When (and how) should a simulation be FAIR?

Bryan Lawrence

NCAS &
University of Reading: Departments of Meteorology and Computer Science

Boulder, CO, via Hangout, 5th May

Outline

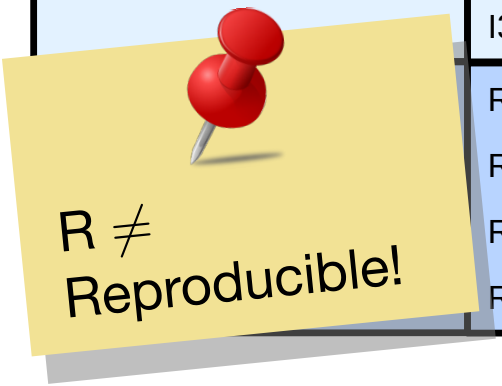
1. Reminder about FAIR, and what does it really mean?
2. Context: The problem with simulations:
   - What is a simulation?
   - What is simulation data?
3. Brief intro to the ES-DOC metadata system
4. Simulation Policy at CEDA
   - CEDA and JASMIN context
   - Overview of the policy.
5. Summary

FAIR principles

FINDABLE	F1. **(meta)data are assigned a globally unique and persistent identifier** F2. data are described with **rich metadata** (defined by R1 below) F3. **metadata** clearly and explicitly include the **identifier of the data** it describes F4. (meta)data are **registered or indexed in a searchable resource**
ACCESSIBLE	A1. (meta)data are **retrievable by their identifier** using a **standardized** communications **protocol**, which is **open, free, and universally implementable**, and **allows for an authentication and authorization procedure**, where necessary A2. metadata are accessible, even when the data are no longer available
INTEROPERABLE	I1. (meta)data use a **formal, accessible, shared**, and broadly applicable language for knowledge representation. I2. (meta)data **use vocabularies** that follow FAIR principles I3. (meta)data include qualified references to other (meta)data
RE-USABLE	R1. meta(data) are richly described with a **plurality of accurate and relevant attributes** R1.1. (meta)data are released with a clear and accessible data usage license R1.2. (meta)data are associated with **detailed provenance** R1.3. (meta)data meet **domain-relevant community standards**

FAIR principles

FINDABLE	F1. **(meta)data are assigned a globally unique and persistent identifier** F2. data are described with **rich metadata** (defined by R1 below) F3. **metadata** clearly and explicitly include the **identifier of the data** it describes F4. (meta)data are **registered or indexed in a searchable resource**
ACCESSIBLE	A1. (meta)data are **retrievable by their identifier** using a **standardized** communications **protocol**, which is **open, free, and universally implementable**, and **allows for an authentication and authorization procedure**, where necessary A2. metadata are accessible, even when the data are no longer available
INTEROPERABLE	I1. (meta)data use a **formal, accessible, shared**, and broadly applicable language for knowledge representation. I2. (meta)data **use vocabularies** that follow FAIR principles I3. (meta)data include qualified references to other (meta)data
	R1. meta(data) are richly described with a **plurality of accurate and relevant attributes** R1.1. (meta)data are released with a clear and accessible data usage license R1.2. (meta)data are associated with **detailed provenance** R1.3. (meta)data meet **domain-relevant community standards**



Yes,but?

1. What is a dataset? (What is the granularity of a dataset?)
2. What should “searchable” mean?
3. What protocol was that? (The open, free, universally implementable one, with AAA.)
4. When and why should *“data be no longer available”?*
5. Which formal accessible (and useful) metadata schema was that?
   - 5.1 Which relevant metadata attributes?”
   - 5.2 What does provenance mean (for a simulation)?”
   - 5.3 Which domain-relevant community standards)?

ESGF - Are these data FAIR?

WCRP CMIP6
World Climate Research Programme

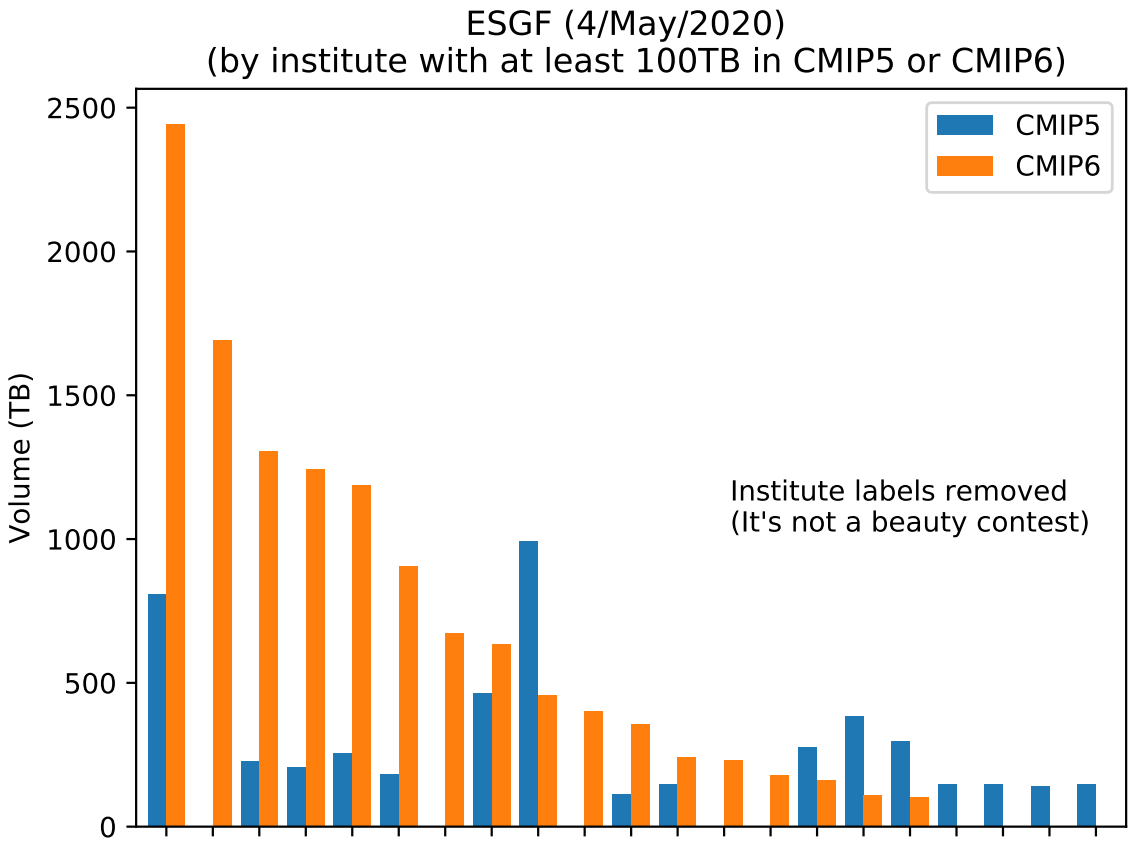
- MIP Era
- Activity
- Model Cohort
- Product

