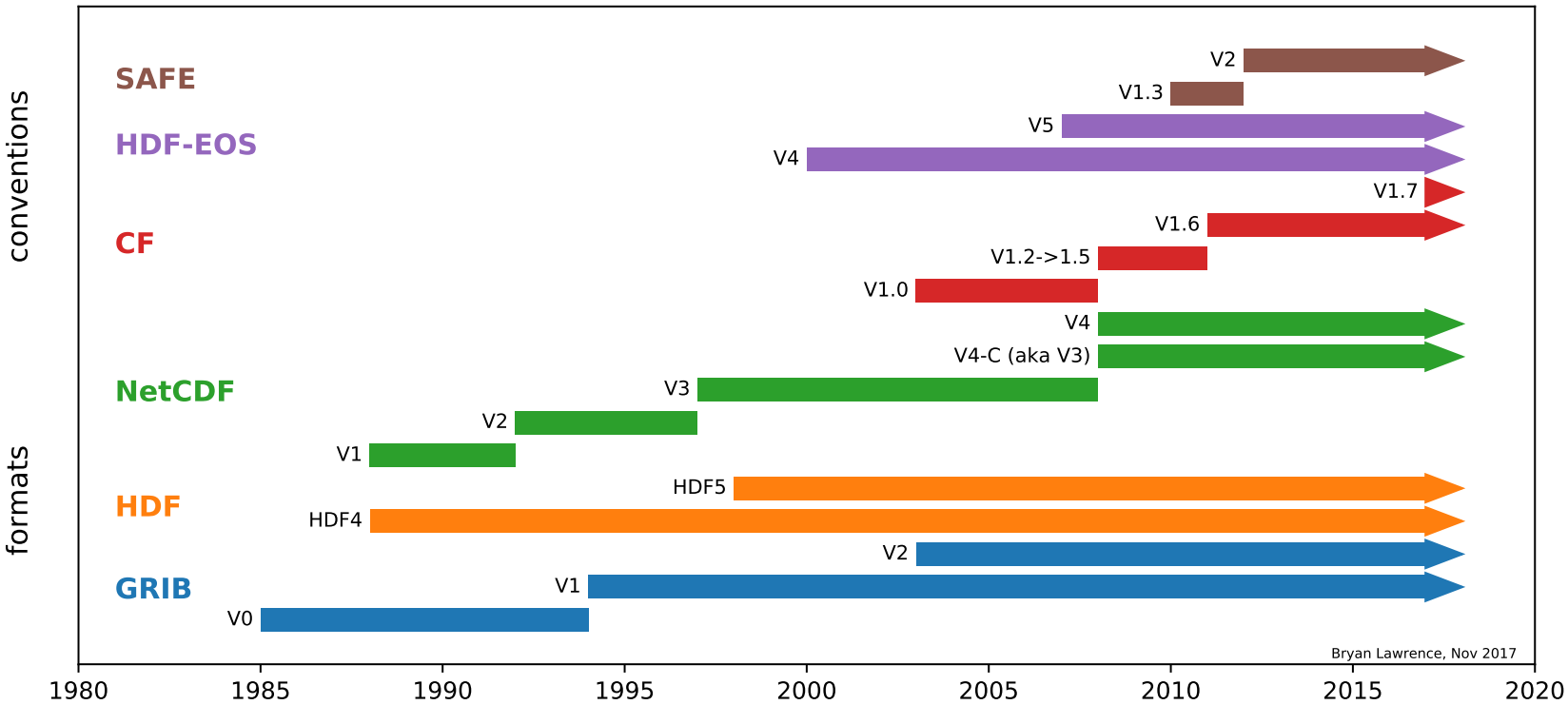
interacting communities



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

Figure adapted from Moss et al, 2010

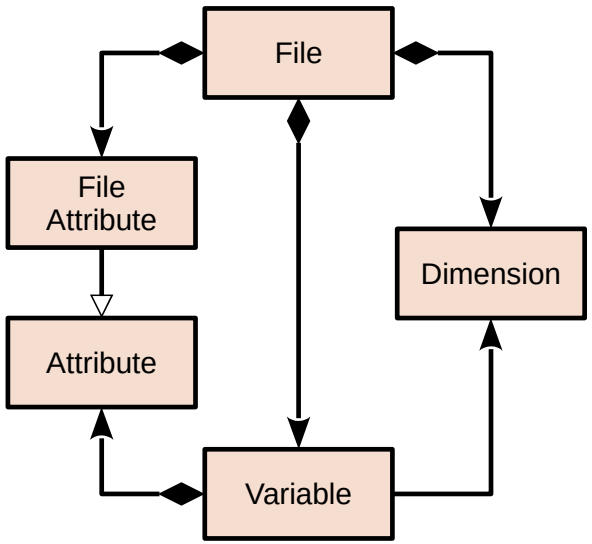
Formats and Conventions



We could discuss these formats and conventions …

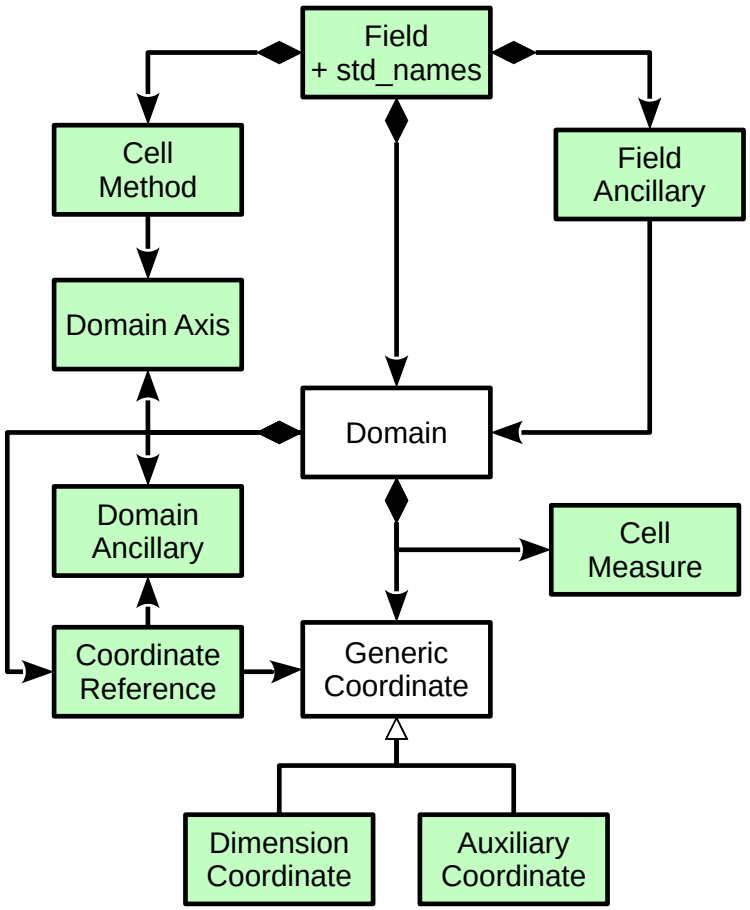
NetCDF3(&4-classic) and CF

…but formats are just (important) buckets …



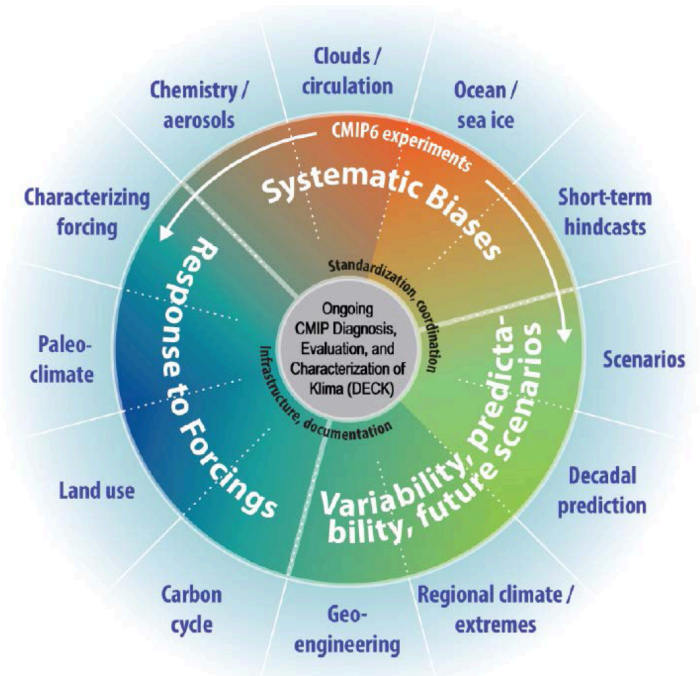
