Science & Technology Facilities Council

OpenAIRE

Open Research Data

Sarah Callaghan*
sarah.callaghan@stfc.ac.uk
@sorcha_ni

Autumn training school Development and Promotion of Open Access to Scientific Information and Research
19 September, 2014, Veliko Tarnovo, Bulgaria

* and a lot of others, including, but not limited to: the NERC data citation and publication project team, the PREPARDE project team, the OpenAIREplus project and the CEDA team

Science & Technology Facilities Council

Who are we and why do we care about data?

The UK's Natural Environment Research Council (NERC) funds six data centres which between them have responsibility for the long-term management of NERC's environmental data holdings.

We deal with a variety of environmental measurements, along with the results of model simulations in:

- Atmospheric science
- Earth sciences
- Earth observation
- Marine Science
- Polar Science
- Terrestrial & freshwater science, Hydrology and Bioinformatics

British Antarctic Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

NGDC
National Geoscience Data Centre

BODC

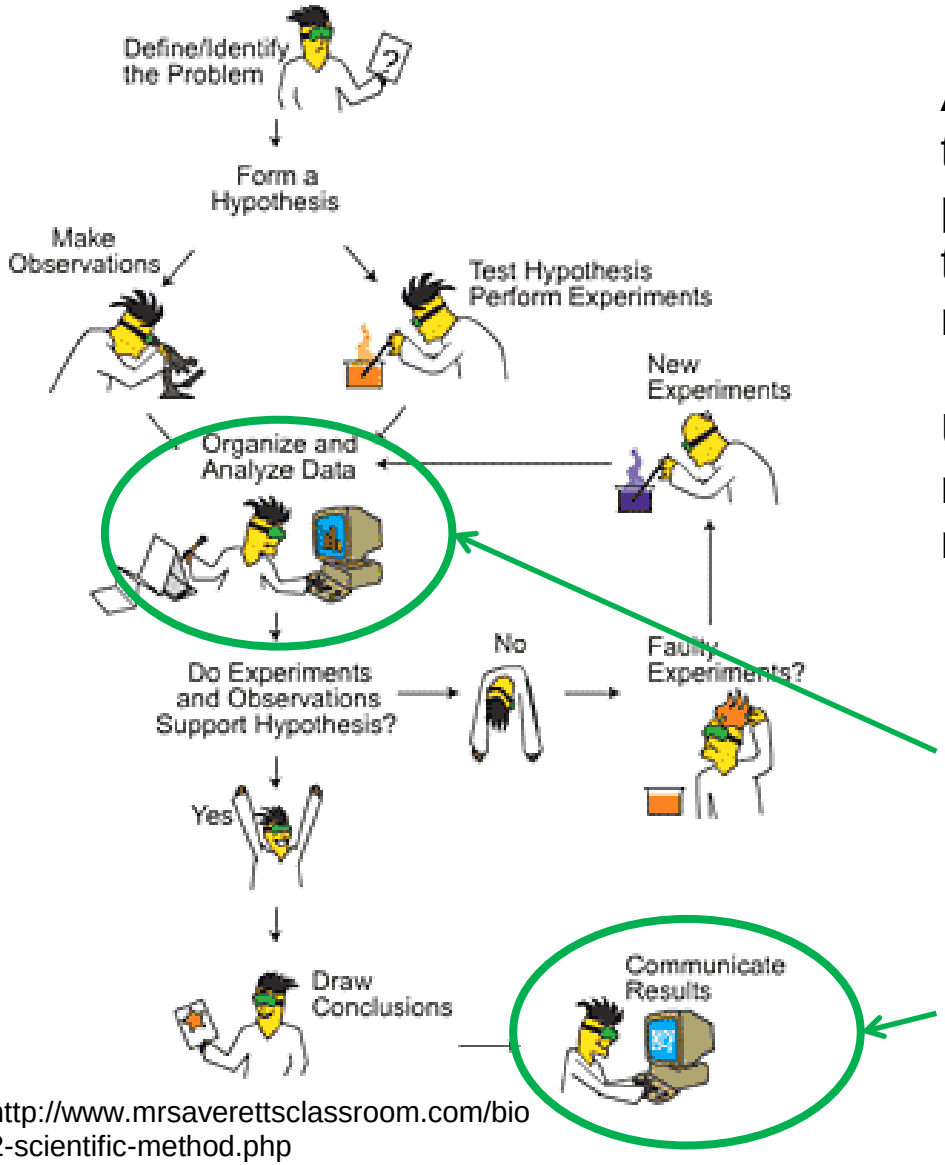
CEH Centre for Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL

NEODC

British Atmospheric Data Centre
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL

UK Solar System Data Centre

The Scientific Method



http://www.mrsaverettsclassroom.com/bio2-scientific-method.php

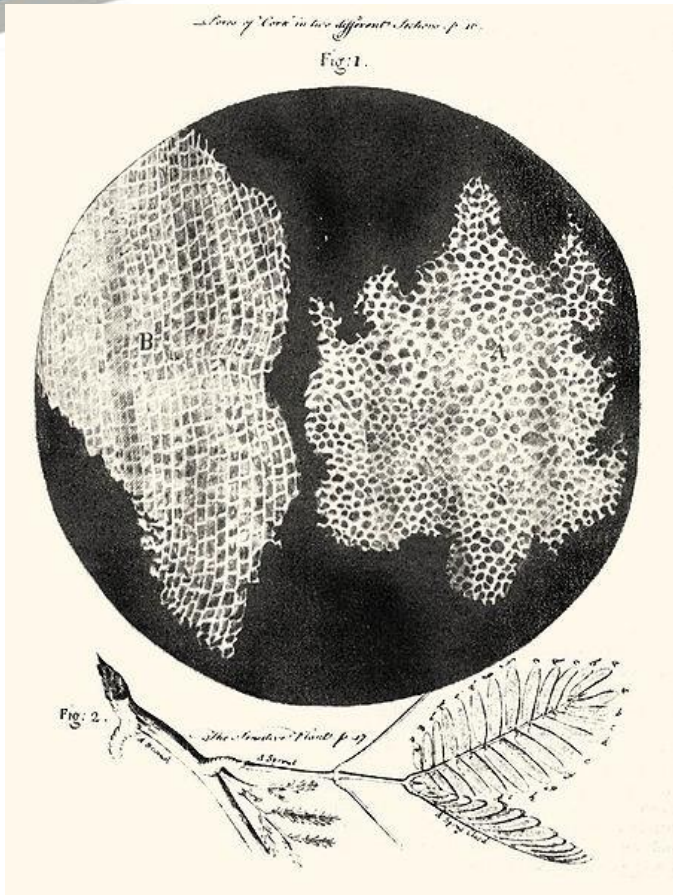
A key part of the scientific method is that it should be reproducible – other people doing the same experiments in the same way should get the same results.

Unfortunately observational data is not reproducible (unless you have a time machine!)

The way data is organised and archived is crucial to the reproducibility of science and our ability to test conclusions.

This is often the only part of the process that anyone other than the originating scientist sees. We want to change this.

Journals have always published data…



Suber cells and mimosa leaves. Robert Hooke, Micrographia, 1665

[*Observations of Stars in the Spiral Nebula.* H. 1622.

The spiral form of this nebula is very distinctly seen in the Pulkova refractor. Unfortunately in the month of March, the best season for the observation of this object, the sky was constantly cloudy; so that I could only get three nights' observations in the months of April and May, when the twilight did not cease for the whole night. It must be attributed to this unfavourable circumstance that the following list of determinations is not so complete as it probably would have been without the twilight. The observations have been made alternately with powers of 138 and 207.

Observations.

Date.	Object.	Magnitude.	Ang. Pos.	No. of measures.	Distance.	No. of measures.
1851, April 7.	N *n*	.....	14° 55′	5	267″·1	4
	N *a*	*a* =(11)	229 24	3	88·0	3
	N *b*	*b* =(11.12)	109 12	3	242·6	3
	a b	.....	93 42	3	298·6	3
April 28.	*a b*	.....	94 23	3	300·8	4
	N *a*	.....	228 36	4		
	N *b*	.....	108 54	4		
	n a	.....	283 42	3		
	n b	.....	153 30	3		
	a d	*d* =(12.13)	323 51	3		
	N *d*	.....	277 27	3		
	a e	*e* =(13)	112 13	3		
	N *e*	.....	161 56	3		
	N *f*	*f* =(12.13)	309 18	3		
	n f	.....	237 31	3		
	a f	.....	335 23	3		
	a g	*g* =(12.13)	215 17	3	115·5	4
	a k	*k* =(12.13)	193 29	3		
	g k	.....	87 5	3		
May 3.	N *k*	*k* =(13.14)	51 47	3		
	n k	.....	173 29	4		
	b k	.....	317 23	3		
	b l	*l* =(11.12)	27 20	4		
	n l	.....	83 17	4	355·2	4
	a e	.....	112 56	4		
	N *e*	.....	161 39	3		
	a m	*m* =(12.13)	172 43	5		
	N *m*	.....	190 44	4		
	b m	.....	238 50	4		
	N *a*	.....	229 12	4	87·0	3
	N *n*	.....	14 47	4	264·2	3

The Scientific Papers of William Parsons, Third Earl of Rosse 1800-1867

…but datasets have gotten so big, it's not useful to publish them in hard copy anymore

Why make data open?

http://www.evidencebased-management.com/blog/2011/11/04/new-evidence-on-big-bonuses/

