Managing Data and Facilitating Science

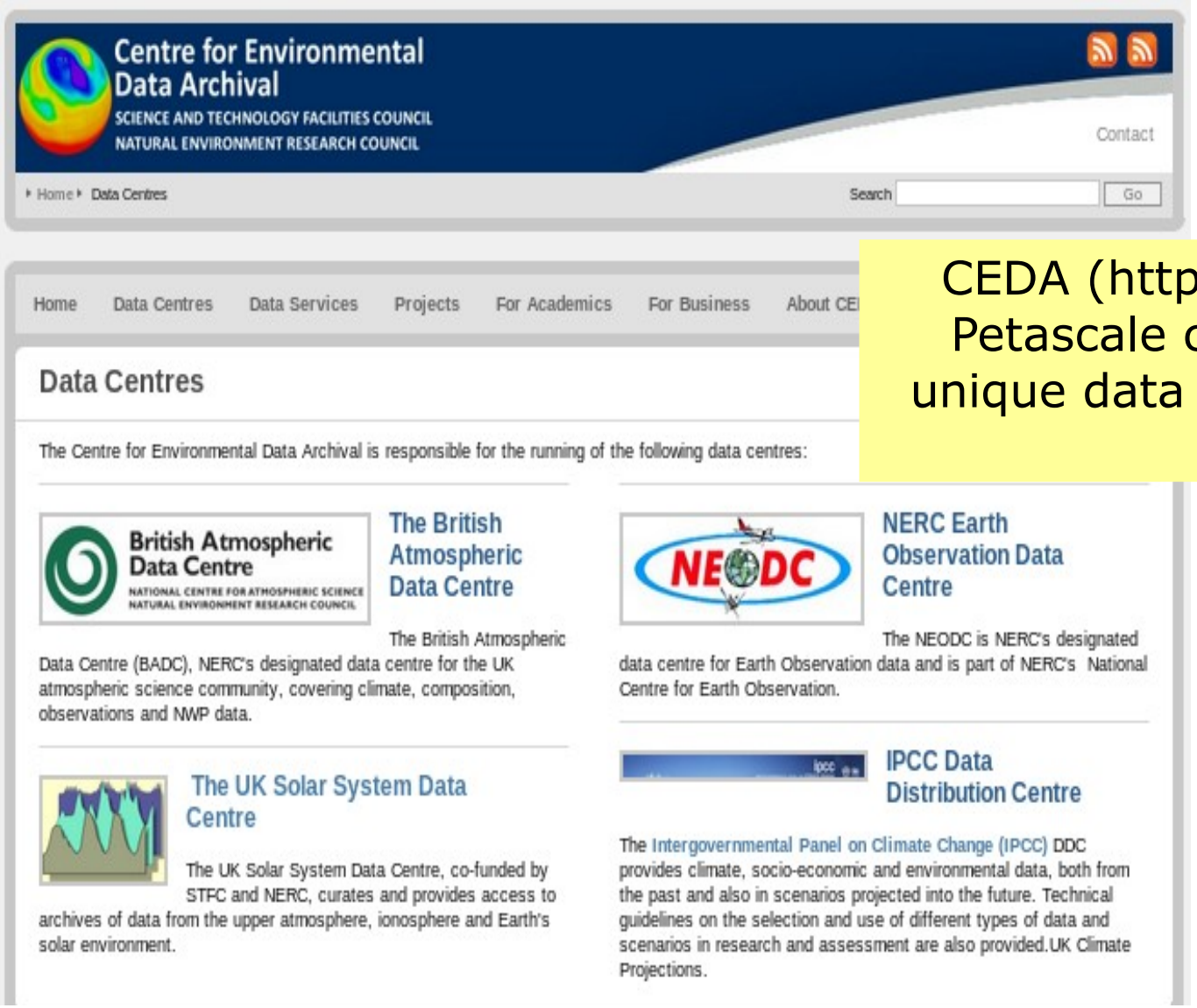
A spectrum of activities in the
Centre for Environmental Data Archival
(CEDA)

Bryan Lawrence

(NCAS, University of Reading, STFC/CEDA)

Victoria Bennett, Sarah Callaghan, Martin Juckes, Sam Pepler and the CEDA Team

What is CEDA?



CEDA (http://www.ceda.ac.uk)
Petascale curated archive with unique data analysis environment (JASMIN)

Context

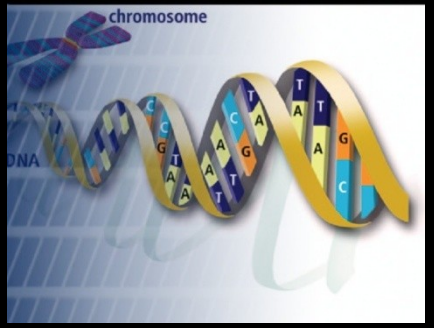
Hydrology:
National Water Archive

Atmosphere:
British Atmospheric Data Centre

Earth observation:
NERC Earth Observation Data Centre

Ocean & marine:
British Oceanographic Data Centre

Bioinformatics:
NERC Environmental Bioinformatics Centre



Earth:
National Geoscience Data Centre

Terrestrial & freshwater:
Environmental Information Centre

Polar:
Antarctic Environmental Data Centre

NERC
SCIENCE OF THE ENVIRONMENT

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

Why is CEDA?

NERC Data Policy

- Ensure the **continuing availability** of environmental data of long-term value for **research, teaching, and for wider exploitation** for the public good, by individuals, government, business and other organisations.
- Support the **integrity, transparency** and **openness** of the research it supports.
- Help in the **formal publication of data sets**, as well as enabling the tracking of their usage to be tracked through citation and data licences.
- **Meet relevant legislation** and government guidance on the management and distribution of environmental information.

Difference between Preservation and Curation

Preservation:

The Phaistos Disk

1700 BC

Preserved, but information content is zero!