To make sense of them we need to interpret the attributes, and relationships between the variables, hence the Climate Forecast conventions!

CF Conventions



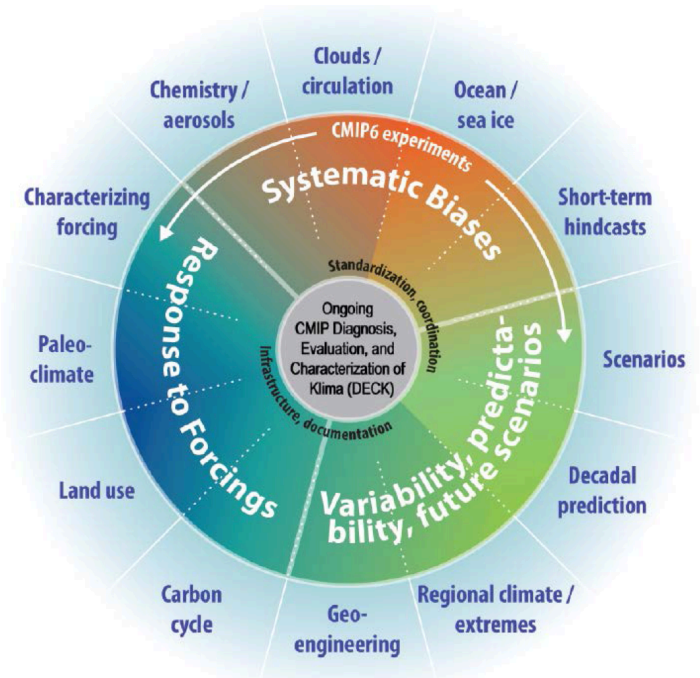
Hassel et al, 2017, GMD!

Model Data Intercomparison Projects — CMIP6



Volumes and Data Rates not yet known, but $>$ 3 PB from UK participants alone!

Model Data Intercomparison Projects — CMIP6

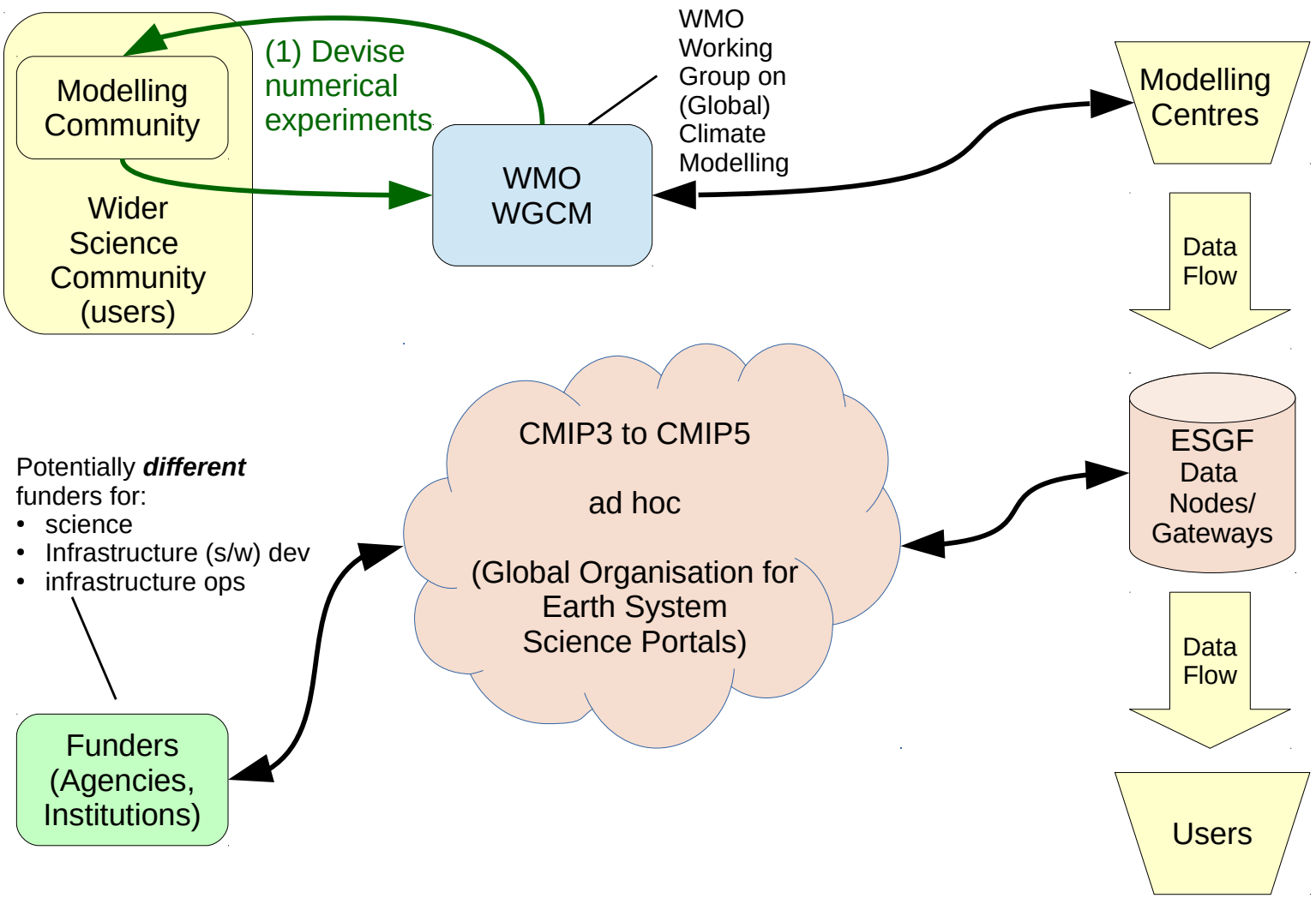


Volumes and Data Rates not yet known, but $>$ 3 PB from UK participants alone!