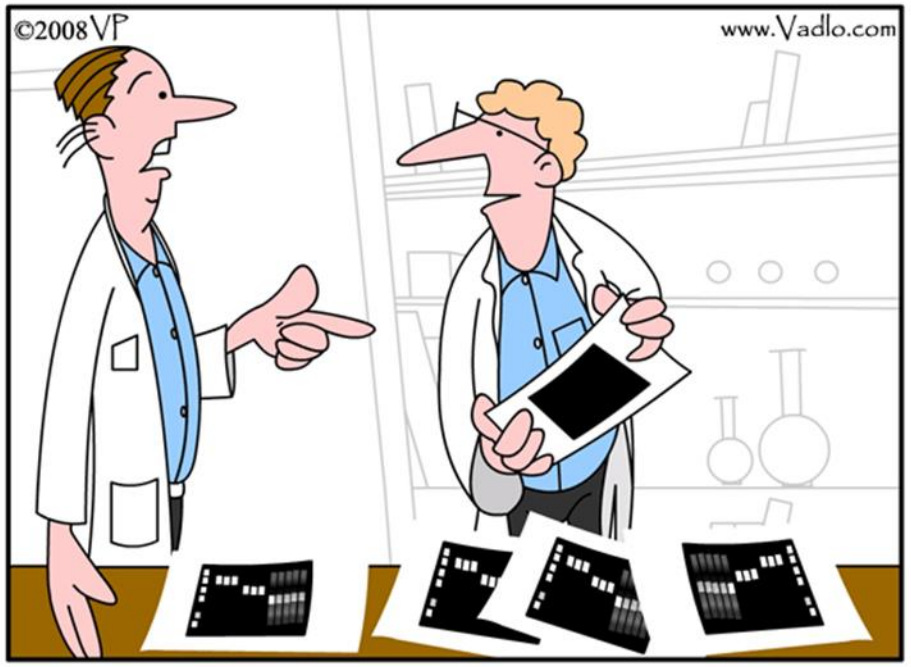
- Pressure from (UK) government to make data from publicly funded research available for free.
  - Scientists want attribution and credit for their work
  - Public want to know what the scientists are doing
  - Good for the economy if new industries can be built on scientific data/research

- Research funders want reassurance that they're getting value for money
  - Relies on peer-review of science publications (well established) and data (starting to be done!)

- Allows the wider research community and industry to find and use datasets, and understand the quality of the data

Need reward structures and incentives for researchers to encourage them to make their data open – data citation and publication

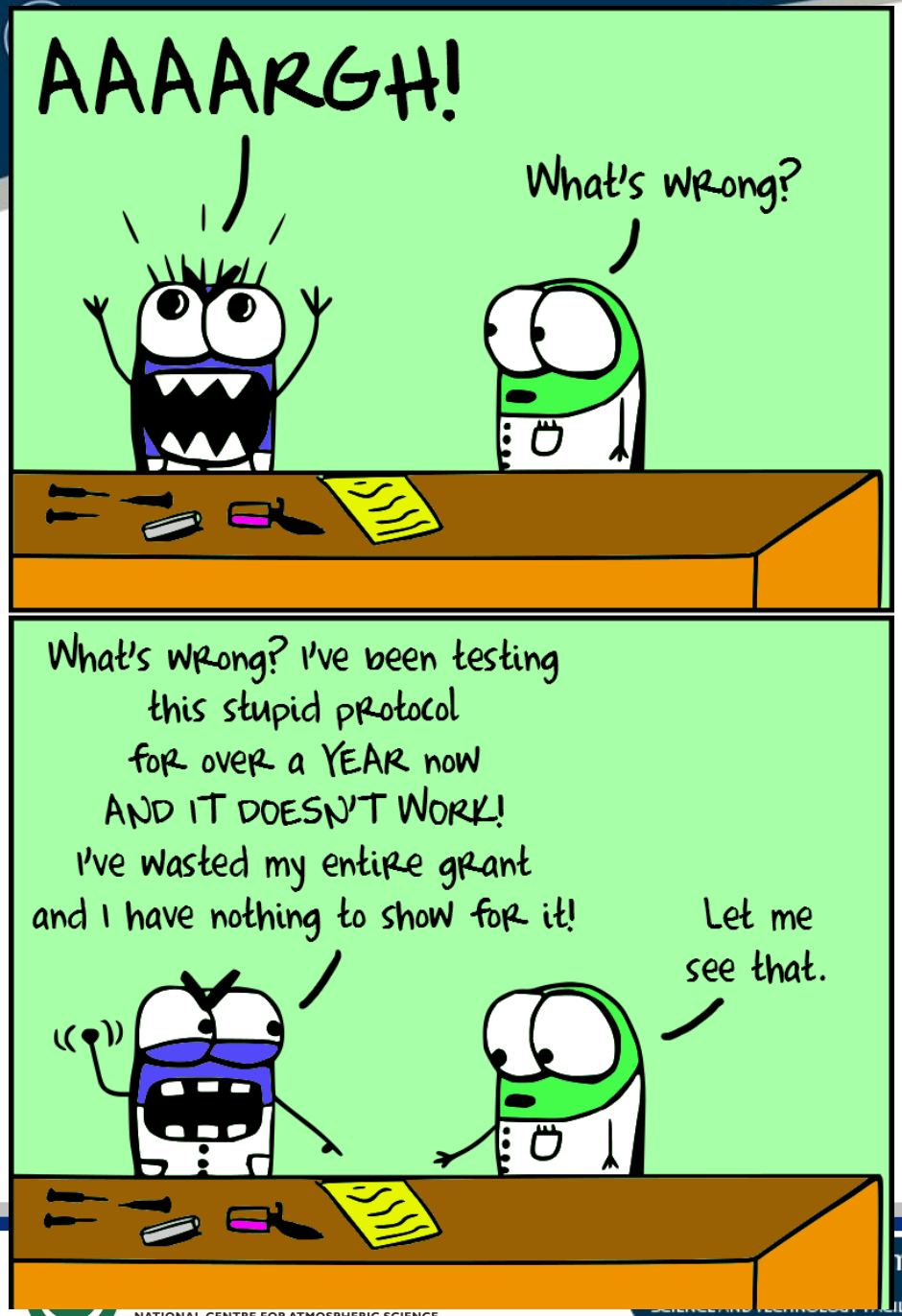
Why bother linking the data to the publication? Surely the important stuff is in the journal paper?