- Source ID
- Institution ID
- Source Type
- Nominal Resolution

- Experiment ID
- Sub-Experiment
- Variant Label
- Grid Label

- Table ID
- Frequency
- Realm
- Variable
- CF Standard Name

- Data Node



Lots of new entrants. Not all the big hitters have got going yet (caveat: some amalgamation of institutes into consortia). Even many of those with large volumes already are only part way through their simulation/publication cycle.

ESGF - Are these data FAIR?

WCRP CMIP6
World Climate Research Programme

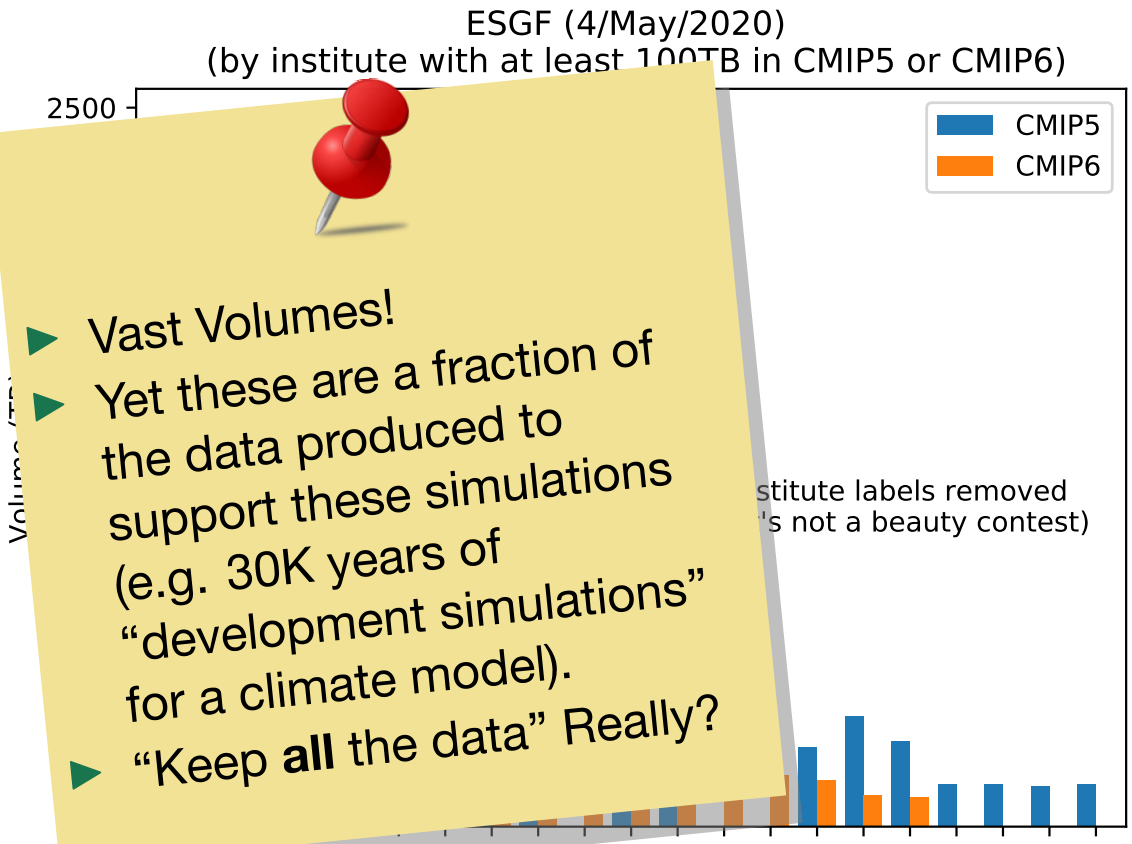
- MIP Era
- Activity
- Model Cohort
- Product