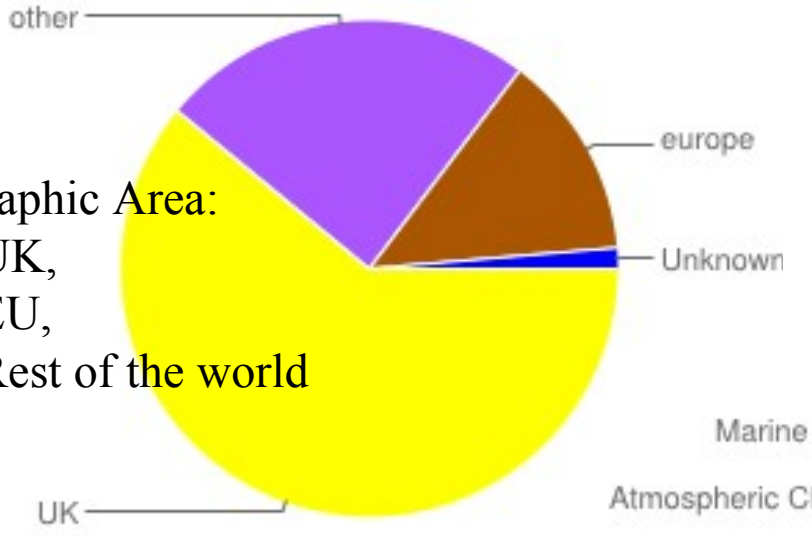
Digital curation: entails (Wikipedia, 29/04/12)

- Collecting *(CEDA: Ingestion)*
- Providing search and retrieval *(Services)*
- Certification of the trustworthiness and integrity *(Documentation/metadata/provenance)*
- Semantic and ontological continuity *(An active process of information migration.)*

Who uses CEDA?
(Consumer Perspective)

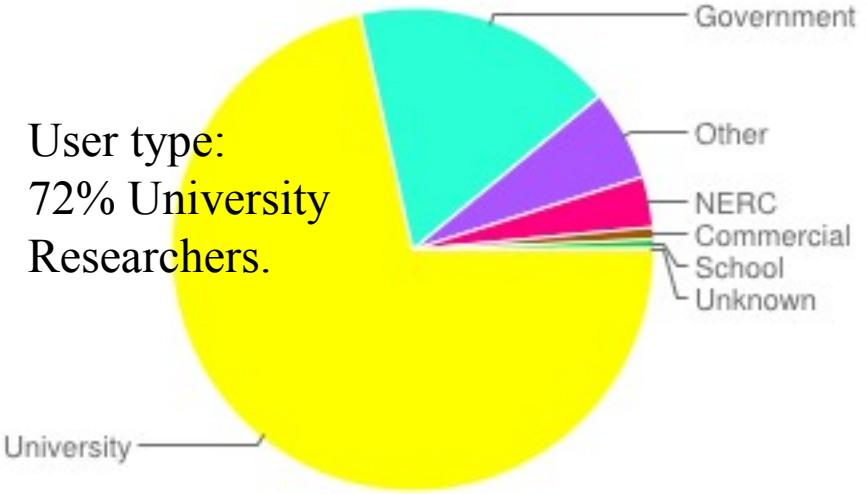
Break down of 3713 users registered for specific CEDA data or services.

We don't have details for the **other 14,000** users!
(April 2012.)

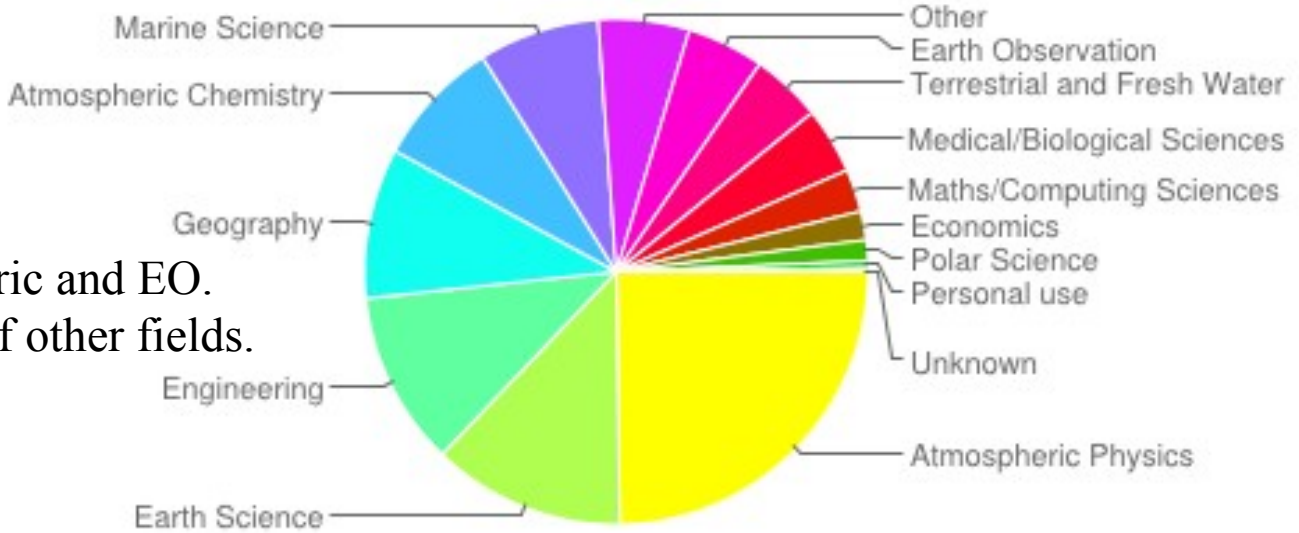


Geographic Area:
61% UK,
13% EU,
24% Rest of the world

User type:
72% University Researchers.



Discipline:
38% Atmospheric and EO.
Full spectrum of other fields.



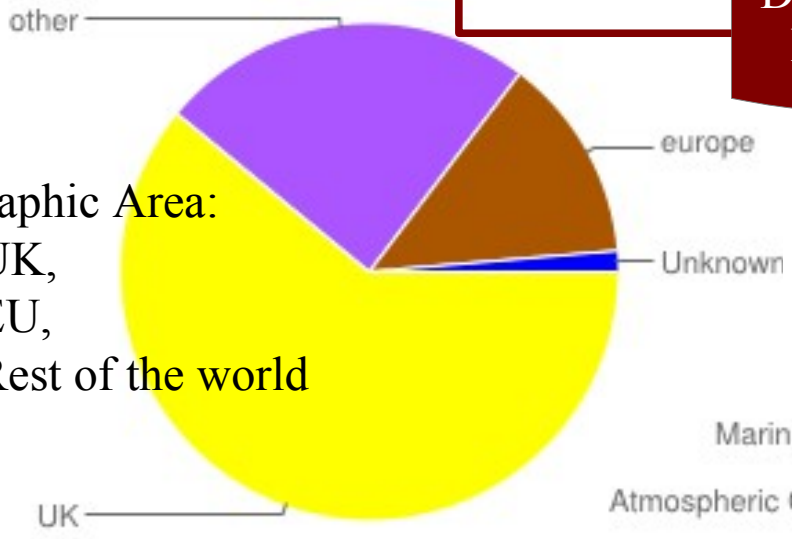
Who uses CEDA?