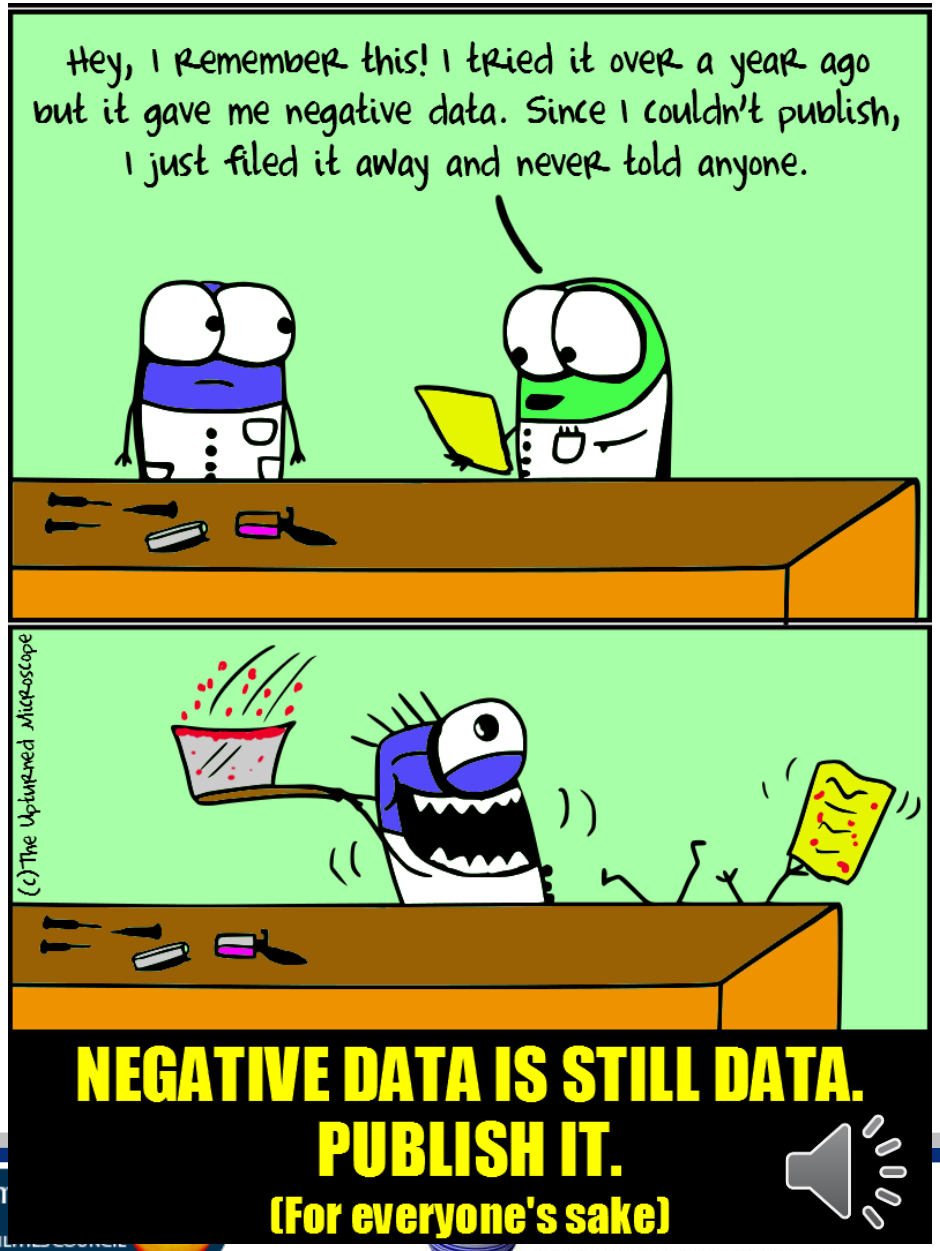


Data don't make any sense, we will have to resort to ***statistics****.*

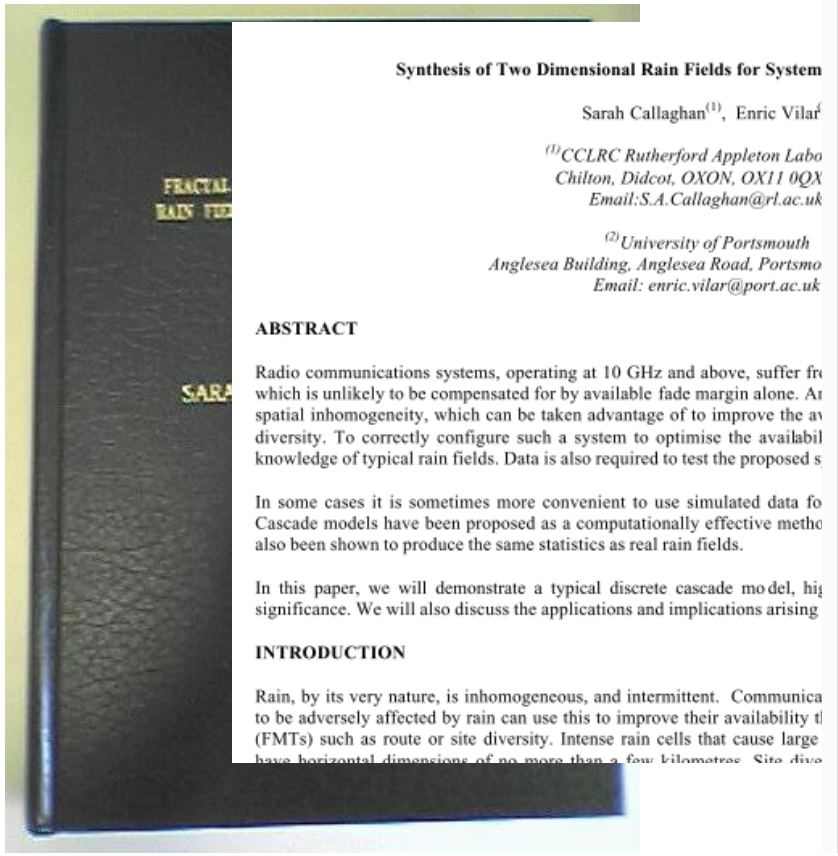
If you can't see/use the data, then you can't test the conclusions or reproduce the results! It's not science!

http://theupturnedmicroscope.com/comic/negative-data/





Most people have an idea of what a publication is



Synthesis of Two Dimensional Rain Fields for System

Sarah Callaghan[(1)], Enric Vilar[(2)]

[(1)]CCLRC Rutherford Appleton Labo
Chilton, Didcot, OXON, OX11 0QX
Email:S.A.Callaghan@rl.ac.uk

[(2)]University of Portsmouth
Anglesea Building, Anglesea Road, Portsmo
Email: enric.vilar@port.ac.uk

ABSTRACT

Radio communications systems, operating at 10 GHz and above, suffer fr
which is unlikely to be compensated for by available fade margin alone. A
spatial inhomogeneity, which can be taken advantage of to improve the a
diversity. To correctly configure such a system to optimise the availabil
knowledge of typical rain fields. Data is also required to test the proposed s

In some cases it is sometimes more convenient to use simulated data fo
Cascade models have been proposed as a computationally effective metho
also been shown to produce the same statistics as real rain fields.

In this paper, we will demonstrate a typical discrete cascade model, hig
significance. We will also discuss the applications and implications arising

INTRODUCTION

Rain, by its very nature, is inhomogeneous, and intermittent. Communica
to be adversely affected by rain can use this to improve their availability t
(FMTs) such as route or site diversity. Intense rain cells that cause large
have horizontal dimensions of no more than a few kilometres. Site dive

RMetS Royal Meteorological Society — Geoscience Data Journal — Open Access

Data Paper

The GBS dataset: measurements of satellite site diversity at 20.7 GHz in the UK

S. A. Callaghan*, J. Waight, J. L. Agnew, C. J. Walden, C. L. Wrench and S. Ventouras

Article first published online: 17 MAR 2013
DOI: 10.1002/gdj3.2

Issue

Geoscience Data Journal
Volume 1, Issue 1, pages 2–6, June 2014

Additional Information (Show All)
How to Cite | Author Information | Publication History | Funding Information

The research presented in this paper was funded by the UK's Ofcom as part of the Spectrum Efficiency Scheme and the support of Ofcom in providing the funding for the GBS experiment is greatly appreciated.

Keywords:
site diversity; radio propagation; fade mitigation techniques

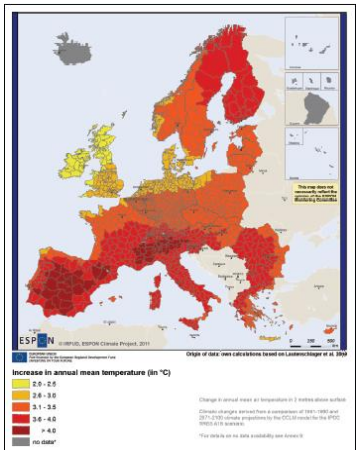
Abstract

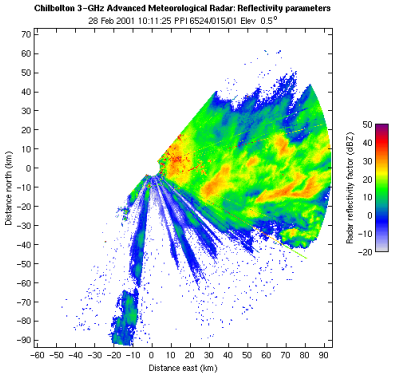
The GBS (Global Broadcast Service) dataset is a series of radio attenuation measurements made at three sites in the UK: Chilbolton and Sparsholt, both in southern UK, and Dundee in Scotland. The aim of the experiment was to make long term measurements of the signal strength received from a 20.7 GHz beacon on the US Department of Defense satellite UFO-9 at multiple sites, in order to determine whether the use of site diversity as a fade mitigation technique would be effective. The dataset spans a period of 3 years, from August 2003 to August 2006 with signal attenuation sampled once per second.

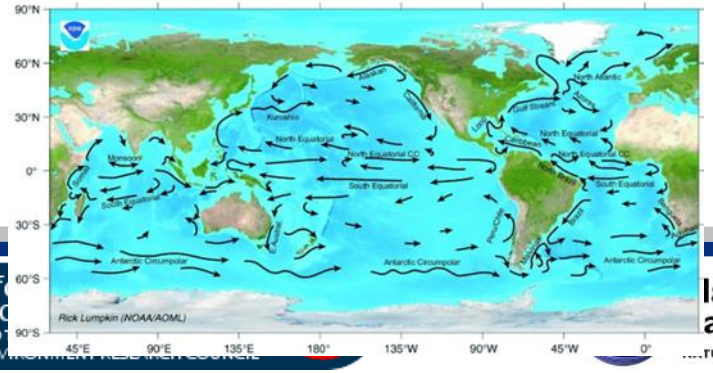
Science & Technology Facilities Council

Some examples of data (just from the Earth Sciences)

1. Time series, some still being updated e.g. meteorological measurements
2. Large 4D synthesised datasets, e.g. Climate, Oceanographic, Hydrological and Numerical Weather Prediction model data generated on a supercomputer
3. 2D scans e.g. satellite data, weather radar data
4. 2D snapshots, e.g. cloud camera
5. Traces through a changing medium, e.g. radiosonde launches, aircraft flights, ocean salinity and temperature
6. Datasets consisting of data from multiple instruments as part of the same measurement campaign
7. Physical samples, e.g. fossils







NATIONAL CENTRE FOR ATMOSPHERIC SCIEI
NATURAL ENVIRONMENT RESEARCH COUNC...

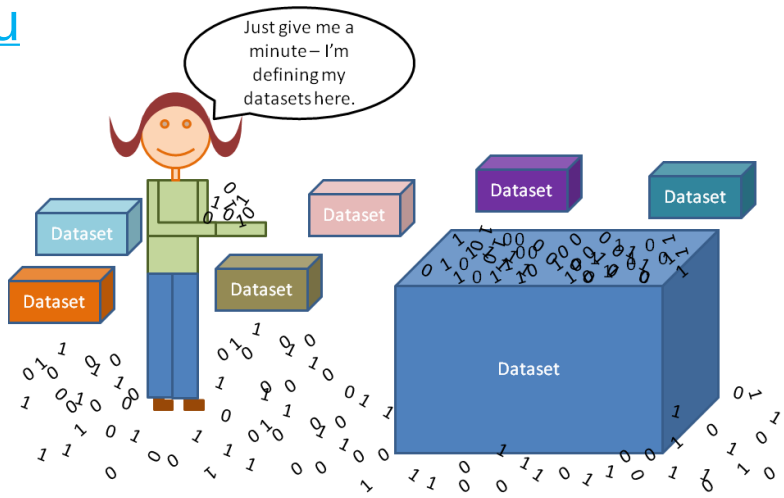
...arth Observation
...TURAL ENVIRONMENT RESEARCH COUNCIL

What is a Dataset?

DataCite's definition (http://www.datacite.org/sites/default/files/Business_Models_Principles_v1.0.pdf):

Dataset: "Recorded information, regardless of the form or medium on which it may be recorded including writings, films, sound recordings, pictorial reproductions, drawings, designs, or other graphic representations, procedural manuals, forms, diagrams, work flow, charts, equipment descriptions, data files, data processing or computer programs (software), statistical records, and other research data."

(from the U.S. National Institutes of Health (NIH) Grants Policy Statement via DataCite's Best Practice Guide for Data Citation).



In my opinion a dataset is something that is:

- The result of a defined process
- Scientifically meaningful
- Well-defined (i.e. clear definition of what is in the dataset and what isn't)

Should ALL data be open?