- Source ID
- Institution ID
- Source Type
- Nominal Resolution

- Experiment ID
- Sub-Experiment
- Variant Label
- Grid Label

- Table ID
- Frequency
- Realm
- Variable
- CF Standard Name

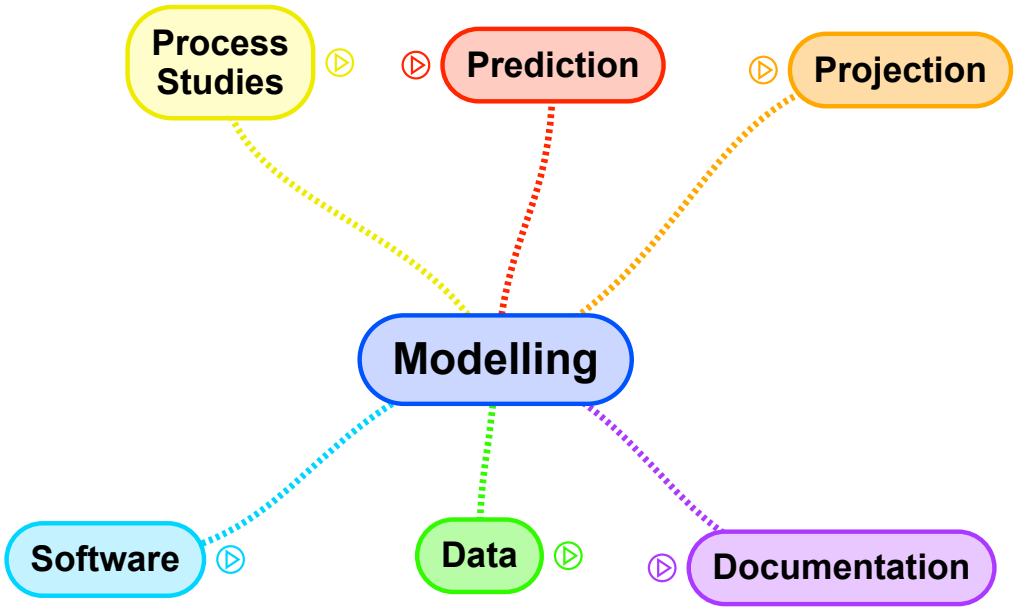
- Data Node



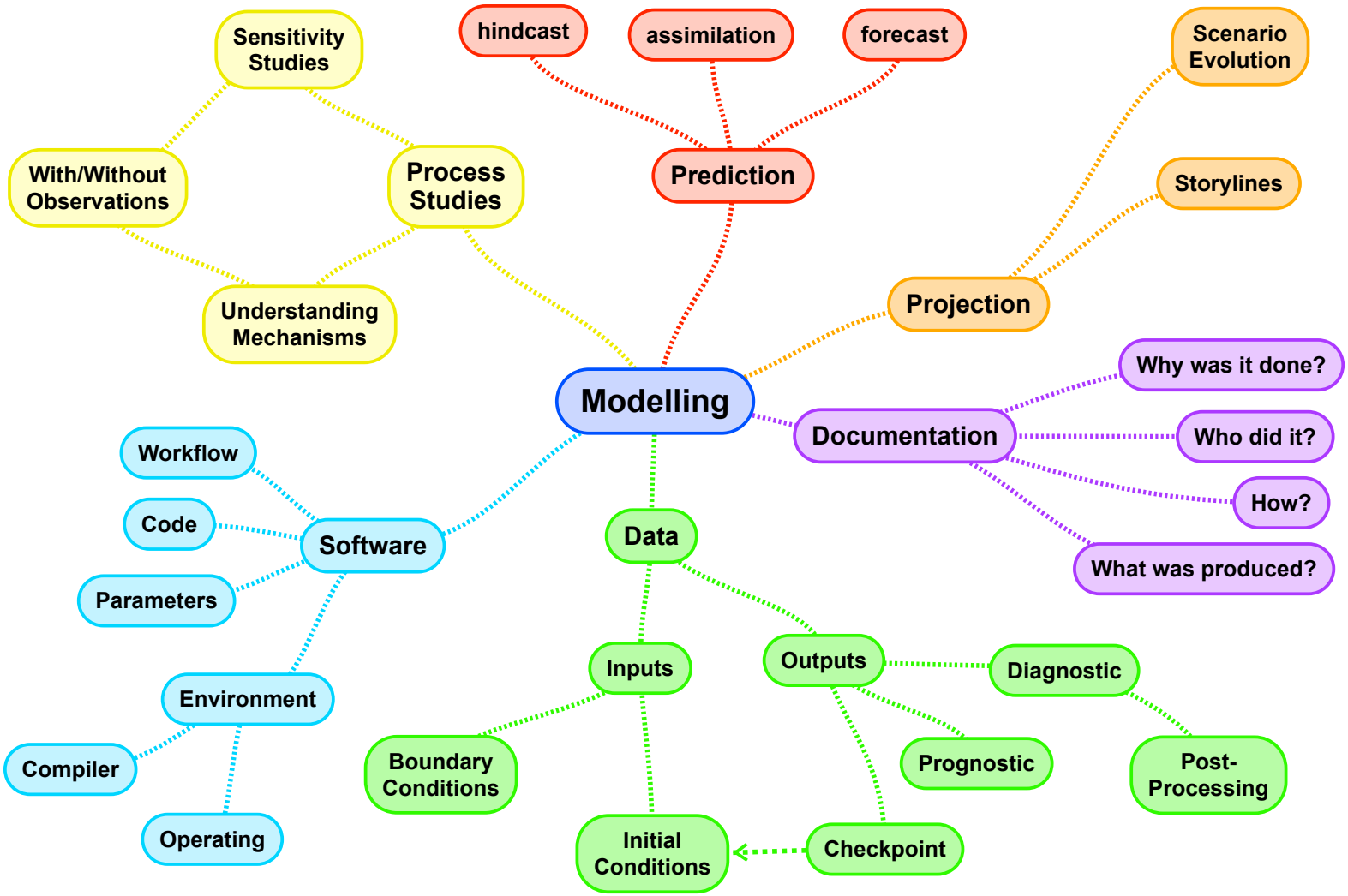
- Vast Volumes!
- Yet these are a fraction of the data produced to support these simulations (e.g. 30K years of "development simulations" for a climate model).
- "Keep **all** the data" Really?

Lots of new entrants. Not all the big hitters have got going yet (caveat: some amalgamation of institutes into consortia). Even many of those with large volumes already are only part way through their simulation/publication cycle.