Requires globally distributed archive: Earth System Grid Federation (ESGF)

Particpants hold and share data. Some participants hold replicants of other data. All using common ESGF particular data formats and services.

How we used to do things

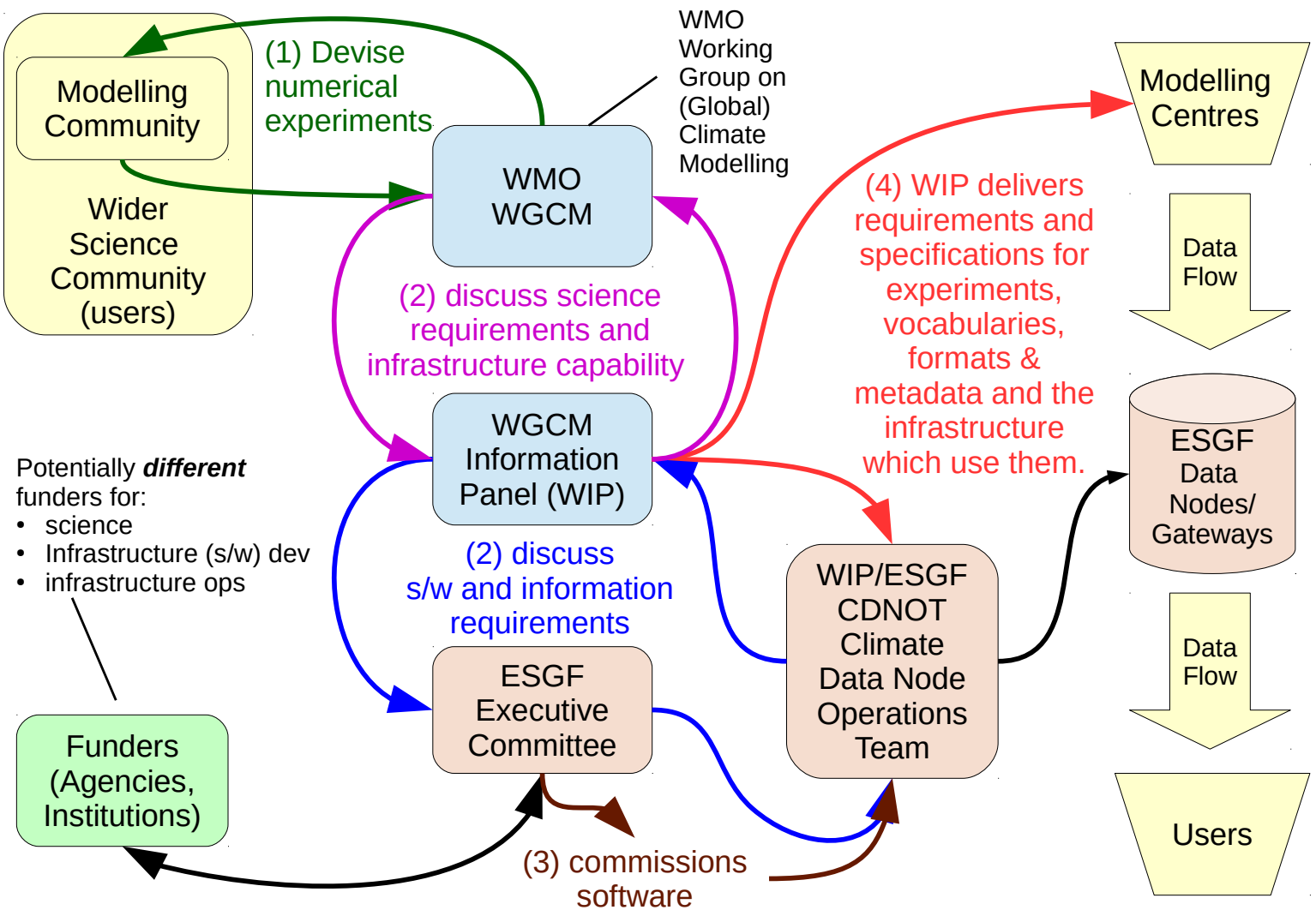


The WIP — WGCM Infrastructure Panel

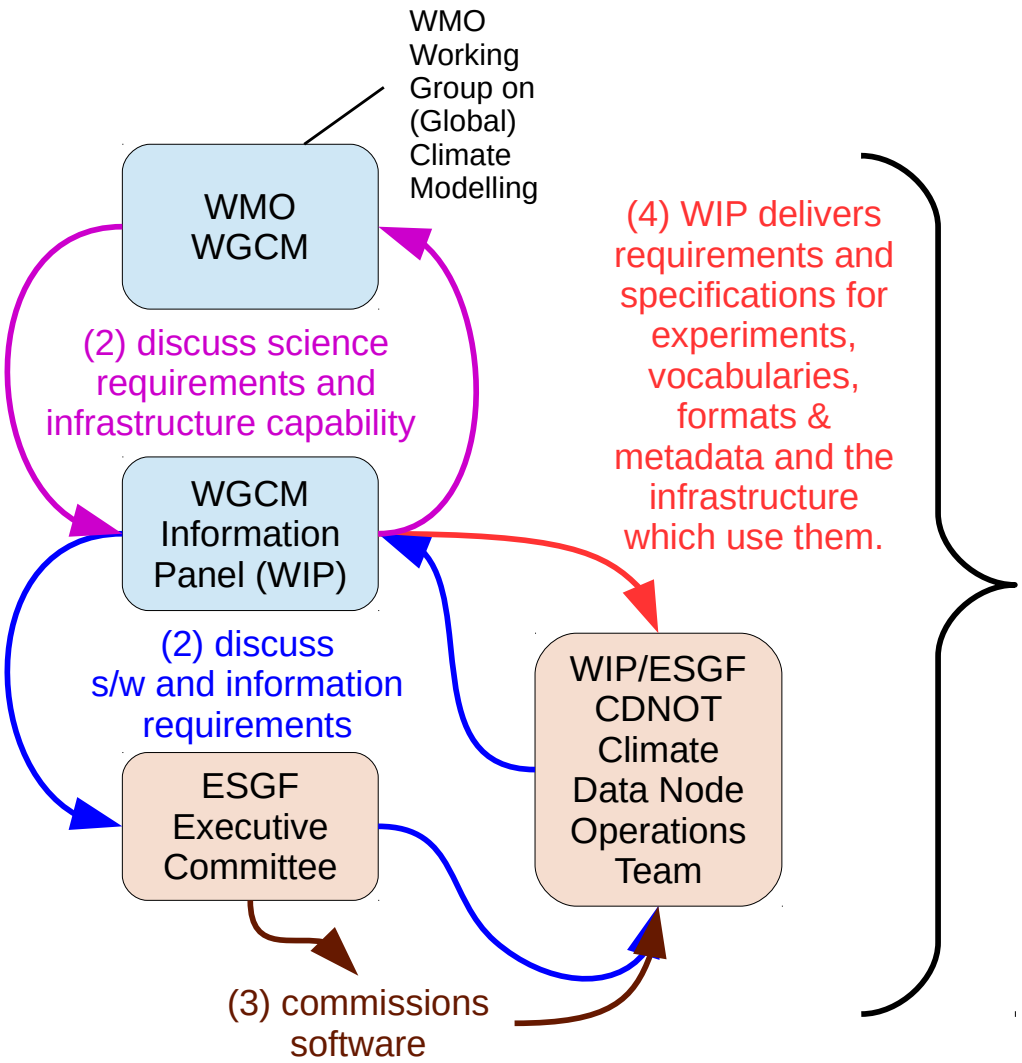
The WGCM (Working Group on Climate Models) Infrastructure Panel