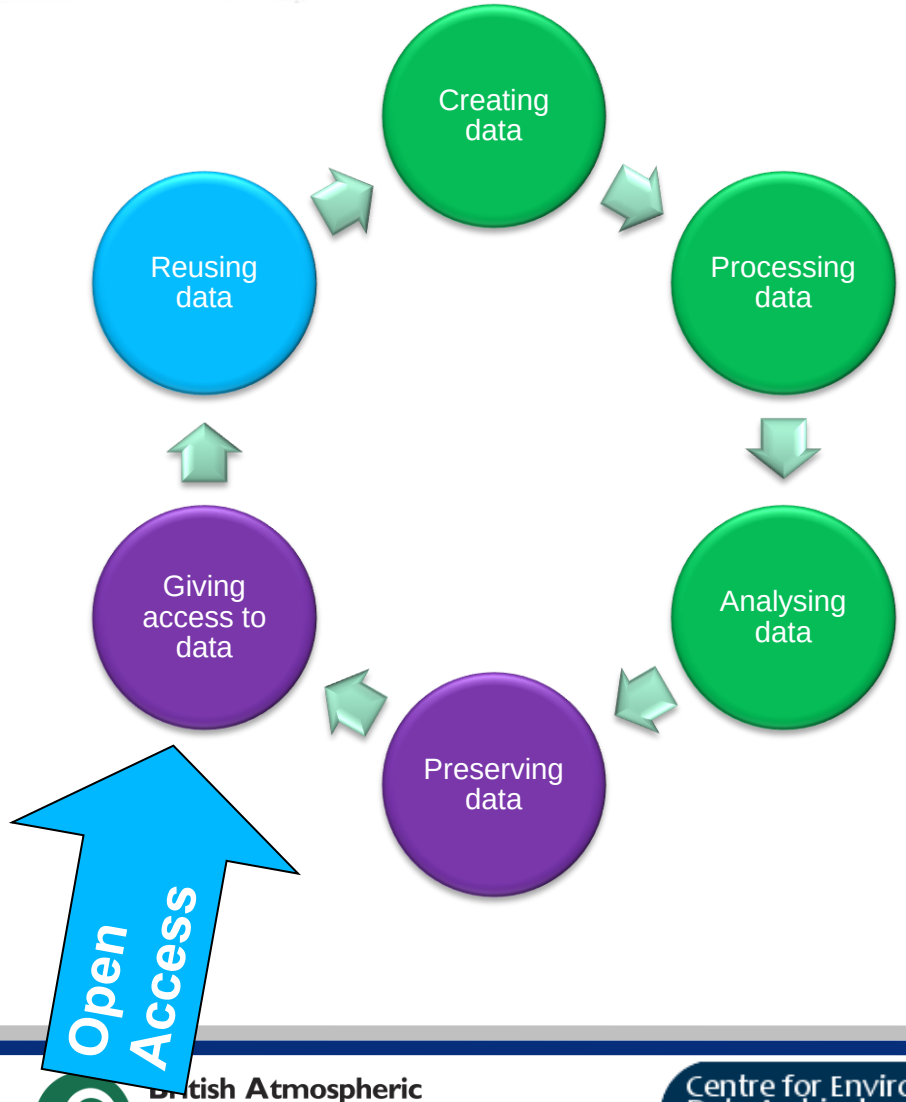
Most data produced through publically funded research should be open.

But!

- Confidentiality issues (e.g. named persons' health records)
- Conservation issues (e.g. maps of locations of rare animals at risk from poachers)
- Security issues (e.g. data and methodologies for building biological weapons)

There should be a very good reason for publically funded data to **not** be open.

The research data lifecycle



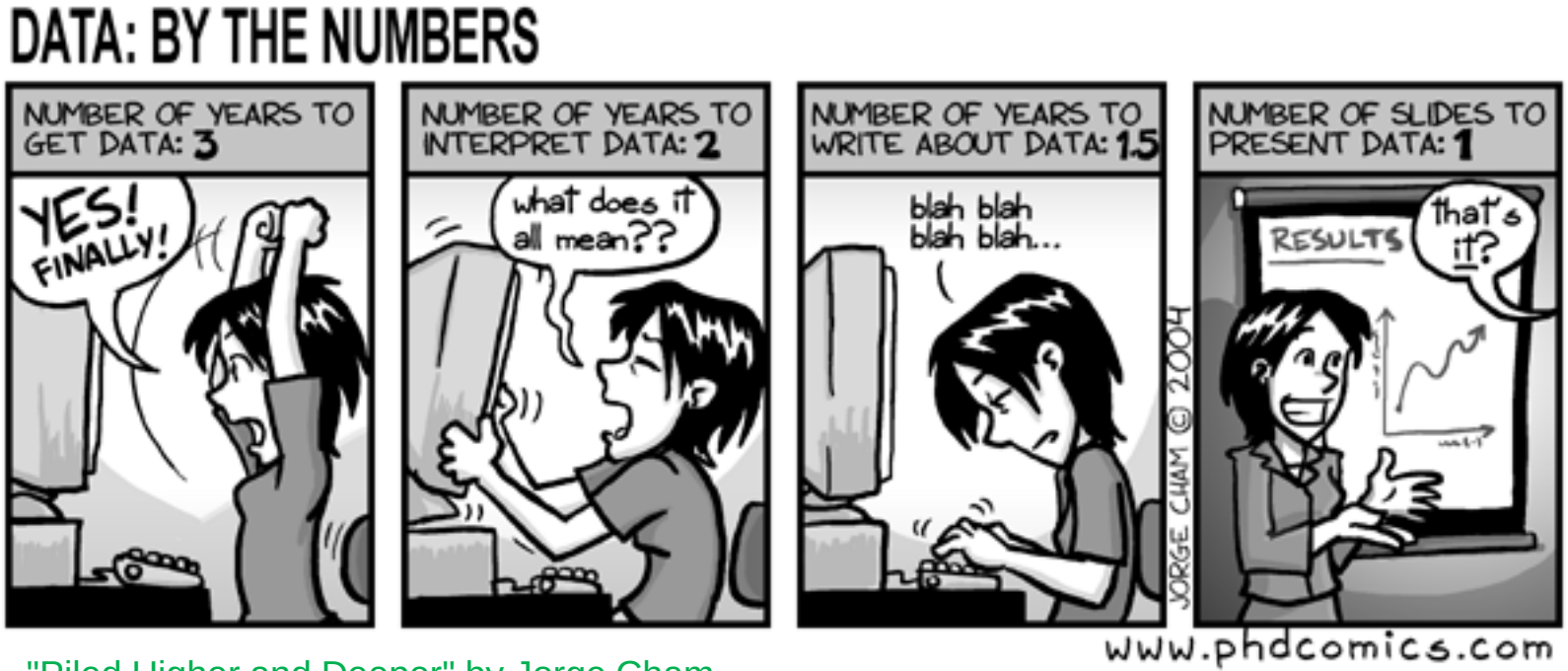
Researchers are used to creating, processing and analysing data.

Data repositories generally have the job of preserving and giving access to data.

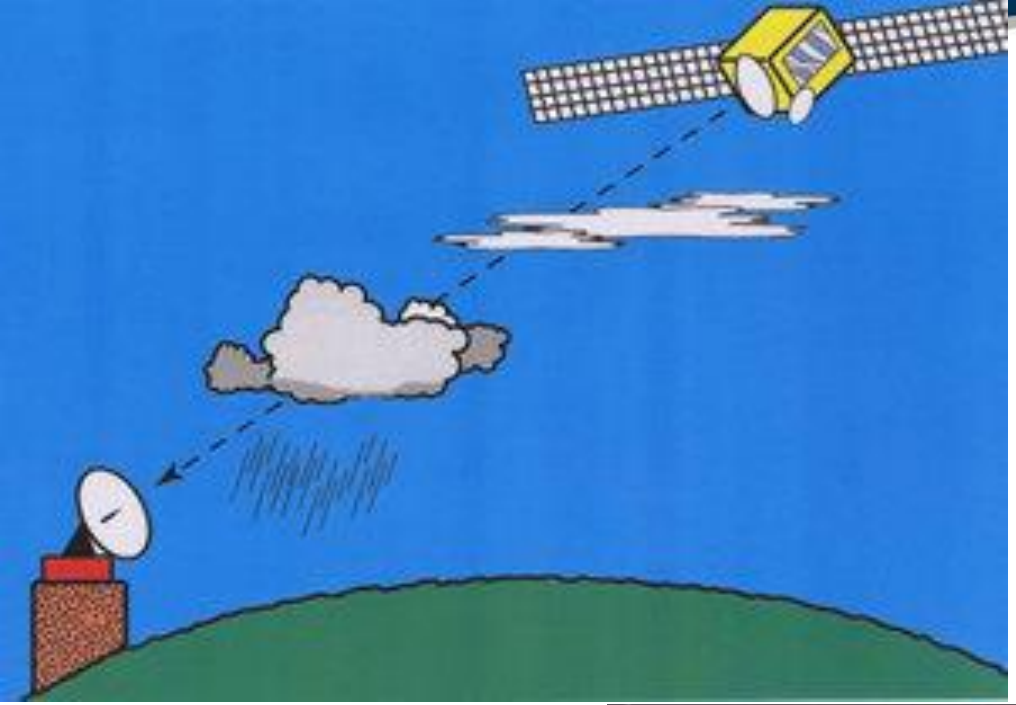
Third parties, or even the original researchers will reuse the data.

See http://data-archive.ac.uk/create-manage/life-cycle for more detail

Creating a dataset is hard work!



"Piled Higher and Deeper" by Jorge Cham
www.phdcomics.com



Creating data: a radio propagation dataset

The problem: rain and cloud mess up your satellite radio signal. How can we fix this?

Italsat F1: Owned and operated by Italian Space Agency (ASI). Launched January 1991, ended operational life January 2001.