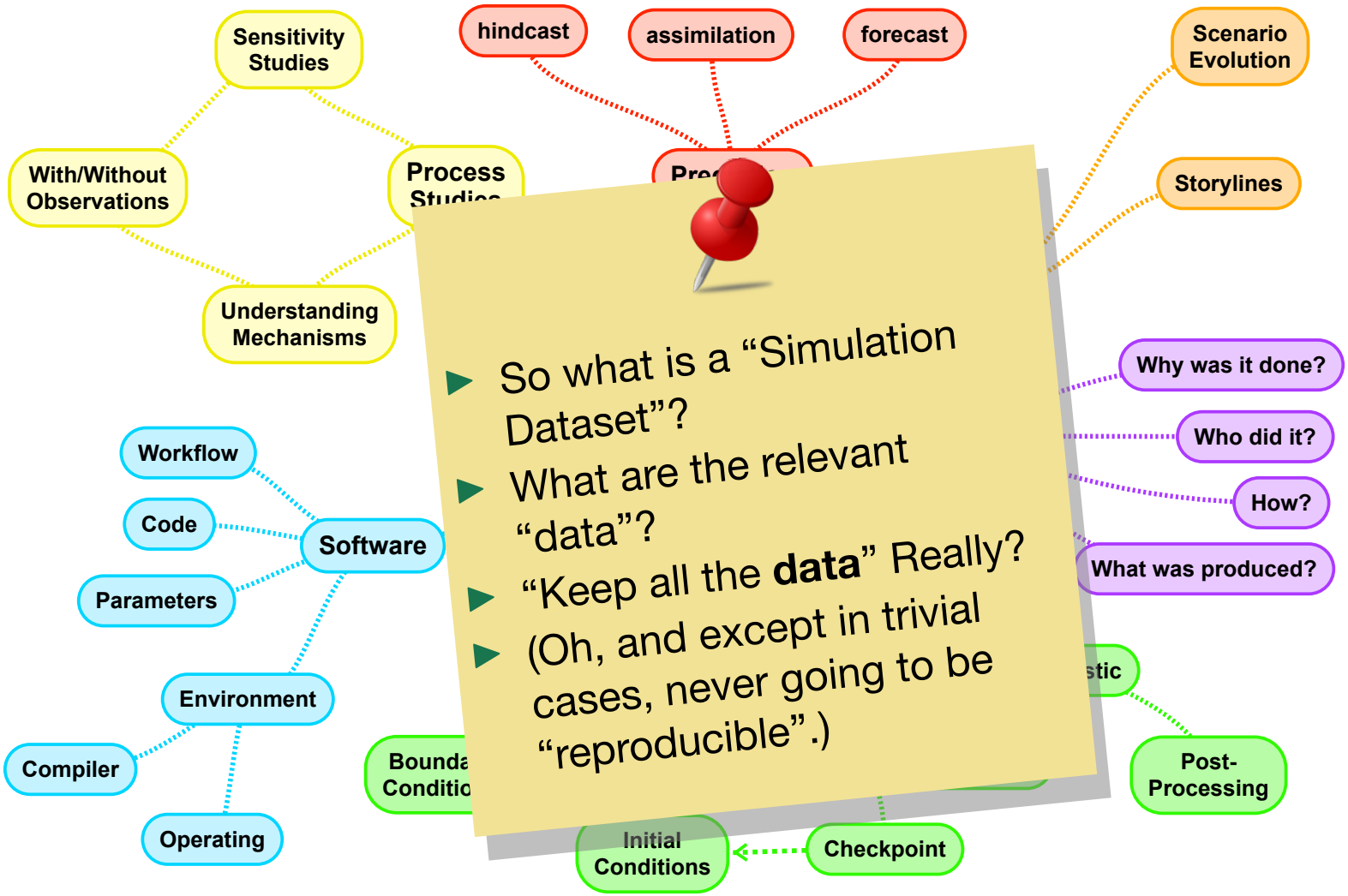
What needs to be FAIR?



What needs to be FAIR?

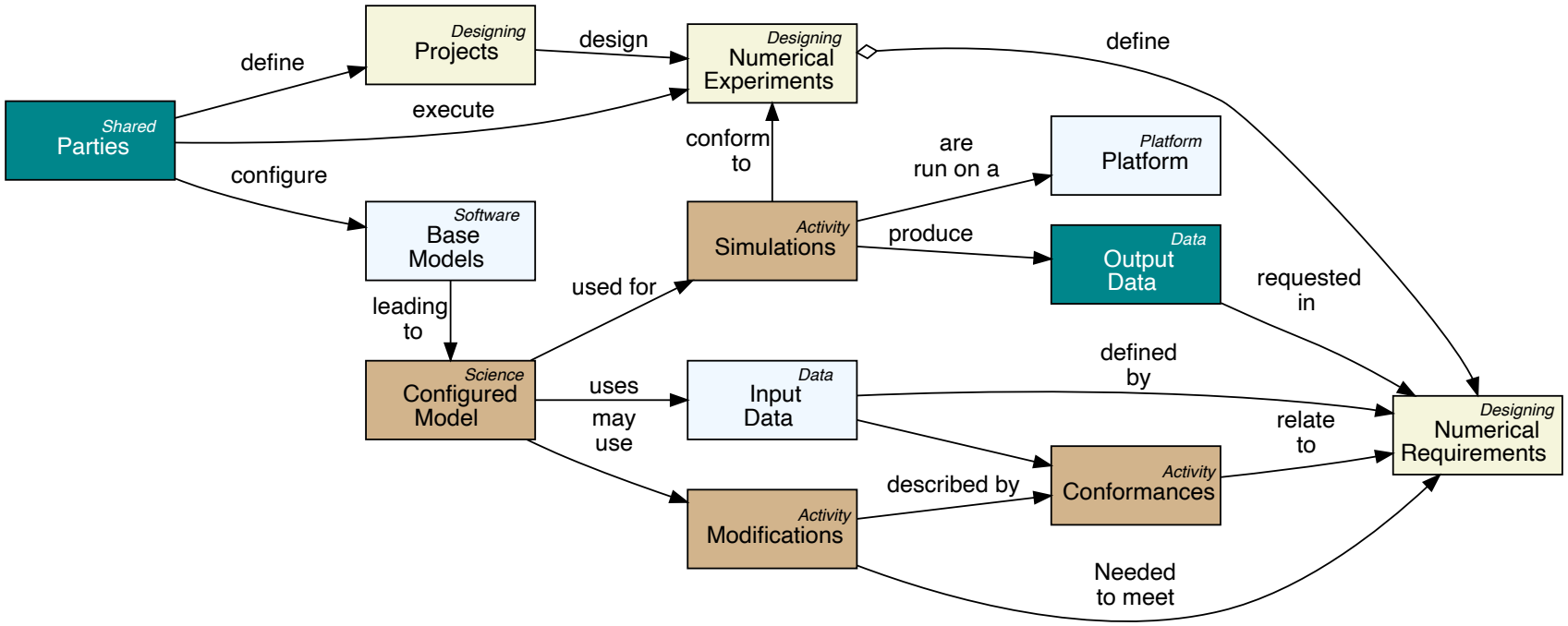


What needs to be FAIR?



- So what is a "Simulation Dataset"?
- What are the relevant "data"?
- "Keep all the **data**" Really?
- (Oh, and except in trivial cases, never going to be "reproducible".)