…to maintain some control over the technical requirements imposed by (the science) . The membership will also include representation of those responsible for the standards and conventions and the IT and software infrastructure underpinning the MIPS. The mission …to promote a robust and sustainable global data infrastructure in support of the scientific mission of the WGCM. Drawing on experts intimately familiar with the scientific goals of the WGCM and aware of the promises and limitations of infrastructural technologies, the WIP will formulate achievable goals for global data infrastructure, ensure coordination of the various groups building components of the system, and advise the relevant institutions on the requirements and commitments needed to maintain its long term vitality.

How it has been working for CMIP6



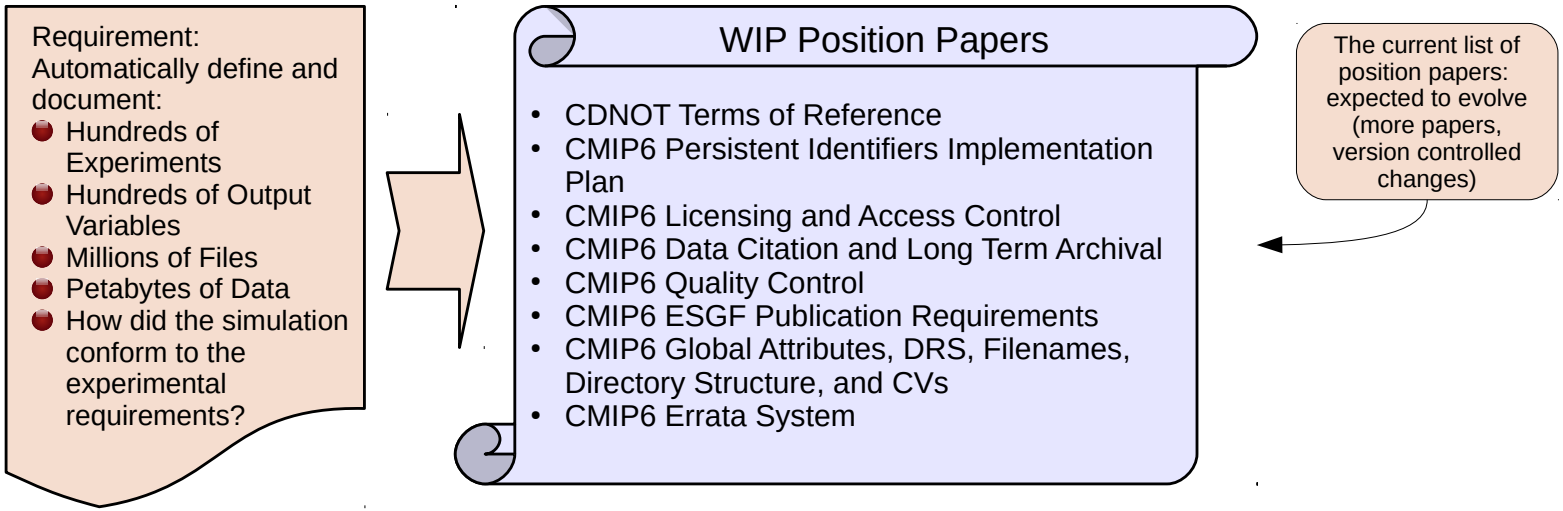
The WIP, feedback, and the virtuous circle



Virtuous Circle with feedback loops!

Science requirements confronting what can be built, deployed, and sustained.

WIP requirements and specifications



WIP requirements and specifications

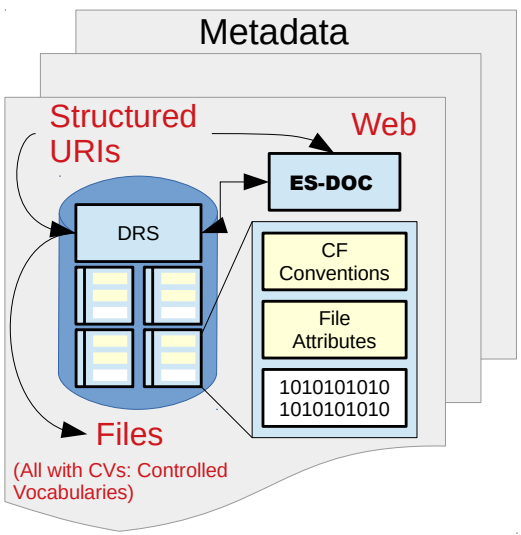
Requirement:
Automatically define and document:

- Hundreds of Experiments
- Hundreds of Output Variables
- Millions of Files
- Petabytes of Data
- How did the simulation conform to the experimental requirements?