(Consumer Perspective)

Break down of 3713 users registered for specific CEDA data or services.

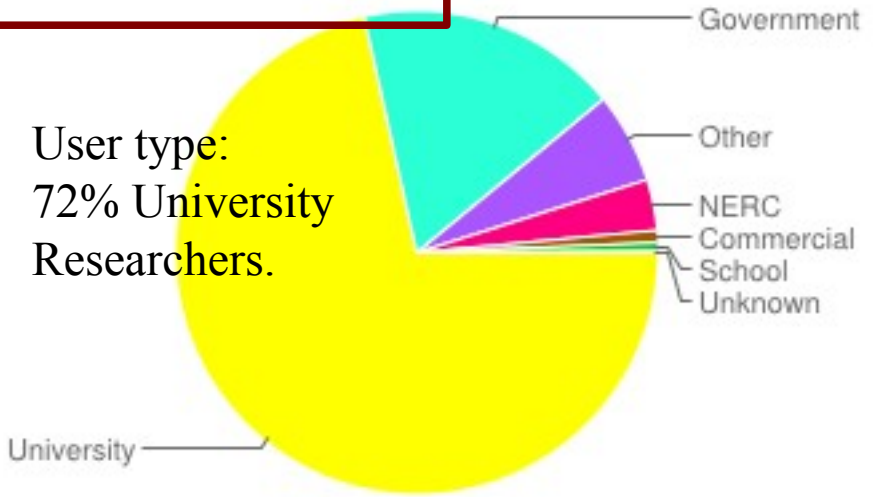
We don't have details for the **other 14,000** users! (April 2012.)

Download Mainly

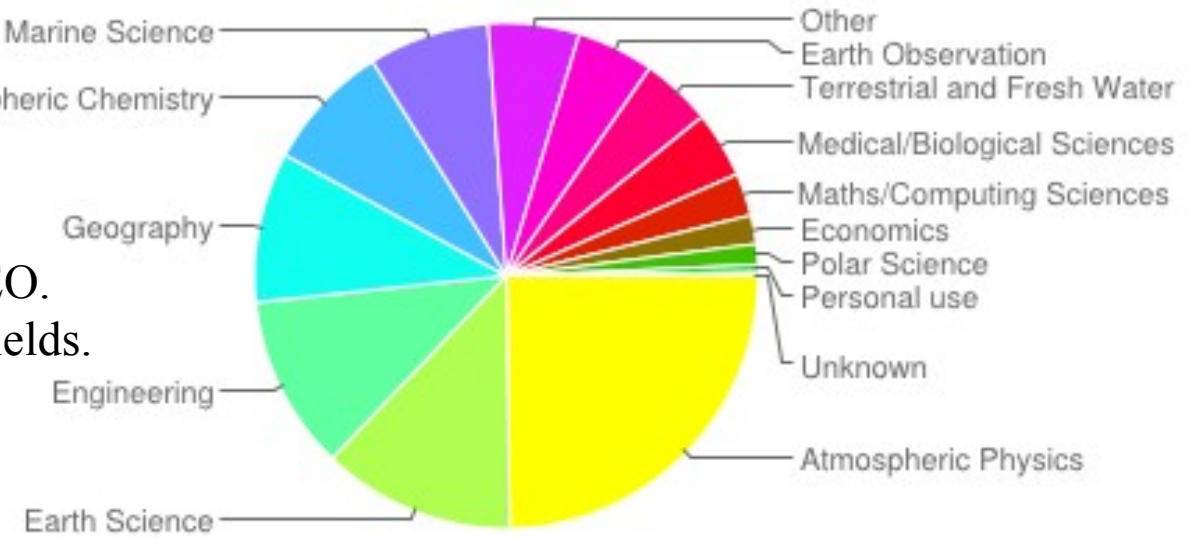


Geographic Area:
61% UK,
13% EU,
24% Rest of the world

User type:
72% University Researchers.



Discipline:
38% Atmospheric and EO.
Full spectrum of other fields.



Science & Technology Facilities Council

BADC: Supporting Aircraft Campaigns

Supporting both UK and European science aircraft databases

(FAAM, ARSF, EUFAR)

Home | My BADC | Data | Search | Community | Help

Get Data | Access Rules | Submit Data | Dataset Index

Application for access to Data from FAAM Campaigns

Last update: 10 June 2013

- The table below lists the FAAM campaigns by chronological order and, for each of them, defines the communities entitled to access the archived non-core and processed core data. Authorised users can apply by clicking on the relevant APPLY link in the table. Application for public processed core data must be done only once. No application is needed to access public non-core data.
- Access to FAAM test flight data as well as FAAM raw core data from field campaigns is restricted to the staff involved in test flights (APPLY here).
- To apply for access to any of these datasets, you should first register as a BADC user (anyone is allowed to register). If you are already a registered BADC user, skip this step. If you have forgotten your BADC user ID and/or password, please contact the BADC helpdesk at badc@rl.ac.uk.
- Both the FAAM facility and FAAM data should be properly referenced in any publications - see the citation section of the useful information page for details.