Simulation Context



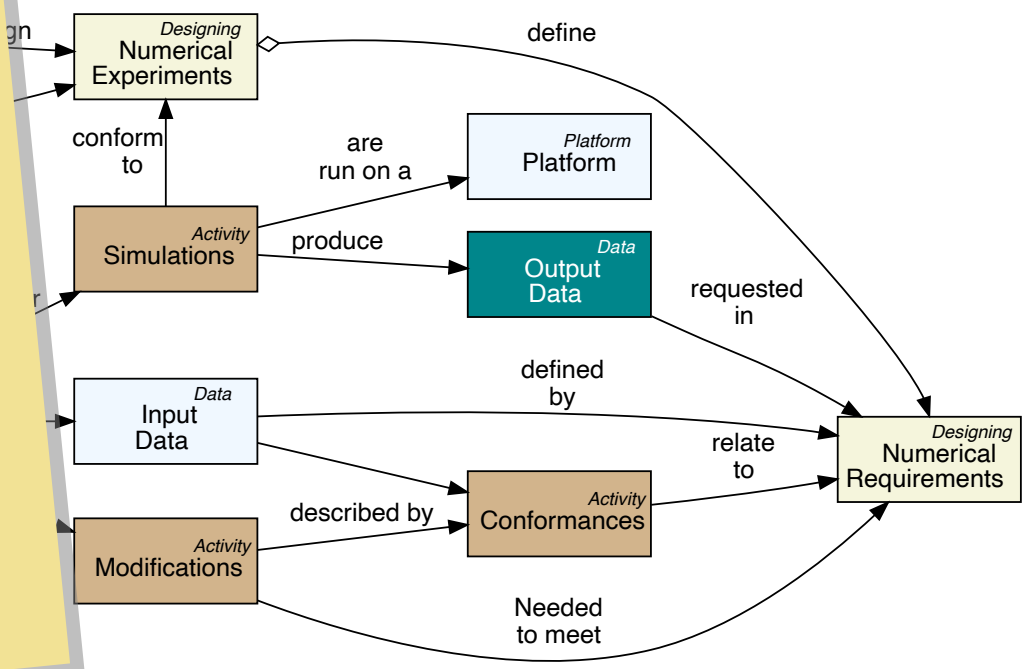
- Traditional Metadata covers only a tiny part of essential documentation!

Simulation Context

es-doc
Earth System Documentation

Why? What? How?

- Experiments may be devised by one group, and run by another, and analysed by a third.
- Participants are separated in space and time. Metadata crucial to information flow.



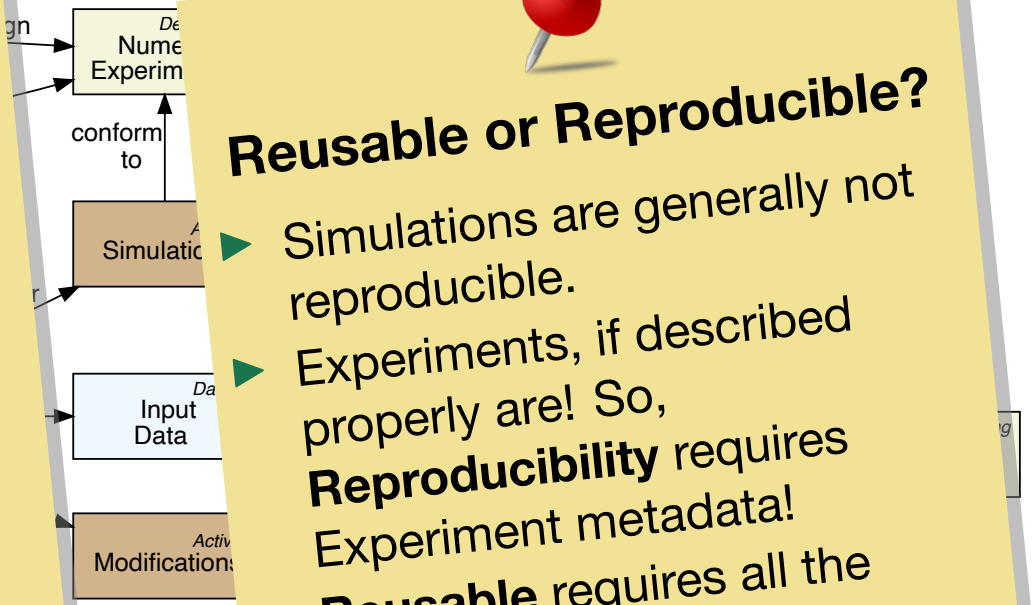
- Traditional Metadata covers only a tiny part of essential documentation!

Simulation Context

es-doc
Earth System Documentation

Why? What? How?

- Experiments may be devised by one group, and run by another, and analysed by a third.
- Participants are separated in space and time. Metadata crucial to information flow.



Reusable or Reproducible?

- Simulations are generally not reproducible.
- Experiments, if described properly are! So, **Reproducibility** requires Experiment metadata!
- **Reusable** requires all the metadata!

- Traditional Metadata covers only a tiny part of essential documentation!

ESDOC Packages

es-doc
Earth System Documentation

designing
- C domain requirements
- C ensemble requirement
- E ensemble types
- E experimental relationships
- C forcing constraint
- E forcing types
- C initialisation requirement
- C multi ensemble
- C numerical experiment
- C numerical requirement
- E numerical requirement scope
- C objective
- C output requirement
- C project
- C simulation plan
- C start date ensemble
- C temporal constraint

shared
- C doc meta info
- C doc reference
- C formal association
- C citation
- C extra attribute
- E nil reason
- C numeric
- C online resource
- C party
- C quality review
- E quality status
- C responsibility
- E role code
- C text blob
- E text blob encoding

iso
- C algorithm
- C lineage
- C process step
- C process step report
- C processing
- C quality evaluation output
- C quality evaluation result
- C quality issue
- C quality report
- E dq evaluation result type
- E ds initiative typecode
- E md cellgeometry code
- E md progress code

drs
- C drs atomic dataset
- C drs ensemble identifier
- C drs experiment
- E drs frequency types
- C drs geographical indicator
- E drs geographical operators
- C drs publication dataset
- E drs realms
- C drs simulation identifier
- C drs temporal identifier
- E drs time suffixes

platform
- C compute pool
- C interconnect
- C machine
- C nic
- C partition
- C performance
- C performance detail
- C project cost
- C storage pool
- E storage systems

software
- C component base
- C composition
- C development path
- C entry point
- C gridspec
- C implementation
- C software component
- C variable
- E coupling framework
- E programming language

cmip
- C cmip dataset
- C cmip simulation

data
- C dataset
- E dataset type
- C variable collection

science
- E model types
- C model
- C realm
- C realm coupling
- C topic
- C topic property
- C topic property set

time
- C calendar
- E calendar types
- C date time
- C datetime set
- C irregular dateset
- E period date types
- C regular timeset
- C time period
- E time units

activity
- C activity
- C axis member
- C child simulation
- C conformance
- E conformance type
- C ensemble
- C ensemble axis
- C simulation
- C uber ensemble

CIM 2.2.0.pre1
29/01/2020 09:05

See Lawrence et. al. 2020, in review (copy on request)