WIP Position Papers

- CDNOT Terms of Reference
- CMIP6 Persistent Identifiers Implementation Plan
- CMIP6 Licensing and Access Control
- CMIP6 Data Citation and Long Term Archival
- CMIP6 Quality Assurance
- CMIP6 ESGF Publication Requirements
- CMIP6 Global Attributes, DRS, Filenames, Directory Structure, and CVs
- CMIP6 Errata System

The current list of position papers: expected to evolve (more papers, version controlled changes)



WIP requirements and specifications

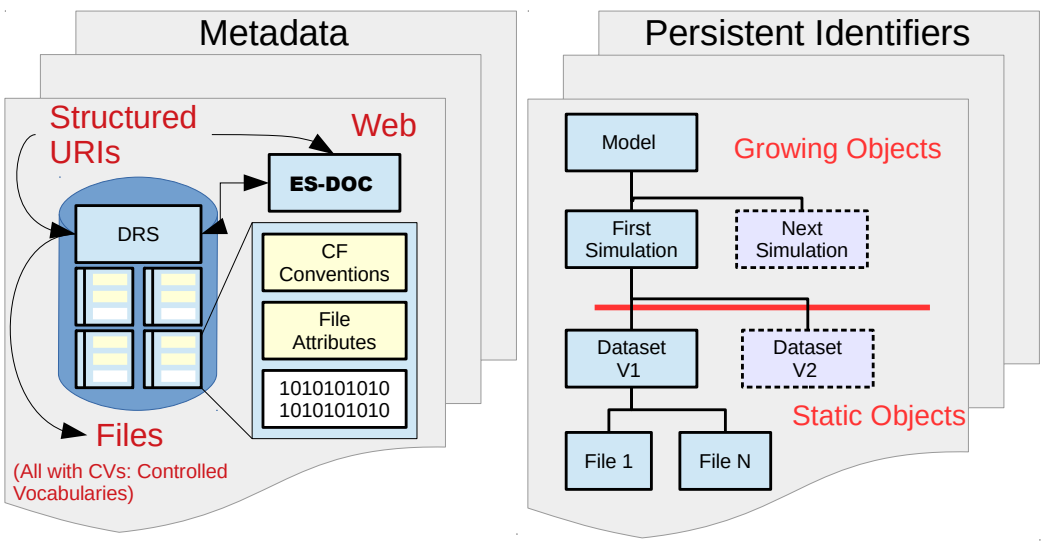
Requirement:
Automatically define and document:

- Hundreds of Experiments
- Hundreds of Output Variables
- Millions of Files
- Petabytes of Data
- How did the simulation conform to the experimental requirements?

WIP Position Papers

- CDNOT Terms of Reference
- CMIP6 Persistent Identifiers Implementation Plan
- CMIP6 Licensing and Access Control
- CMIP6 Data Citation and Long Term Archival
- CMIP6 Quality Assurance
- CMIP6 ESGF Publication Requirements
- CMIP6 Global Attributes, DRS, Filenames, Directory Structure, and CVs
- CMIP6 Errata System

The current list of position papers: expected to evolve (more papers, version controlled changes)



WIP requirements and specifications

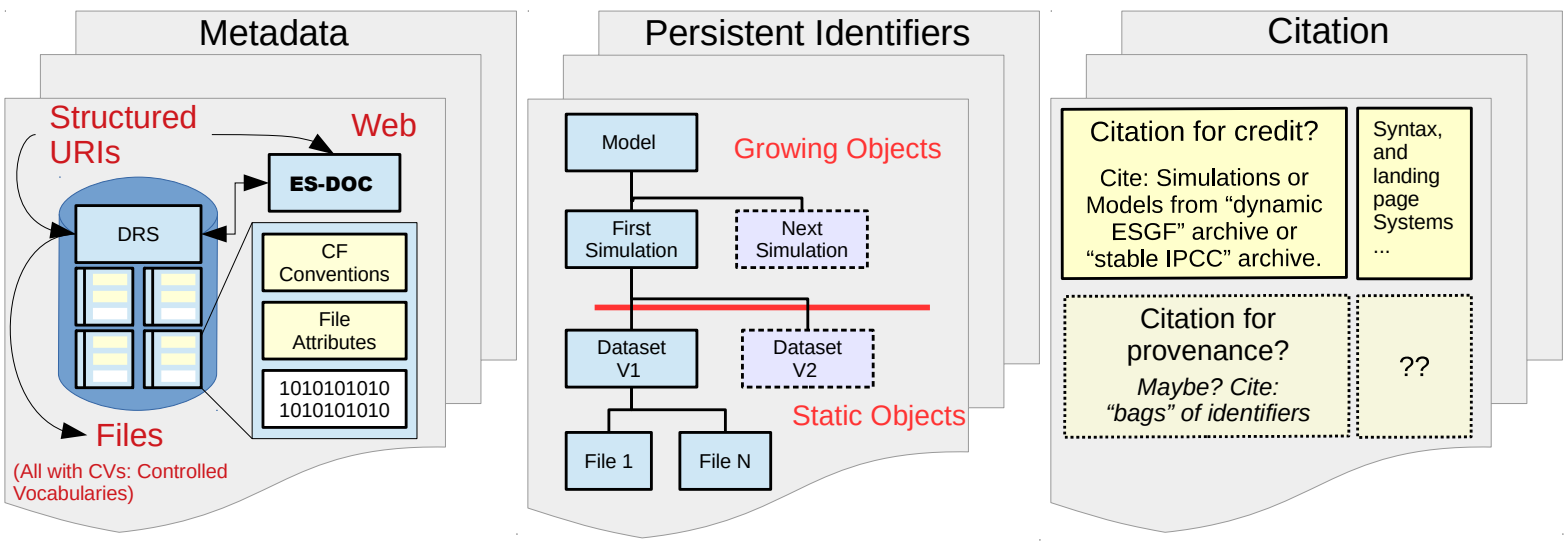
Requirement:
Automatically define and document:

- Hundreds of Experiments
- Hundreds of Output Variables
- Millions of Files
- Petabytes of Data
- How did the simulation conform to the experimental requirements?