Campaign (Click on campaign name to view web page)	Dates (Click on Year number to view full schedule)	Aircraft base	Flight numbers	Access to processed core data	Access to non-core data
ITOP-UK, (UTLS Ozone project)	12 Jul - 3 Aug 2004	Horta, Azores	B028-39	Public, on application – APPLY (*)	Public
ADRIEX	27 Aug - 6 Sep 2004	Treviso, Italy	B043-50	Public, on application – APPLY (*)	Public
EAQUATE	14-18 Sep 2004	Cranfield, UK	B051-54	Public on application – APPLY (*)	Public
AUTEX-WINTEX	27 Oct 2004 - 11 Feb 2005	Cranfield, UK	B056, B058, B060, B062-63, B065-68, B084-86, B251-267	Public on application – APPLY (*)	Public
NU-WAVE	17-24 Nov 2004	Cranfield, UK	B059, B061	Public on application – APPLY (*)	Public

Home | My BADC | Data | Search | Community | Help

Get Data | Access Rules | Submit Data | Dataset Index

Get Data

Login Help

Username: *Not logged in*

Download multiple files How to use * Depth: 1 GO

Current directory: / badc / eufar / data / aircraft

Dataset: *Combined Data from the Projects Funded Under The European Facility for Airborne Research in Environmental and Geo-sciences (EUFAR) Project* Details

awi-polar5	*POLAR 5 (Basler BT-67)- Operated by: AWI (Alfred Wegener Institute)*
dlr-do228-101	*Dornier DO228 - 101 - Operated by: DLR (Deutsches Zentrum fur Luft- und Raumfahrt e.V.)*
dlr-do228-dcffu	*Dornier 228-212/1 - Operated by: DLR (Deutsches Zentrum fur Luft- und Raumfahrt e.V.)*
dlr-g550-halo	*G550 HALO (Gulfstream) - Operated by: DLR (Deutsches Zentrum fur Luft- und Raumfahrt e.V.)*
enviscope-learjet	*Learjet 35A - Operated by: Enviscope GmbH*
enviscope-partenavia	*Partenavia P68B - Operated by: Enviscope GmbH*
faam-bae146	*BAe146-300 - Operated by: FAAM (Facility for Airborne Atmospheric Measurements)*
fub-ask16	*ASK16 Motorglider - Operated by: FUB (Freie Universitat Berlin, Institut fur Weltraumwissenschaften)*
fub-c207	*Cessna T207A - Operated by: FUB (Freie Universitat Berlin, Institut fur Weltraumwissenschaften)*
fzk-enduro	*ENDURO - Operated by: FZK (Forschungszentrum Karlsruhe)*
gtk-3in1twinotter	*3in1Twin Otter - Operated by: GTK (Geological Survey of Finland)*
ibimet-skyarrow	*SkyArrow 650 TCNS - Operated by: IBIMET (CNR - Institute of Biometeorology)*

BADC: ENES, GO-ESSP and CMIP5

Science Driver: CMIP5

Fifth Coupled Climate Model Intercomparison Project.

Petascale globally distributed data (2 PB, 600K datas, 4.5M files)

Partnerships

European Network for Earth Simulation (ENES, IS-ENES project)

Global Organisation for Earth System Science Portals (GO-ESSP, ESGF project)

Information Challenge

Data Documentation,

Workflow (versioning, replication)

User Management (Access control)

Infrastructure Challenge

Network bandwidth (incoming and outgoing)

Storage (volume and performance)

Analysis Environment?

Science & Technology Facilities Council

JASMIN:Joint Analysis System

J is for Joint

Jointly *delivered* by

RALSpace (CEDA) and SCD.

Joint *users* (initially):

NERC aca community & Met Office

Joint *users* (target):

Industry (data users & service providers)

Europe (wider environ. academia)

A is for Analysis

Private (Data) Cloud

Compute Service

Web Service Provision

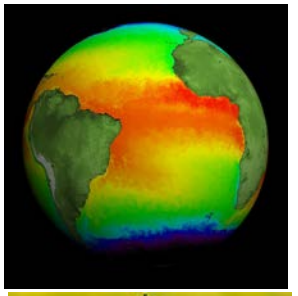
For

Atmospheric Science

Earth Observation

Environmental Genomics

… and more.



S is for System

£10m investment

at RAL

#1 in the world for big data analysis capability?

Opportunities

JASMIN is a collaboration platform!

within NERC (who are the main investor)

between communities (Space and Climate via CEMS)

with industry (cloud providers, SMEs)

Across Europe (ENES etc)

(CEMS:the facility for Climate and Environmental Monitoring from Space)

The main (day) jobs at CEDA have and continue to be:

(1) to ingest and deliver data, and

(2) **to collect and manage information about that data**

(3) Provide data services, and

(4) Facilitate Science Programmes

(from small, e.g. one PI, to large, e.g. CMIP5)

There is a new job: To provide an environment to manipulate data, both the data in the managed archive, and user "managed" data.

Data Infrastructure Depends on Information Infrastructure

We're dealing with data for and in three domains:

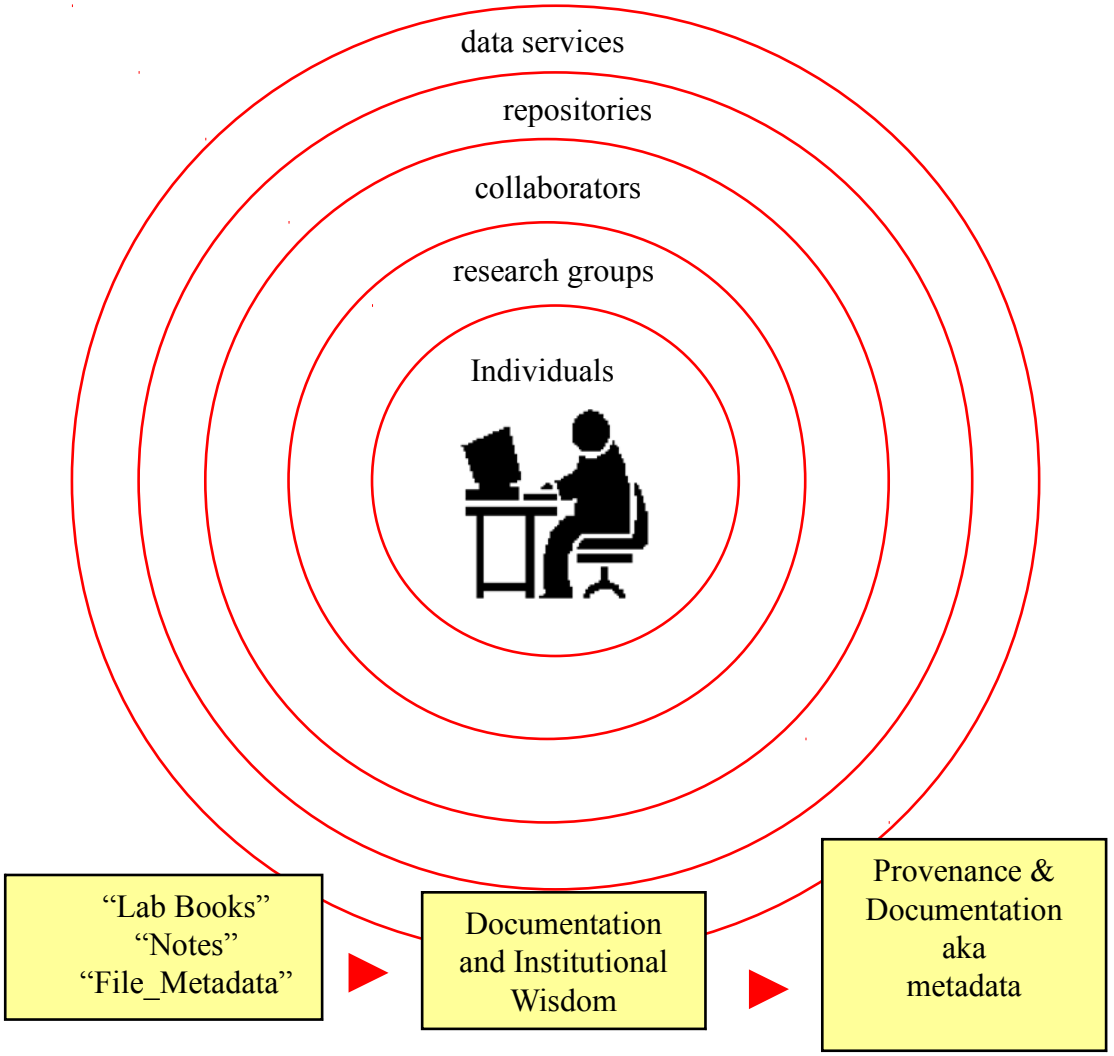
- » Institutional (Science)
- » Federated (Science, but different communities)
- » Shared (Services)