British Atmospheric Data Centre
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL

The receive cabin at Sparsholt in Hampshire

Inside the receive cabin – the instruments my data came from

Creating/processing data

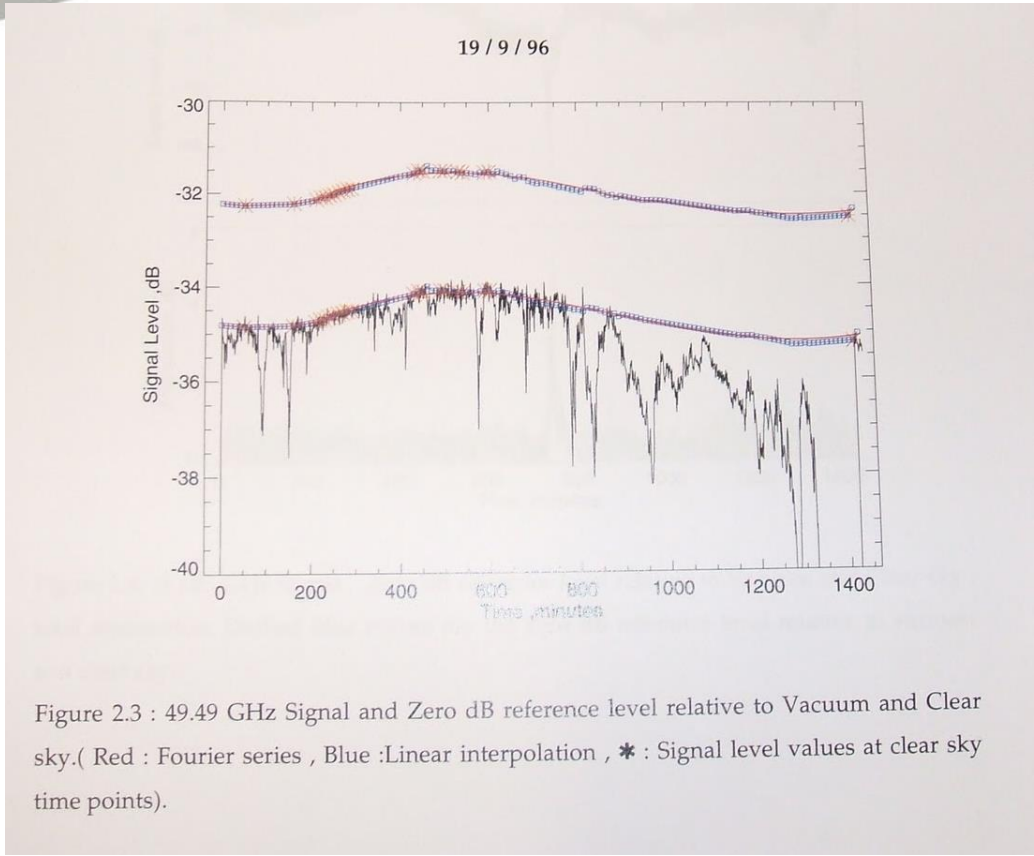
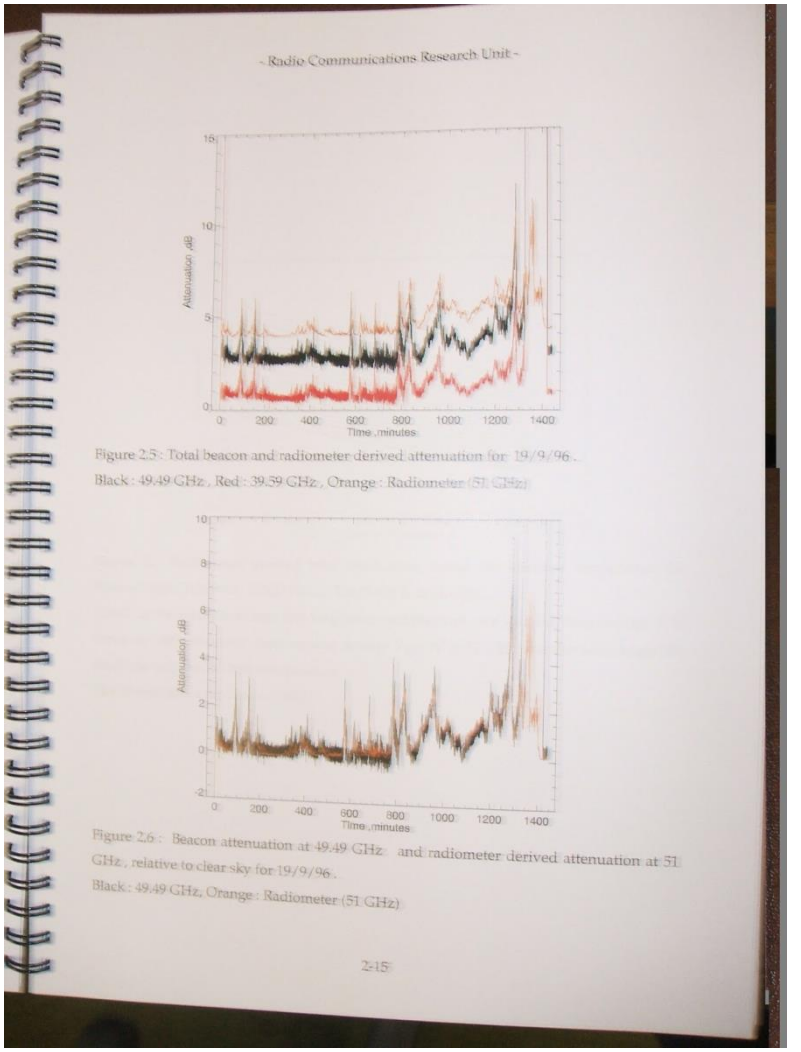


Figure 2.3 : 49.49 GHz Signal and Zero dB reference level relative to Vacuum and Clear sky.(Red : Fourier series , Blue :Linear interpolation , ✱ : Signal level values at clear sky time points).

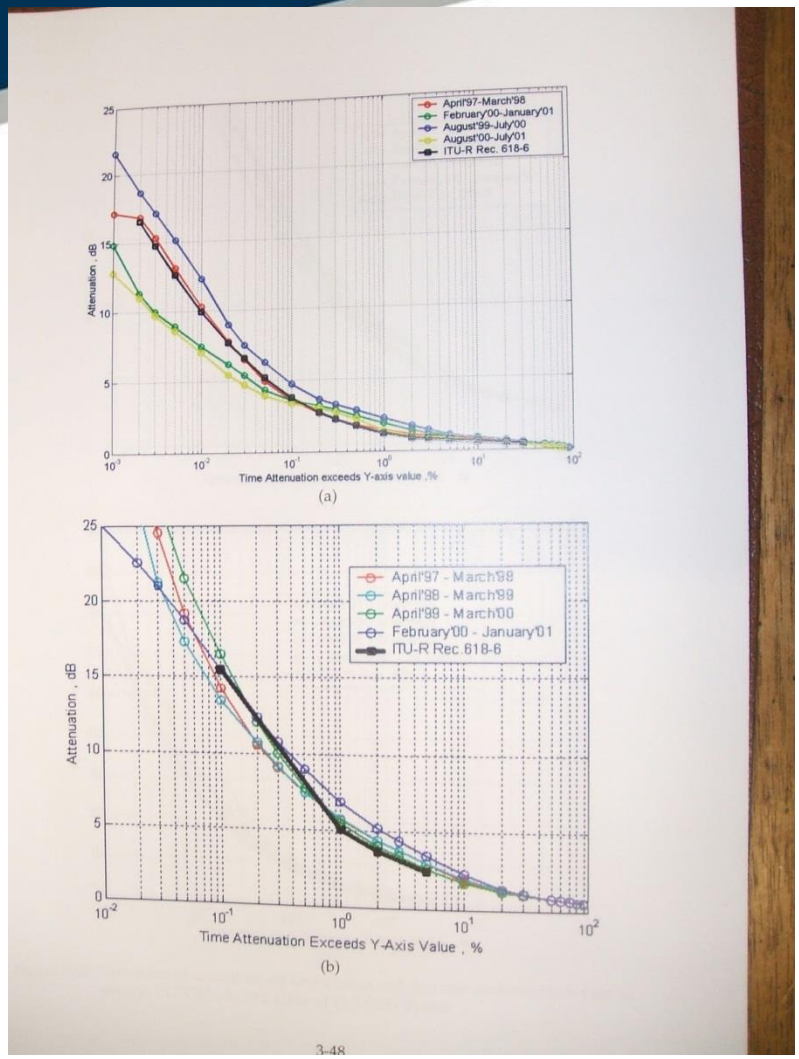
One day's worth of raw data from one of the receivers

My job was to take this...



...turn it into this....

Analysing data



...with the final result being this.

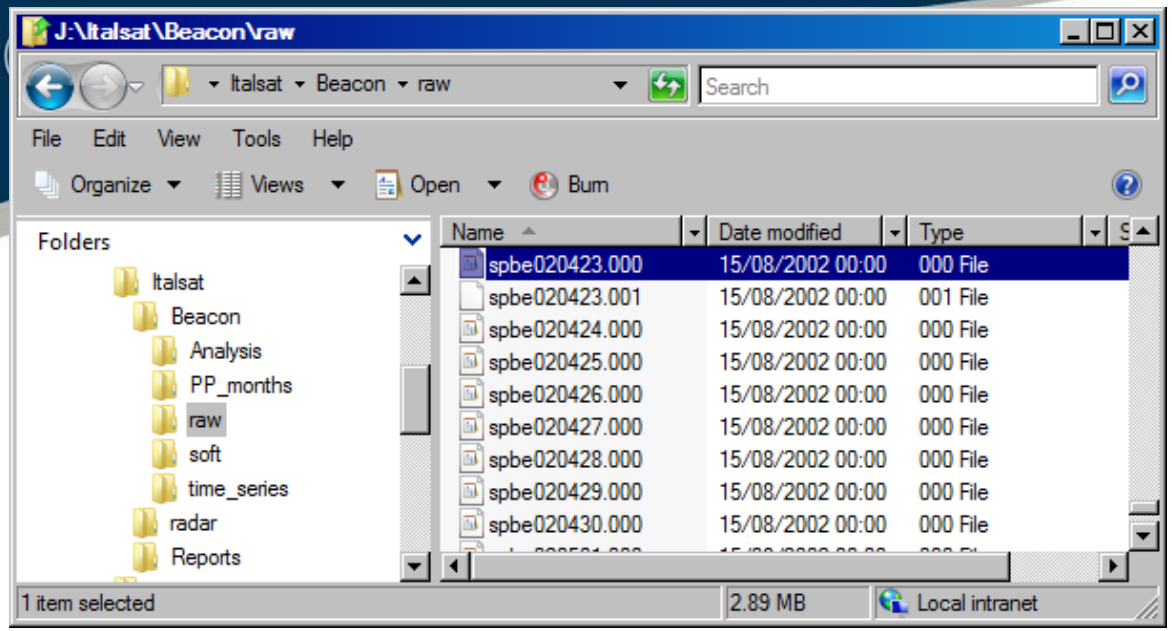
…a process which involved 4 major steps, 4 different computer programmes, and 16 intermediate files for each day of measurements.

Each month of preproccessed data represented somewhere between a couple of days and a week's worth of effort.