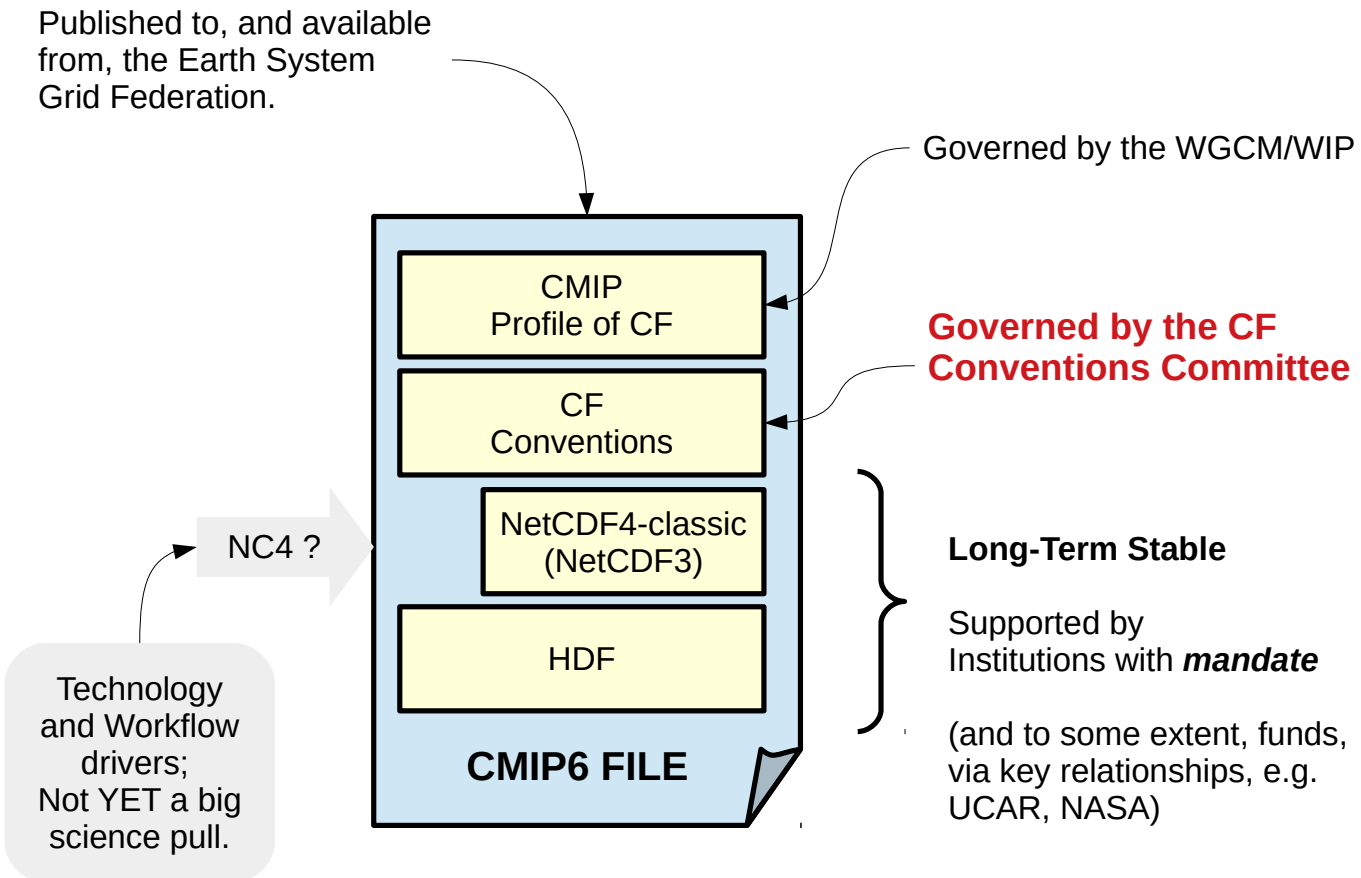
WIP Position Papers

- CDNOT Terms of Reference
- CMIP6 Persistent Identifiers Implementation Plan
- CMIP6 Licensing and Access Control
- CMIP6 Data Citation and Long Term Archival
- CMIP6 Quality Assurance
- CMIP6 ESGF Publication Requirements
- CMIP6 Global Attributes, DRS, Filenames, Directory Structure, and CVs
- CMIP6 Errata System

The current list of position papers: expected to evolve (more papers, version controlled changes)

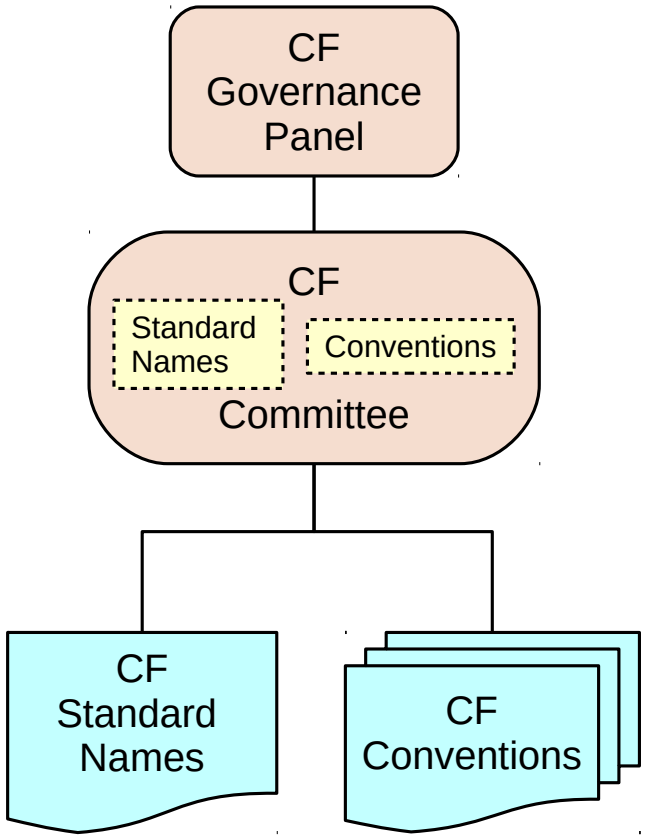


Strong Dependence on CF



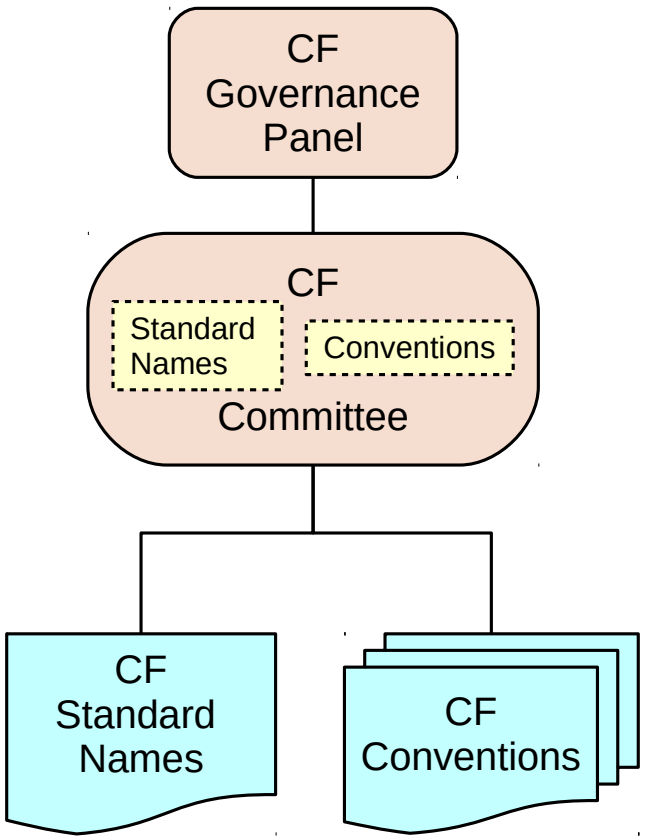
CF governance established by community itself

CF grew out of a need to work together in the context of climate model data intercomparison, but has taken on a much wider role since inception.