Between each we have collaborations, formal or informal, and between each, we have more and more layers of information, conventions and common behaviours!

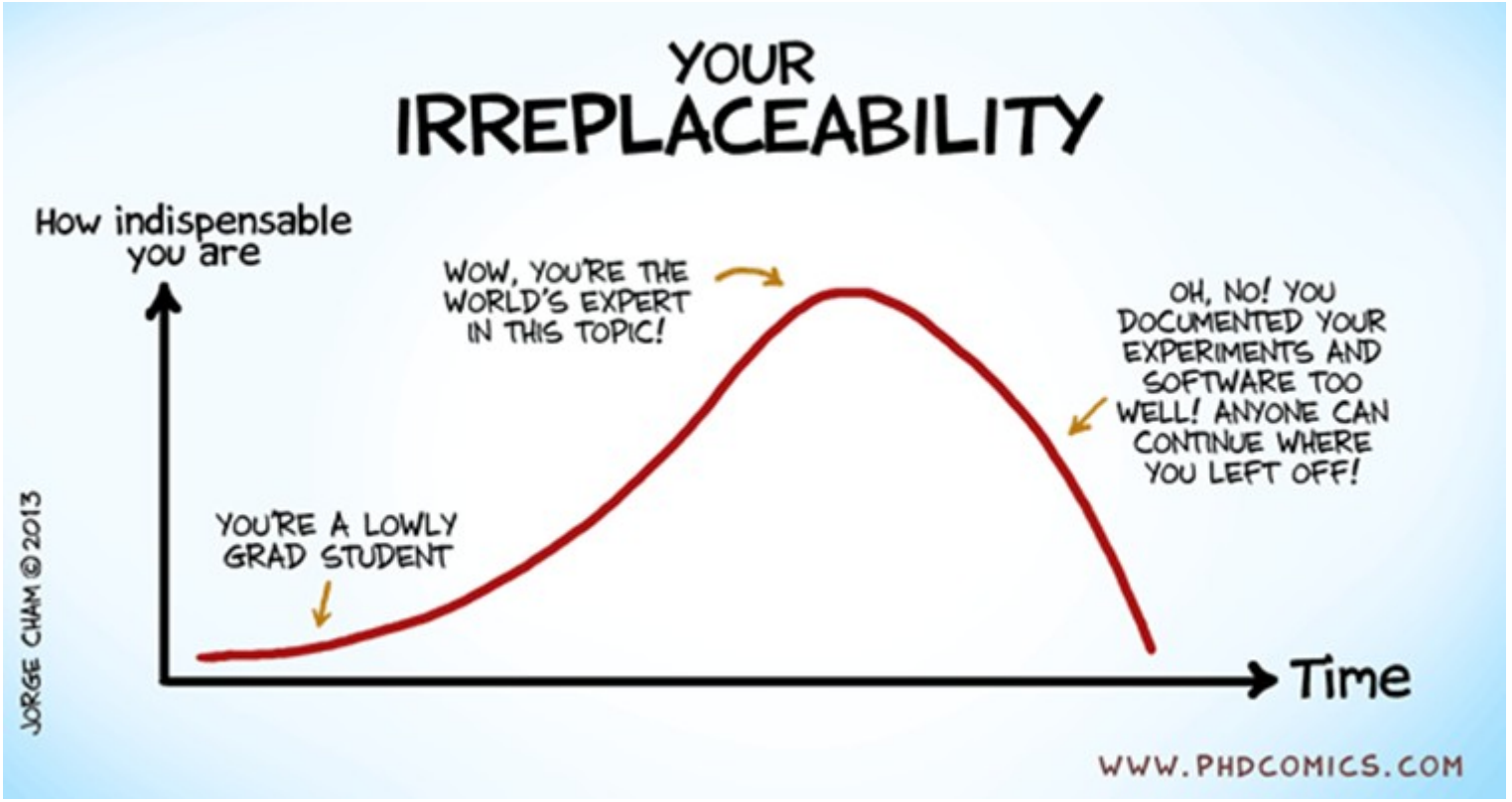
The hard part is that the producer on the inside doesn't always know (or want to produce) what the next layer up wants or needs.

Crucial to remember that in this environment "if it isn't in the metadata" the outer layers "WONT KNOW IT"!

- » Can't rely on institutional wisdom and/or notes on portals which don't flow outwards!