It was a job where attention to detail was important, and you really had to know what you were looking at from a scientific perspective.

Preserving data (the wrong way!)

Part of the Italsat data archive – on CDs in a shelf in my office

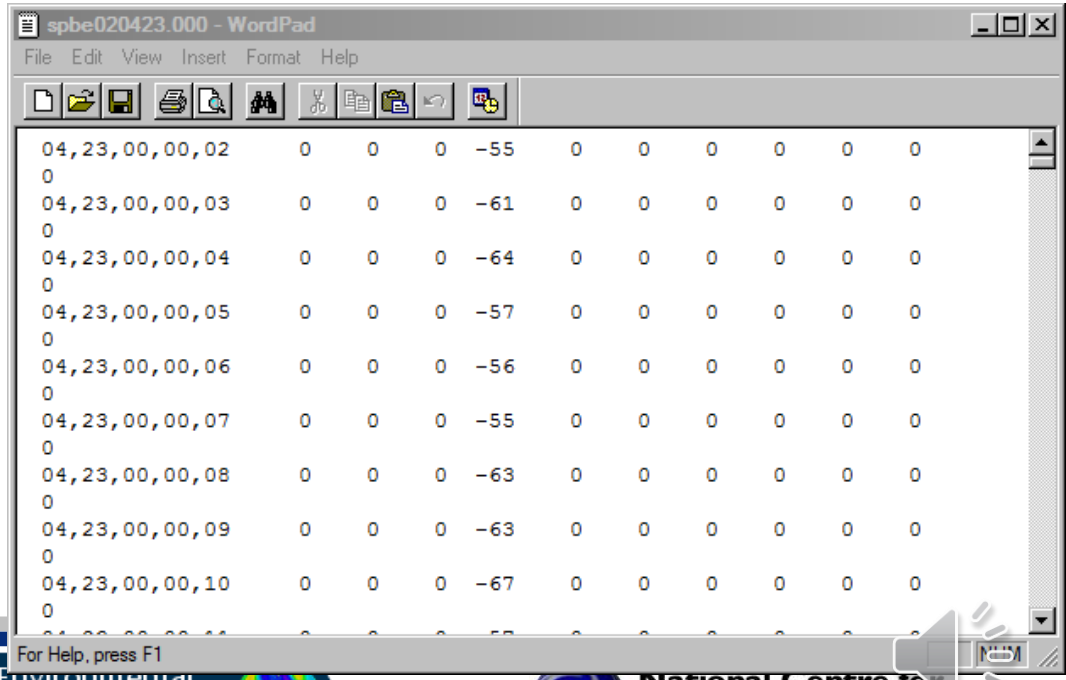


What the processed data set looks like on disk

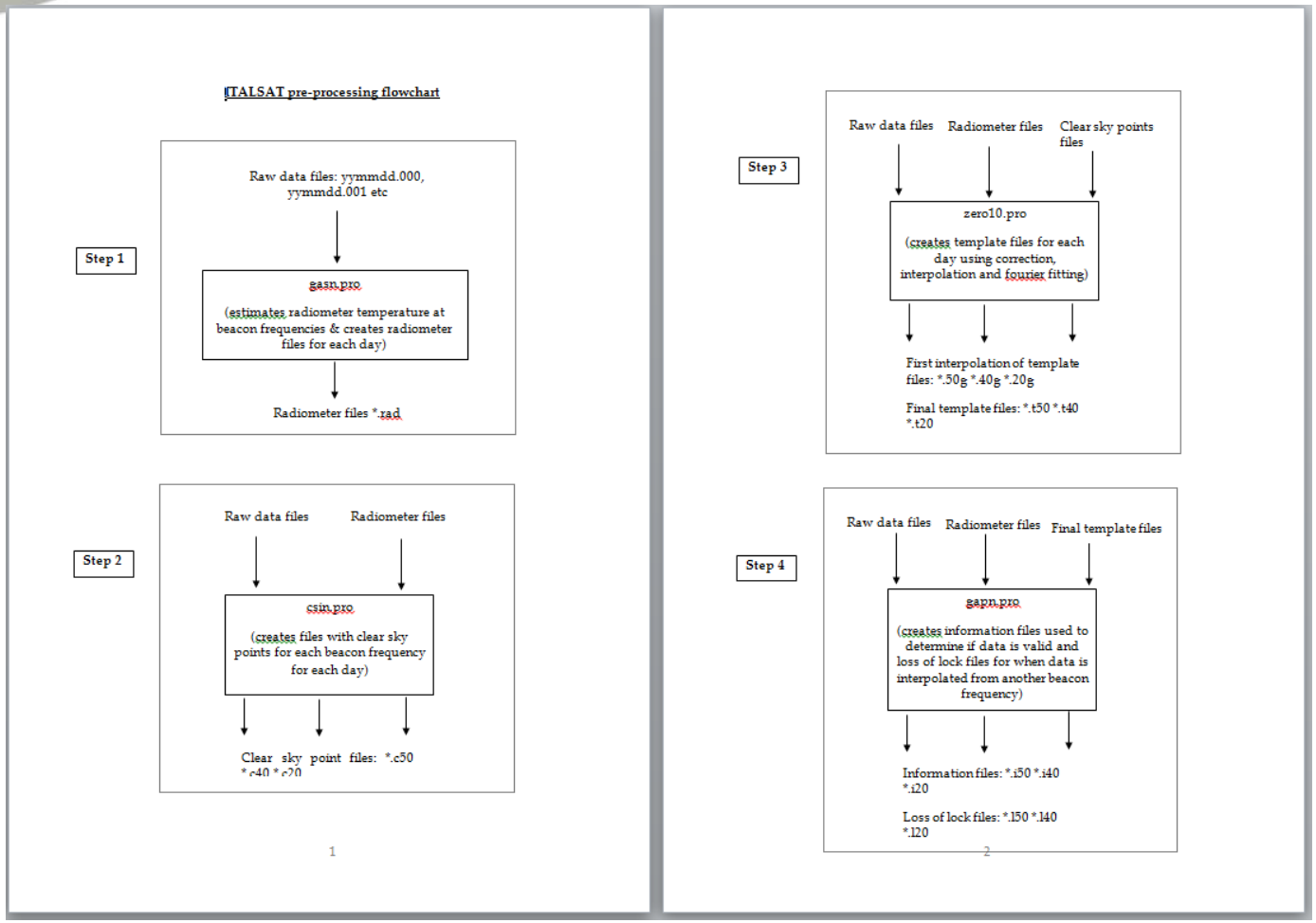
What the raw data files looked like.

(I do have some Word documents somewhere which describe what all this is…)

I could make these files open easily, but no one would have a clue how to use them!



Example documentation



ITALSAT pre-processing flowchart

Step 1

Raw data files: yymmdd.000, yymmdd.001 etc

gasn.pro

(estimates radiometer temperature at beacon frequencies & creates radiometer files for each day)

Radiometer files *.rad

Step 2

Raw data files | Radiometer files

csin.pro

(creates files with clear sky points for each beacon frequency for each day)

Clear sky point files: *.c50 *.c40 *.c20

1

Step 3

Raw data files | Radiometer files | Clear sky points files

zero10.pro

(creates template files for each day using correction, interpolation and fourier fitting)

First interpolation of template files: *.50g *.40g *.20g

Final template files: *.t50 *.t40 *.t20

Step 4

Raw data files | Radiometer files | Final template files

gapn.pro

(creates information files used to determine if data is valid and loss of lock files for when data is interpolated from another beacon frequency)

Information files: *.i50 *.i40 *.i20

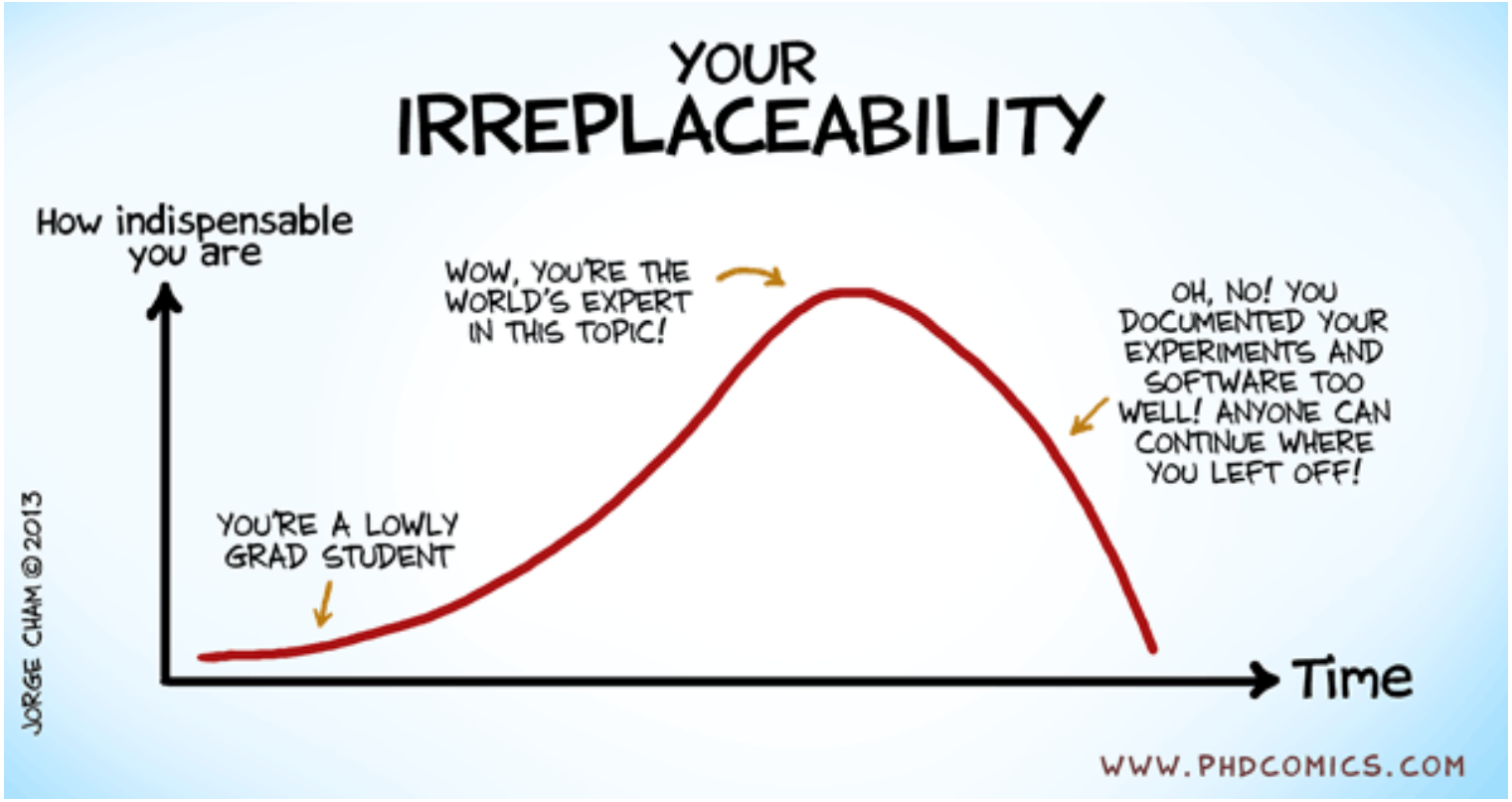
Loss of lock files: *.l50 *.l40 *.l20

2

Note the software filenames in the documentation.