Lawrence, B. N., R. Drach, B. E. Eaton, J. M. Gregory, S. C. Hankin, R. K. Lowry, R. K. Rew, and K. E. Taylor. "Maintaining and advancing the CF standard for earth system science community data." (2005).

Attempts to anchor CF with third parties — 1

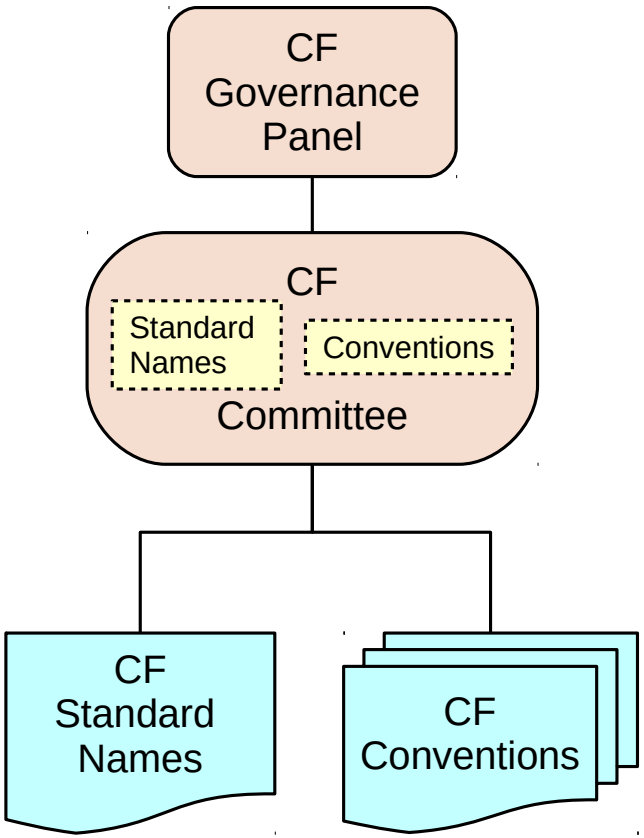


OGC®
Making location count.
www.opengeospatial.org

OGC network Common Data Form (netCDF) standards suite

CF-netCDF3 Data Model Extension Standard 11-165r2

Attempts to anchor CF with third parties — 1



OGC®
Making location count.
www.opengeospatial.org

OGC network Common Data Form (netCDF) standards suite

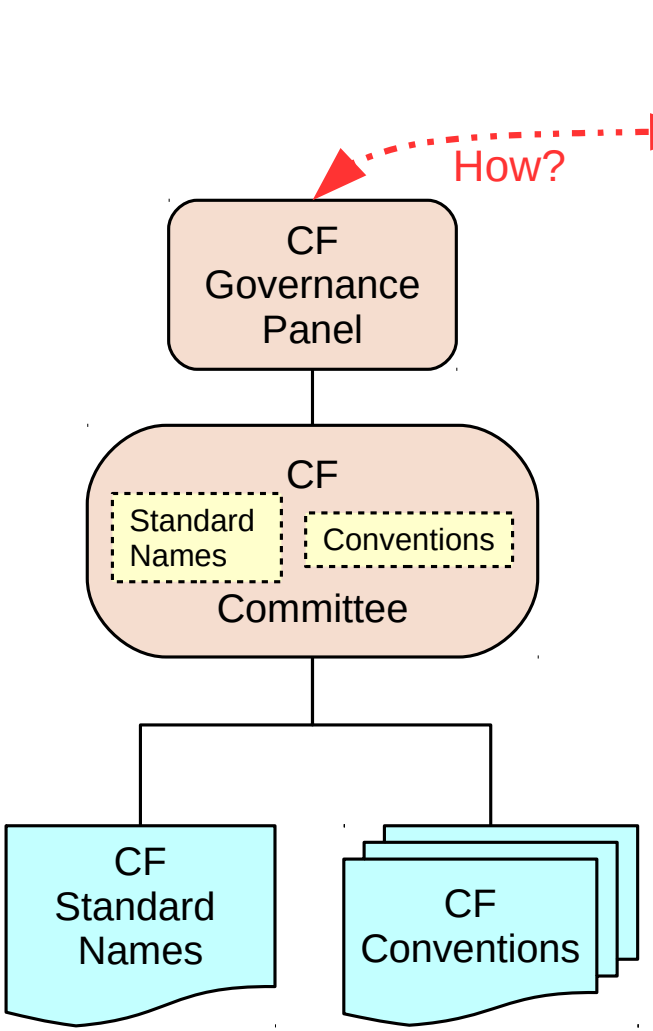
CF-netCDF3 Data Model Extension Standard 11-165r2

Developed by a body that took no cognisance of exiting governance procedures;

Ignored by the SCIENCE community!

No part of ongoing evolution!