ESDOC Packages

es-doc
Earth System Documentation

designing
- C domain requirements
- C ensemble requirement
- E ensemble types
- E experimental relationships
- C forcing constraint
- E forcing types
- C initialisation requirement
- C multi ensemble
- C numerical experiment
- C numerical requirement
- E numerical requirement scope
- C objective
- C output requirement
- C project
- C simulation plan
- C start date ensemble
- C temporal constraint

shared
- C doc meta info
- C doc reference
- C formal association
- C citation
- C extra attribute
- E nil reason
- C numeric
- C online resource
- C party
- C quality review
- E quality status
- C responsibility
- E role code
- C text blob
- E text blob encoding

iso

platform
- pute pool
- connect
- ine
- on
- mance
- mance detail
- cost
- pool
- systems

software
- C component base
- C composition
- C development path
- C entry point
- C gridspec
- C implementation
- C software component
- C variable
- E coupling framework
- E programming language

cmip
- C cmip dataset
- C cmip simulation

me
- types
- et
- ateset
- types
- C regular timeset
- C time period
- E time units

activity
- C activity
- C axis member
- C child simulation
- C conformance
- E conformance type
- C ensemble
- C ensemble axis
- C simulation
- C uber ensemble

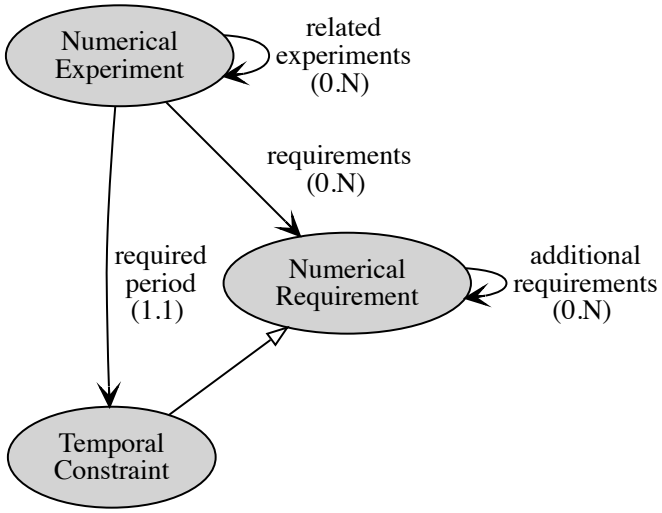
C topic property set

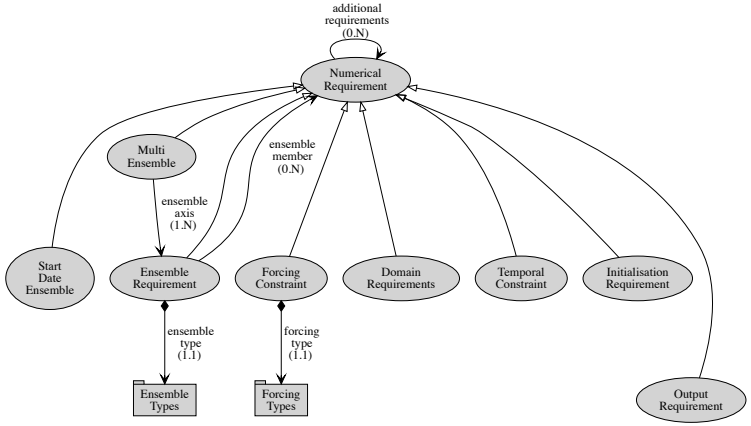
CIM 2.2.0.pre1
29/01/2020 09:05

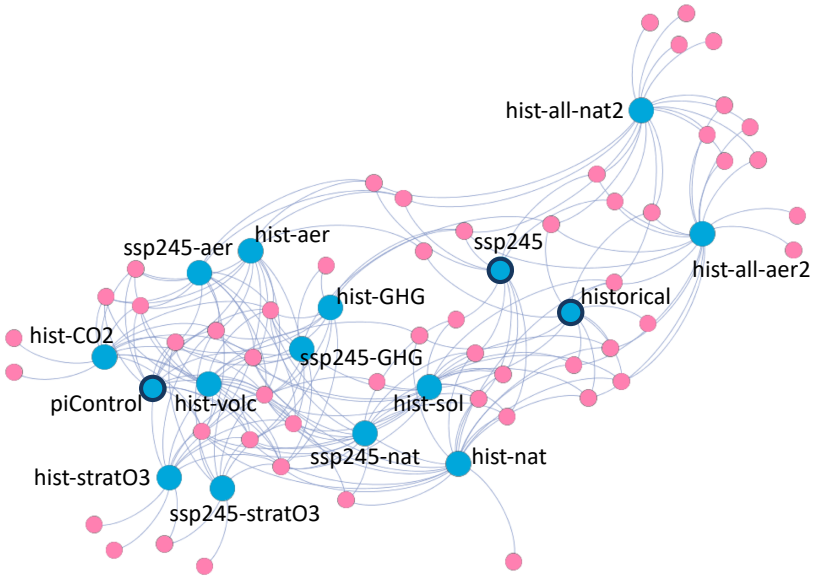
- All packages defined in Python
- Python libraries for creating and manipulating metadata
- RDF export available.
- Extensible.
- (Open. Under active development on GitHub)

See Lawrence et. al. 2020, in review (copy on request)

Metadata Bonus!