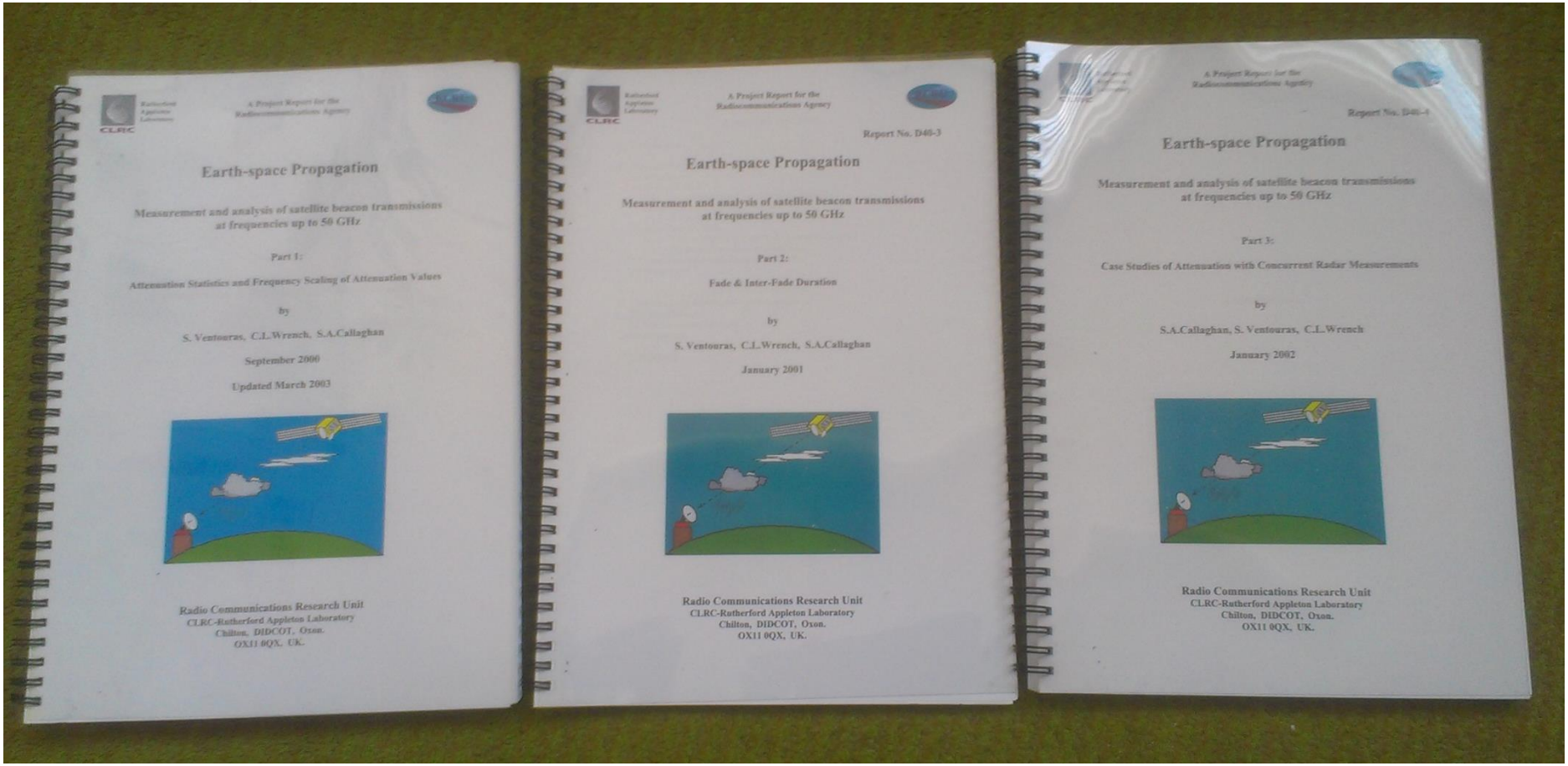
I still have the IDL files on disk somewhere, but I'd be very surprised if they're still compatible with the current version of IDL

Documentation can sometimes produce mixed feelings



"Piled Higher and Deeper" by Jorge Cham
www.phdcomics.com

Publications – grey literature



Publications – journal paper

Where's the data?

RADIO SCIENCE, VOL. 41, RS2007, doi:10.1029/2005RS003252, 2006

Long-term statistics of tropospheric attenuation from the Ka/U band ITALSAT satellite experiment in the United Kingdom

S. Ventouras,[1] S. A. Callaghan,[1] and C. L. Wrench[1]

Received 9 February 2005; revised 9 December 2005; ac

[1] Long-term statistics of tropospheric attenu measurements made in the south of England u 49.5, 39.6, and 18.7 GHz; coincident rainfall r the receiving ground station. A method to rem beacon signals and to establish the reference le total attenuation has been presented in detail. estimated to be ~±0.5 dB. A new method for statistics has been proposed and validated aga 18.7, 39.6, and 49.5 GHz. For both locations, tl predictions compared with the established Inte recommendation method. A significant monthl the attenuation and rainfall statistics and should the design and use of future slant path systems. are subject to diurnal variations; however, for t seem to follow a particular pattern.

Citation: Ventouras, S., and C. L. Wrench (2006), Lo ITALSAT satellite experiment in the United Kingdom,

Table 4. Annual Measured and Predicted Total Attenuation Statistics for Sparsholt, UK[a]

	Total Attenuation, dB								
	49.5 GHz			39.6 GHz			18.7 GHz		
Outage, %	Measured	ITU-R, 0.01%	New Method, All Distribution	Measured	ITU-R, 0.01%	New Method, All Distribution	Measured	ITU-R, 0.01%	New Method, All Distribution
30	3.05	3.09	2.96	0.99	1.06	0.94	0.46	0.42	0.38
20	3.40	3.67	3.50	1.31	1.46	1.29	0.61	0.54	0.46
10	4.38	4.89	4.42	1.96	2.33	1.93	0.84	0.78	0.61
5	5.87	6.30	5.48	3.00	3.34	2.64	0.96	1.05	0.76
3	7.11	7.38	6.47	3.84	4.14	3.30	1.10	1.26	0.89
2	8.14	8.48	7.86	4.54	4.95	4.30	1.36	1.46	1.01
1	10.34	10.53	10.58	6.03	6.50	6.38	1.85	1.85	1.50
0.50	13.28	12.86	13.45	7.98	8.33	8.66	2.45	2.30	2.09
0.30	15.99	15.16	15.78	9.83	10.15	10.54	2.91	2.77	2.59
0.20	18.50	17.39	17.80	11.47	11.92	12.20	3.25	3.25	3.06
0.10	23.45	22.17	21.69	14.95	15.73	15.49	3.91	4.30	4.02
0.050				19.23	20.63	19.49	5.21	5.72	5.28
0.030				23.04	24.98	23.00	6.46	7.04	6.42
0.020							7.50	8.26	7.51
0.010							9.91	10.71	9.75
0.005							12.91	13.59	12.42
0.003							15.04	15.95	14.58
0.002							16.62	17.93	16.34
0.001							17.87	21.42	17.52

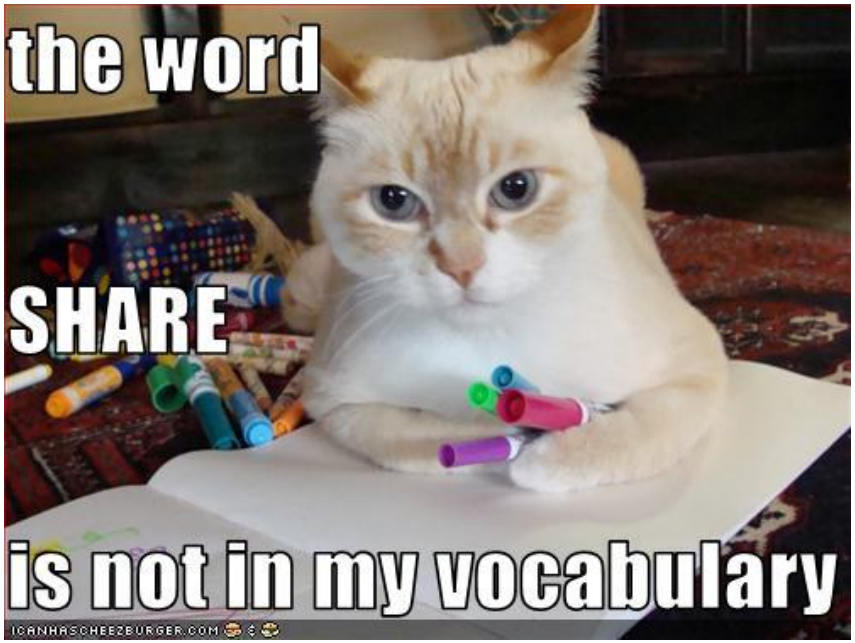
[a]For measured statistics, 49.5 and 39.6 GHz were averaged over 4 years, and 18.7 GHz was averaged over 3 years. For predicted statistics, ITU-R, 0.01% refers to Recommendation P.618-8, and New Method, All Distribution is a proposed combination method, whole rain distribution for rain attenuation statistics.