Attempts to anchor CF with third parties — 2



WMO

WCRP World Climate Research Programme

WWRP

GCOS GLOBAL CLIMATE OBSERVING SYSTEM

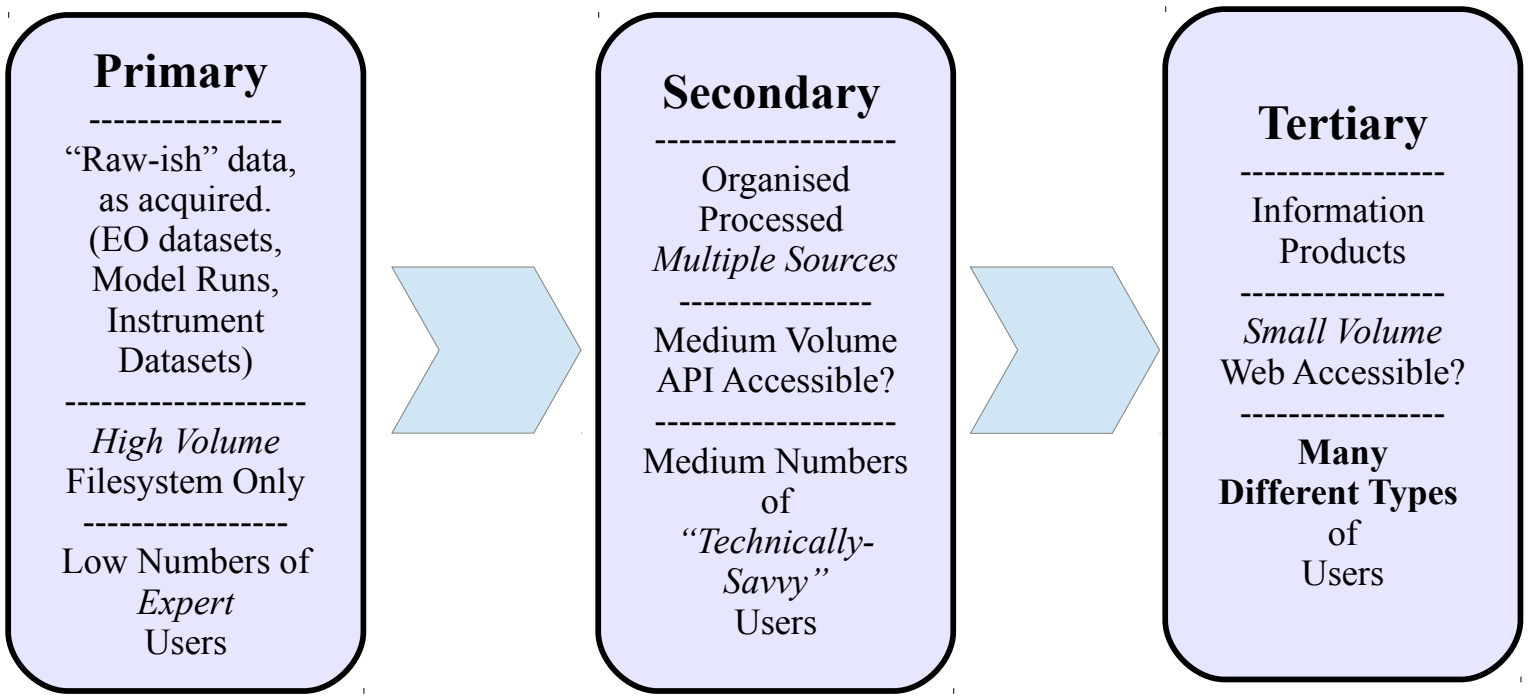
OGC®

OGC network Common Data Form (NetCDF) standards suite

CF netCDF3 Data Model Extension Standard 11-165r2

(Not operational bodies!)

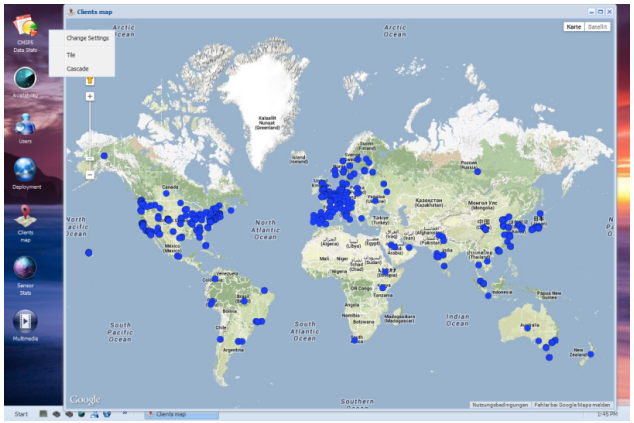
Transforming data into information



Need to avoid concentrating on standards and services which only address Tertiary Users, particularly at the cost of Secondary and Primary users!

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

Earth System Grid Experience



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

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

Started with **Individual End Users**

- Limited resources (bandwidth, storage)

Moved to **Organised User Groups**

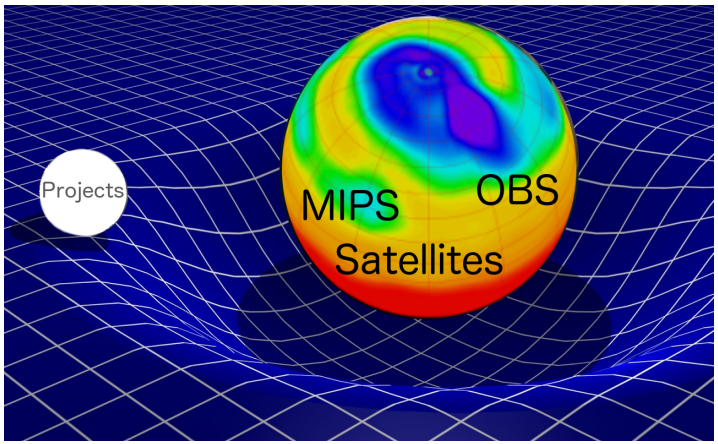
- Organize a local cache of files
- Most of the group don't access ESGF, but access cache.

Then **Data Centre Services**

- Provide access to a replica cache
- May also provide compute by data
- CEDA, DKRZ, etc

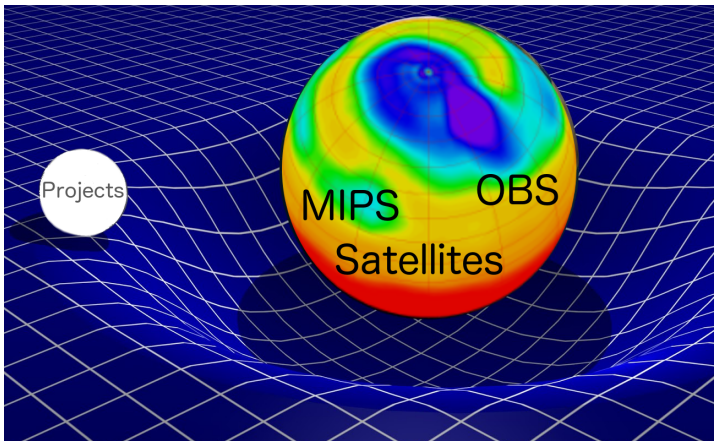
Trend from download at home, to exploit a cache, to exploit a managed cache with compute!

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 data and 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 data and computational environment.

ESGF OGC® Making location count. www.opengeospatial.org etc.

LOTUS

Optimised High Performance Data Analysis Environment

Community Cloud

Customisable

(with high performance route to archive)

CEDA Data Services

Remote access to archive & catalogues. Download etc

Service Nodes | ESF Nodes

CEDA Archives

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

What doesn't work so well

1. Relationship between funders and infrastructure/standardisation efforts is not good as it needs to be.
2. Some participants try and dominate and we don't have good formal methods of dealing with this …
3. …but even then, most processes are not agile enough.
4. Despite best efforts tension between need for "software and information" expertise and domain knowledge usually results in "engineers" and "scientists" who may not share the same vocabulary rather than properly multi-skilled individuals.
   - Reward structures for multi-skilled individuals are poorly conceived.
5. Disruptive influence of technology change; technology life-cycles are faster than than some informal consensus processes let alone standardisation.
6. Key individuals do not have enough time to contribute, and often no backup for their expertise.
   - Insufficient investment of both money and people.

What works well

1. Commitment from originators and maintainers of CF conventions, the WIP, from Unidata and The HDF Group, despite ongoing fragility of funding.
2. Relying on slowly changing format standards and resisting change for change's sake:
   - CF Design Principle: *Conventions should be developed only as needed, rather than anticipating possible needs.*
3. Formally recognising the tension between scientific demand and technology capability (both in terms of what can be done, and what it costs).
4. Standards such as CF which are "of the people" and "by the people" (people = scientists **and** engineers).
5. Anchoring responsibility in the scientific organisation (WGCM) rather than in a technology organisation.