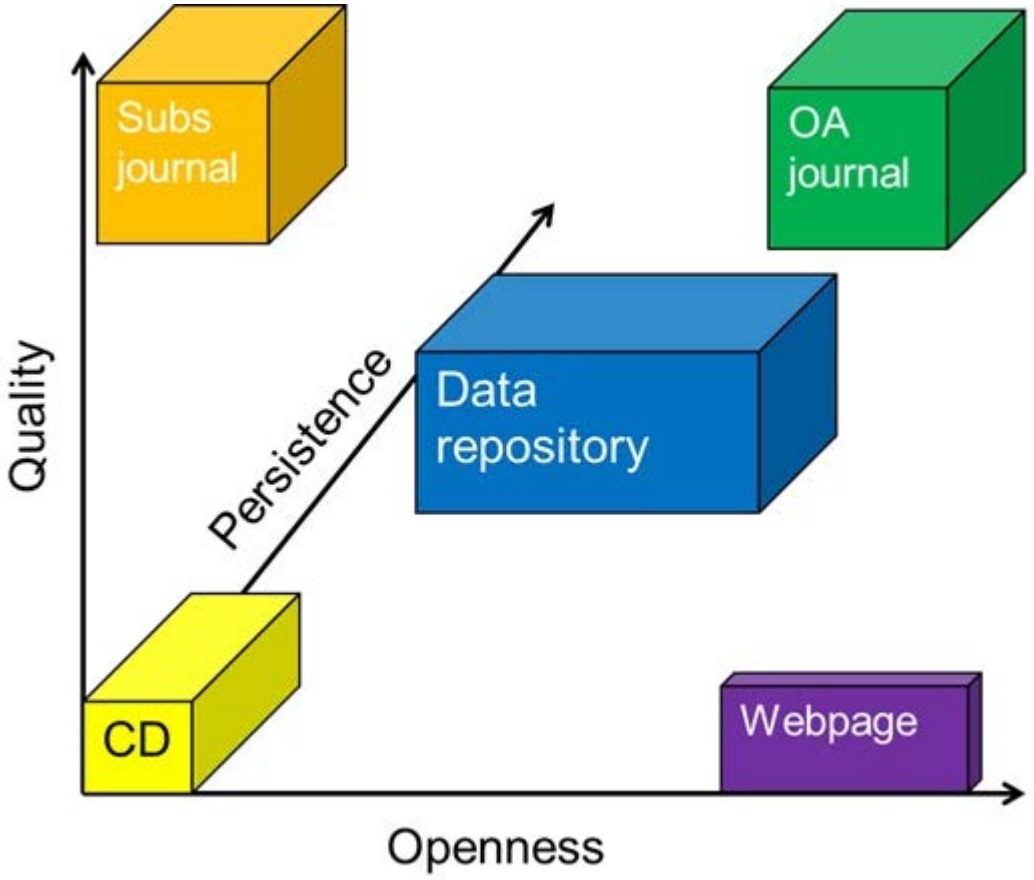


Documentation can be problematic



Data Publication

Publishing = making something public after some formal process which adds value for the consumer (e.g. peer review) and provides commitment to persistence



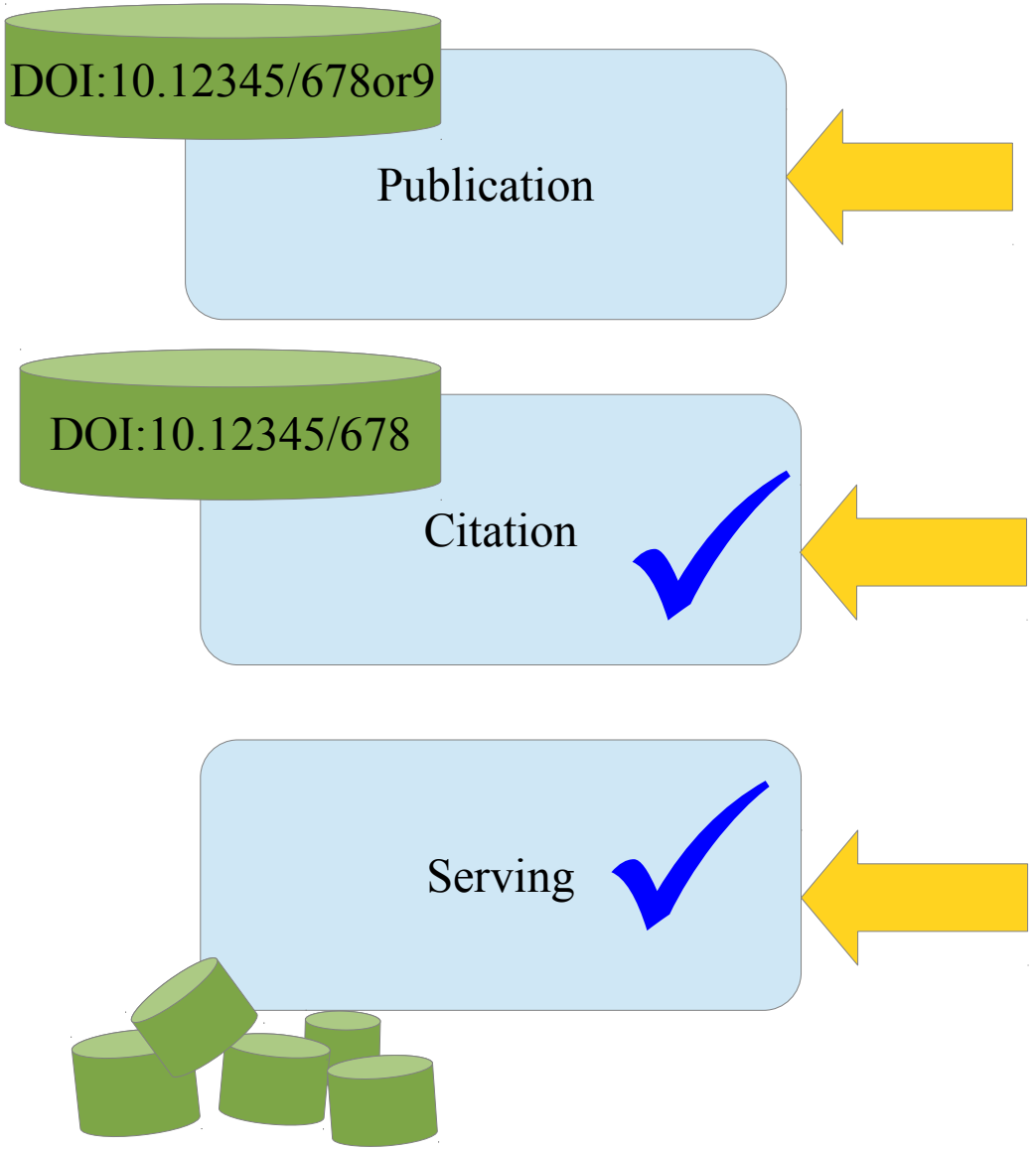
We want to encourage researchers to make their data:

- Open
- Persistent
- Quality assured:
  - through "scientific peer review" via a journal publication
  - or repository-managed processes

Unless there's a very good reason not to!