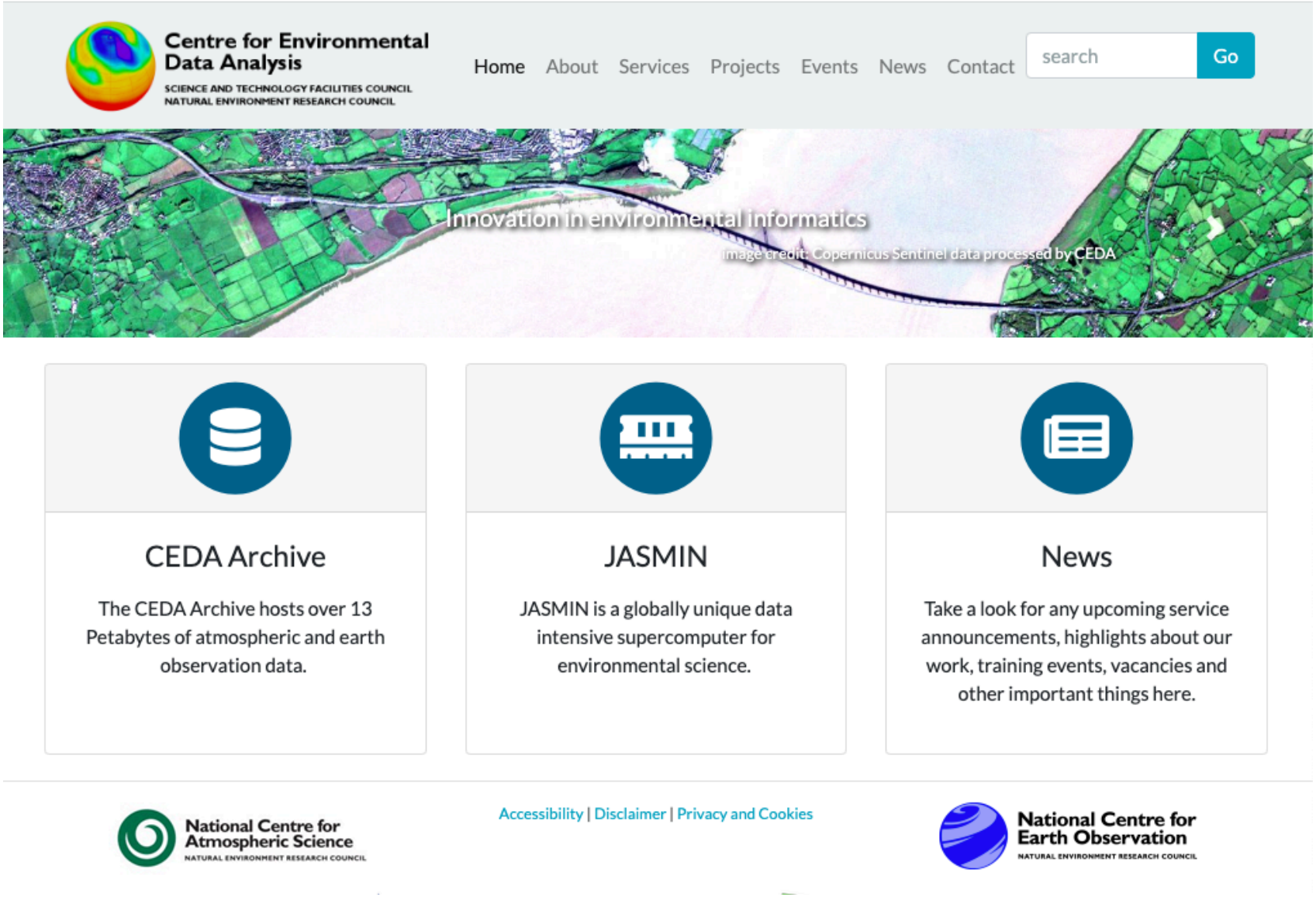




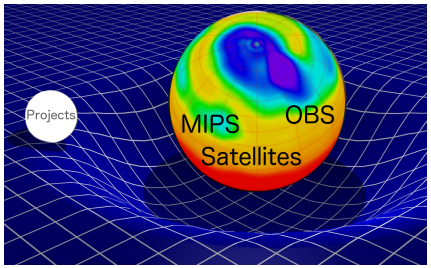


Metadata triples expose explicit relationships between experiments in CMIP6 based on forcing constraints.
(Pascoe et al 2020, GMD, to appear.)

CEDA

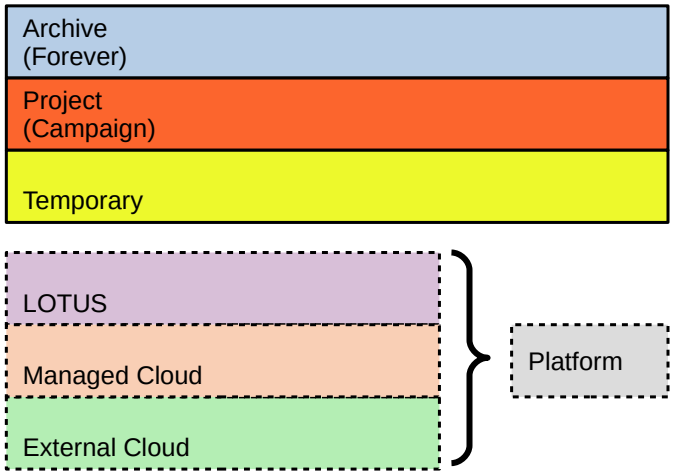

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL
Home About Services Projects Events News Contact
search
Go
Innovation in environmental informatics
Image credit: Copernicus Sentinel data processed by CEDA
CEDA Archive
The CEDA Archive hosts over 13 Petabytes of atmospheric and earth observation data.
JASMIN
JASMIN is a globally unique data intensive supercomputer for environmental science.
News
Take a look for any upcoming service announcements, highlights about our work, training events, vacancies and other important things here.
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL
Accessibility | Disclaimer | Privacy and Cookies
National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

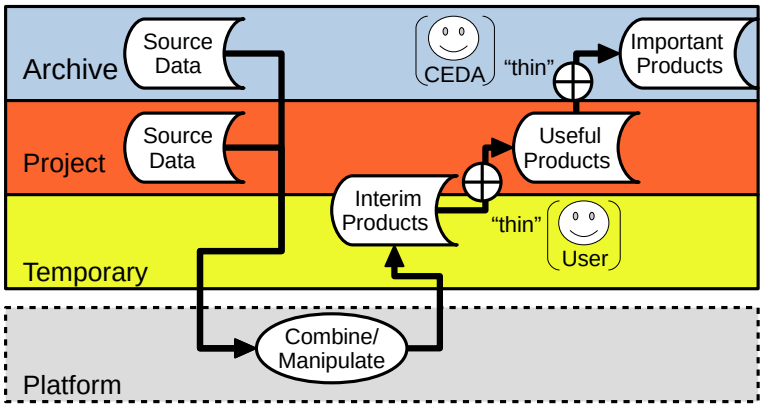
JASMIN: Data Commons and Current Curation Context



- CEDA Archive is curated with potentially centennial scale persistence (with data review).
- Project disk organised into shared "Group Work Spaces" (GWS).

- Simulations produced elsewhere and transferred to GWS, and in very special cases (CMIP) ingested directly into archive.
- Users have disk and can persist data on tape, but this is different from the archive, which is persisted on disk and on tape. What simulation data should go into the archive?