Why? BECAUSE THEN THEY TAKE METADATA SERIOUSLY? (AND THEY GET CREDIT!)

Peer Review



Scientific Review
Needs domain expertise
Evaluates the "quality" of the dataset
Partnerships with academic journal publishers

Technical Review
Checks on data format, use of standards, completeness of metadata.
Evaluates the "usability" of the dataset.
Freezes the dataset to make it Publishable.

The Day Job
We take data in, collect and create metadata, archive and manage the data,
and give them out again to other users.
Data is active – updated, modified, added to.

Rewards

For the Data Producer

- Attribution and credit for the hard work of creating the data and putting it into the archive with supporting metadata
- Citation counts used to evaluate impact of data - support career development
- Improved discoverability = increased opportunity for collaboration/more funding

For everyone else

- Reassurance that money isn't being wasted recreating data that already exists
- Transparency and accountability - supports and maintains the scientific record
- Improves usability and discoverability of data

… but using Publication as a tool is just one of the tools in our toolbox!

Parsons and Fox:
Is Data Publication the Right Metaphor?

Data Science, 2013

… describe several metaphors for managing, sharing, and stewarding data:

- Data Publication, Big Iron, Science Support, Map Making, and Linked Data.
- P&F address strengths and weaknesses, but particularly question the applicability of a "publication" approach to making data broadly available, but note

(Making one wonder who claimed "data publication" was the "right" or only one ...)

- "No one metaphor satisfies enough key data system attributes and that multiple metaphors need to co-exist in support of a healthy data ecosystem"

The problem we (CEDA) have with P&F is that we recognise all the words, but the analysis doesn't work for us ...

Metaphors: Publication

What we do	Parsons and Fox
Datasets are the result of discrete processes (e.g. experimental campaigns) though may be dynamic (added/modified as time passes).	Define discrete, well-described data sets
Technical review carried out in-house by science support staff. Scientific peer review carried out (or not) as part of the journal article publication process	Ideally with a certain level of quality assurance or peer-review.
Datasets are the fundamental foundation of the research article and the basis for the published conclusions.	The data sets often provide the basis for figures and tables in research articles and other publications.
Datasets exist and are used without having associated peer-reviewed journal articles. A dataset can and should be published in its own right (cf publishing negative results)	Published data are then considered to be first-class, reference-able, scientific artefacts, and they are often closely associated with peer-reviewed journal articles.

Metaphors: Big Iron

What we do	Parsons and Fox
CMIP5 (2 PB); UPSCALE (one model campaign ~ 0.5 PB)	Typically deals with massive volumes of data that are relatively homogenous and well defined but highly dynamic and with high throughput.
Now our own dedicated “HPD” environment.	Is a large, sophisticated, well-controlled, technical infrastructure potentially involving supercomputing centers, dedicated networks, substantial budgets, and specialized interfaces.
Currently concentrating on download, migrating to analysis in our environment.	Focus is on large volumes, reducing actual data transfer, computational scaling, etc.
CMIP5 is absolutely about persistence for us. IPCC role!	Historically less emphasis placed on archiving, but it is an increasing concern.
NetCDF, CF Conventions CMOR (etc)	Rely heavily on data and metadata standards and typically use relational (e.g., MySQL) and hierarchical (e.g. HDF) data structures and organizational schemes

Metaphors: Science Support

What we do	Parsons and Fox
The day job. We interact with researchers as early as possible as we're not directly connected to their labs/research station.	Embedded, operational support structure typically associated with a research station or lab.
We concentrate on data management, and providing services to make it and use of data easier for the researcher.	Data management is seen as a component or function of the broader "science support" infrastructure of the lab or the project. May include: facilities management, field logistics, administrative support, systems administration, equipment development, etc.
CEDA creates/manages disciplinary specific datasets with an international user base	Focus on creating community collections by characterizing important fundamental processes or particular representative conditions over time.
Because we have so much data, we standardise in terms of file formats, metadata models etc.	The data are organized in myriad ways, usually geared towards a specific set of intended uses and local reuse in conjunction with other local data

Metaphors: Map Making

What we do	Parsons and Fox
	Map Making shares attributes of the other paradigms
	Important metaphor is not the final product or the production process but the representation of the data and their associated science questions through a geographical perspective, notably the map.
We have datasets which are dynamical maps e.g. met. radar, remote sensing	Data in this approach tend to be more fixed in time, i.e. they are more geared toward describing geospatial features rather than dynamic processes.
Some use of OGC standards to expose our data into GIS.	Focus tends to be on cartographic visualization and intercomparison with uneven attention to preservation. Data are well standardized around a map- (or grid-) based model with an associated (geo)database.

Metaphors: Linked Data

What we do	Parsons and Fox
Linked data principles more commonly used for the metadata about the dataset rather than the data itself.	The “data” in Linked Data are defined extremely broadly and are envisioned as small, independent bits with specific names (URIs) interconnected through defined semantic relationships (predicates) using model and language standards (e.g. the Resource Description Framework, RDF).
	Focus almost entirely on enabling interoperability and capitalizing on the interconnected nature of the web.
	Major emphasis on open data.
Total focus on this issue (even for metadata)	Scant attention is paid to preservation, curation, or quality.
	Uses a graph model not a hierarchical or relational model of data organization. Lends itself well to very distributed and interdisciplinary connections
	Requires substantial agreement on the formal semantics, i.e. ontologies, to be useful for diverse audiences.

Right Ideas? Wrong interpretation?

Table 2. Summary of strengths and weaknesses of the data management worldviews.

	Data Publication	*Big Iron*	*Science Support*	*Map Making*	*Linked Data*
Trust	good	moderate	good	moderate	Poor
Discovery	poor	moderate	poor	moderate	Good
Preservation	good	poor	variable	poor	Poor
Access	moderate	moderate	poor/moderate	good	Good
Usability	moderate	moderate	good	moderate/good	Moderate
Governance	poor	good	poor	moderate	Poor
Credit and accountability	good	poor/ moderate	variable	poor/moderate	Variable

Curation?

?

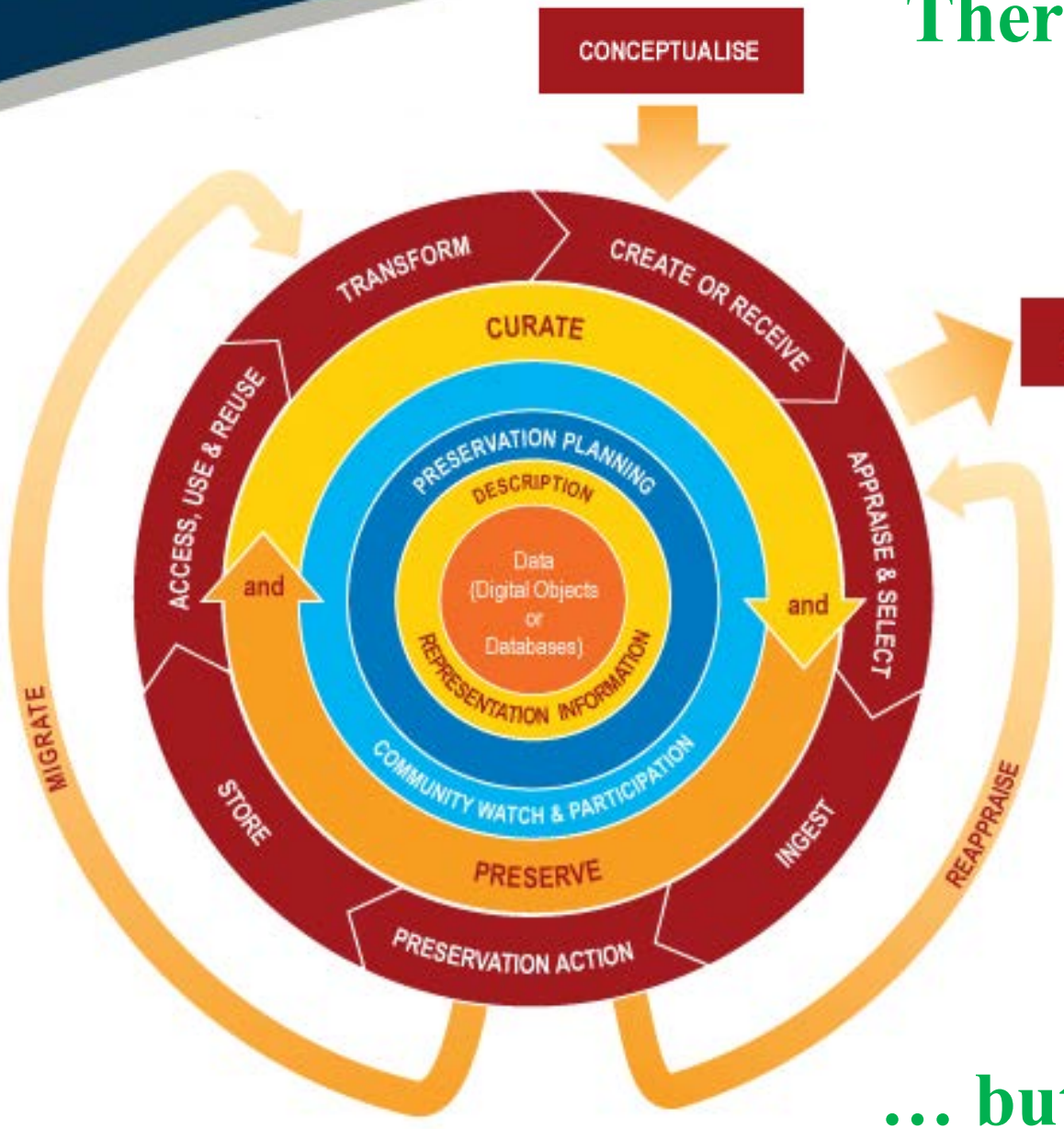
ETC

Out Of Scope

Missing:
Provenance?

We use big iron to provide access to data publication with science support delivering curation, with map making and linked services on top …
THESE ARE NOT ORTHOGONAL CONCEPTS

There are other metaphors …



The Digital Curation Centre's Curation Lifecycle Model provides a graphical, high-level overview of the stages required for successful curation and preservation of data from initial conceptualisation or receipt through the iterative curation cycle.

… but they're limited too!

http://www.dcc.ac.uk/resources/curation-lifecycle-model

Metaphors: The Conclusion

Parsons and Fox were wrong:

- In their (they admit "simplistic") analysis of Data Publication they draw rather too many conclusions, based on (we think) an overly rigorous expectation of homogeneity and applicability for this particular metaphor (primarily because "it's the most mature"?)
- We see Data Publication as one of the many tools we have in CEDA, but it's by no means "mature" nor all-encompassing, so

Parsons and Fox were right:

- Data publication is not THE right Metafor, *but it is a useful one!*
- They recommend mixing and matching as a solution … all the metaphors are in play, but this means

Parsons and Fox were wrong:

- We can't cleanly delineate strengths and weaknesses of non-orthogonal concepts

"It is high time that the global data infrastructure was recognized as "operational" and resourced as such."

A National Strategy for Advancing Climate Modeling, 2011

Some of our activities are operational, and some are not. Some are well resourced, and some are not.

It's a bit early to analyse the strengths and weaknesses on the basis of practice … many of these concepts apply to things we (as a community) are experimenting with … sometimes because an experimental (soft money funded) activity is the only way we can progress ...

Limitations of Effort

So we want:

Trust

Discovery

Preservation/Fixity

Curation

Access

Usability

Governance

Credit and Accountability

Provenance

A, B, C, D Metadata

But these are hard things, and we are only learning how to do some of them …

Whatever combination of activities (aka metaphors) we use to deliver these, we all have varying incomplete implementations ...

Big problems in expectation management?

Should we try to do less? Google Foo ...