Analysis Workflow

CEDA Policy I

https://help.ceda.ac.uk/article/4300-archiving-of-simulations-guide

(From c 2005, the days of the BADC, definitely needs updating!)

Context

- Simulations are generally, but not always, analogues of the "real" world that may provide insights on physical causal relationships.
- Where simulations represent predictions of the real world or where they incorporate real measurements to improve estimates of the state of the real world (e.g. assimilation products) their wider value (in the long term, or to a larger community) is enhanced
- Where simulations have more confusing relationships with the real world (as would be the case with "sensitivity" experiments where either the boundary conditions or the relations within the model are idealised), their wider value is less obvious.

CEDA Policy II

Key Insights:

- Some things are obviously important (but we still may not be able to afford them).
- Some things are obviously not important.
- The difficulty is in between. In between it is a matter of choice and economics. The choice is a function of “Does the provider want to do the metadata generation?” and “Does the curator see a potential community of users?”

CEDA Policy II

These led to three sets of questions:

- If the answer to one or more of the following questions (on later slides) is yes, then simulated data are candidates for professional data management beyond that provided by the investigating team responsible for producing the data.
- If the answer to any of the following questions (on later slides) is yes, then the simulated data should not be archived, but could still be candidates for data management to aid exploitation within a larger project.
- If the answer to any of the following questions (on later slides) is yes, then value judgements will need to be made about how much, if any, of the simulated data should be archived.

Questions that suggest simulations should be curated

- Is there — or is there likely to be in the future — a community of potential users who might use the data without having one of the original team involved as co-investigators (or authors)?
- Does some particular simulation have some historical, legal or scientific importance that is likely to persist? (Some simulations may become landmarks, in some way, along the route of scientific knowledge. They may also have been quoted to make a statement that might be challenged — either scientifically or legally – and should therefore be kept for evidential reasons.)
- Is the management of the data by a project team likely to be too onerous for them or result in duplication of effort with other NERC funded activities?
- Is it likely that the simulation will be included in future inter-comparisons for which NERC funding will be sought?
- Does the simulation integrate observational data in a manner that adds value to the observations?

Questions that suggest simulations should be curated

- Is there — or is there likely to be in the future — a community of potential users who might use the data without having one of the original team involved as co-investigators (or authors)?
- Does some particular simulation have some historical, legal or scientific importance that is likely to persist? (Some simulations may become landmarks, in some w knowledge. They may also have been c t might be challenged — either s uld therefore be kept for evidential rea
- Is the management of y to be too onerous for them or re other NERC funded activities?
- Is it likely that the simulation will be included in future inter-comparisons for which NERC funding will be sought?
- Does the simulation integrate observational data in a manner that adds value to the observations?

The obvious missing question:

"Can this be afforded?"

Questions which suggest curation can/should be avoided

- Is the data produced by a trivial algorithm that could be easily regenerated from a published algorithm description?
- Is the data unlikely to ever be used in a peer-reviewed publication, or as evidence to support any public assertions about the environment?
- Is the data known to be of poor quality or to have little scientific validity?
- Is it impossible to adequately document the methodology used to produce the data in a way that is accessible to users of the data outside the producing team?
- Is the simulated data produced in a sensitivity experiment rather than as a predictive or retrospective analysis of a real system?
- Is the data likely to be of short-term use, and in the case of loss, more easily (in terms of physical and financial effort) replaceable by rerunning the simulation

Questions which suggest curation can/should be avoided

- Is the data produced by a trivial algorithm [illegible] be easily regenerated from a p[illegible]
- Is the data unlikely to [illegible] d publication, or as evidence to support a[illegible] environment?
- Is the data known to b[illegible] scientific validity?
- Is it impossible to ade[illegible] ogy used to produce the data in a [illegible] f the data outside the producing team?
- Is the simulated data p[illegible] ent rather than as a predictive or retrospective analysis of a real system?
- Is the data likely to be of short-term use, and in the case of loss, more easily (in terms of physical and financial effort) replaceable by rerunning the simulation

"Is it impossible …" and by implication:

"Is it likely adequate documentation will not be made available?"

Value Judgements

The following value judgements apply to both the middle-ground, and the datasets that have passed the "candidates for" criteria:

- Would storage of the data be prohibitively expensive?
- Would storage of statistical summaries rather than individual data items provide adequate evidential information about the simulation? (e.g. while it might normally be desirable to store all ensemble members, would ensemble and/or temporal means be adequate in a situation where storage of the individual members at full time resolution might be prohibitively expensive).
- Would historical preservation be satisfied by archiving only the data which supported published figures, or is future use likely to include data processing?

Other aspects of the policy

Other aspects of the policy include:

1. Guidelines for how to archive simulation dataset
   - These guidelines have informed the development of ES-DOC, although ES-DOC is not yet mandatory for all simulation data at CEDA, although that is the aim.
2. Guidelines for how to assess expected curation lifetimes
   - (Although we have never removed data from the archive, and experience at large sites suggest this is not cost-effective, at least with current tape cost trajectories.)
3. Custodial responsibilities
   - (Mainly to address the situation where the data is not being held at a designated data centre — aka repository.)

The entire policy is somewhat dated and needs review, but it has stood the test of time rather well.

Summary

1. Not all simulations can or should be FAIR.
2. Unless all the simulation workflow can be properly documented, the data produced is not really re-usable across time.
   - …and if communities are not willing to document their simulations, expensive national resources should not be made available for their support!
3. The ES-DOC metadata system has been developed to address the needs of simulation documentation (not just CMIP projects).
4. Simulations are not really reproducible, experiments are!
5. Policy about curation and FAIRness depends on clear guidelines, economics, and the willing to create metadata.
6. It is easier to decide on what not to curate, than what can be curated.

Acknowledgements

- NERC: the UKRI Natural Environment Research Council has funded a network of designated data centres for over two decades, including the British Atmospheric Data Centre and the NERC Earth Observation Centre, which amalgamated to become the Centre for Environmental Data Analysis.
- The original CEDA policy was written by myself with Anne De Rudder, Jamie Kettleborough, and Kevin Marsh.
- ES-DOC (previously Metafor) has been supported by a number of US and European projects as well as institutional funding (particularly from NCAS and IPSL).
  - The current work is being supported by NCAS, IPSL, and the IS-ENES3 project with funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084.