What it all came down to:

Composite image from Flickr user bnilsen and Matt Stempeck (NOI), shared under Creative Commons license

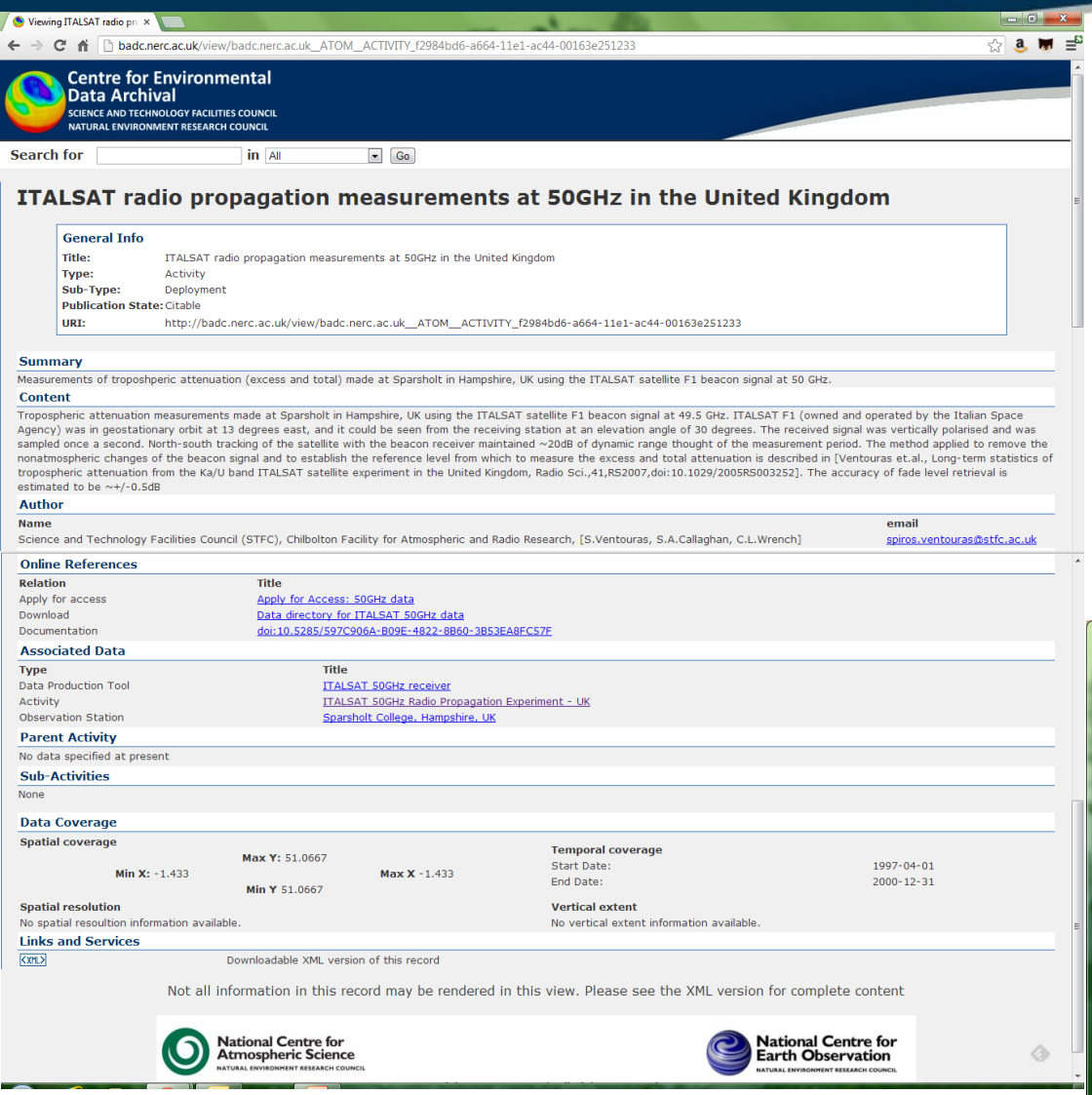
And I wasn't even preserving my data properly!

As for giving access to the data…

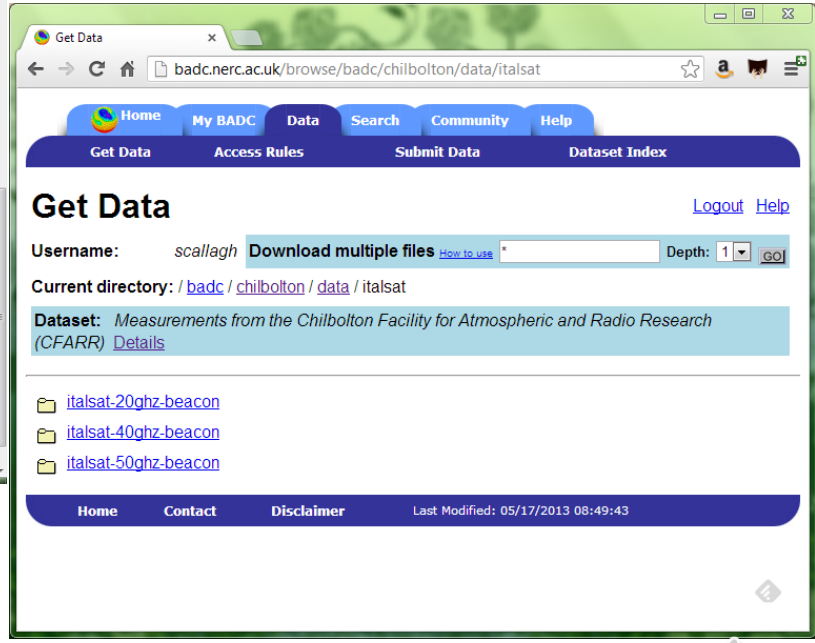


I did share, but there was a lot of non-disclosure agreements (I am not a lawyer!)

And I didn't feel like I got the credit for it.(The first publication based on the data wasn't written by me, and I didn't even get my name in the acknowledgements.)



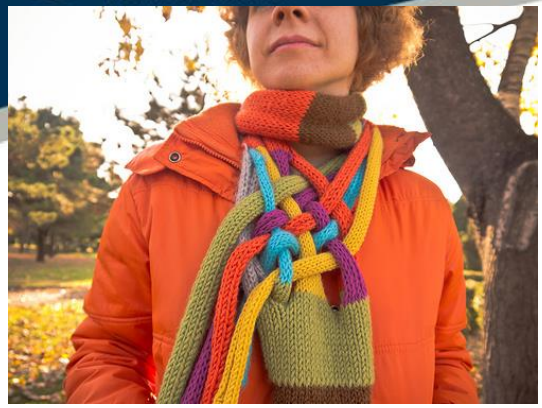
Good news: the data is all open (and documented) on the BADC now



Another example: How is my scarf like a dataset?

- The raw material it's made from doesn't contain information
- But the act of knitting encodes information into the scarf
- The scarf is the result of a well defined process (knitting) and has a particular method used to create it
- I need to be able to describe it
- I need to be able to find it
- I need to store it properly so it doesn't get lost, or corrupted (i.e. eaten by moths or shredded by mice)
- I might need to recreate it so I need to keep information about it
- I put a lot of time and effort into making it, so I'm very attached to it!

Just like not all scarves are the same, not all datasets are the same!



http://www.flickr.com/photos/nazliceti ner/6448303541/

http://www.flickr.com/photos/maco_nix/5 019885742/

How the dataset was created and used will determine how open it can be.

http://www.flickr.com/photos/lo vefibre/3251690074/

http://www.flickr.com/phot os/halfbisqued/80841459 76/

http://www.flickr.com/phot os/lucathegalga/2282305 884/

http://www.flickr.com/ photos/ujkakevin/230 3531028/

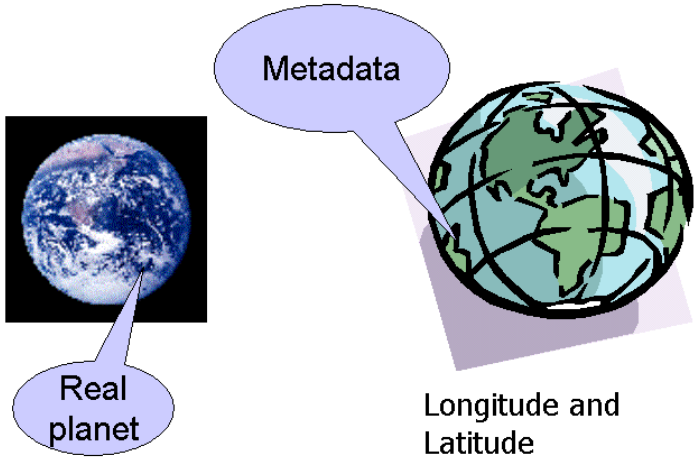
Metadata

It is generally agreed that we need methods to:

- define and document datasets of importance.
- augment and/or annotate data
- amalgamate, reprocess and reuse data

To do this, we need **metadata** – data about data

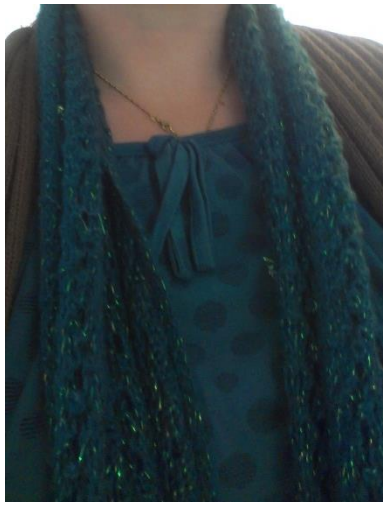
For example:
Longitude and latitude are metadata about the planet.
• They are artificial
• They allow us to communicate about places on a sphere
• They were principally designed by those who needed to navigate the oceans, which are lacking in visible features!



http://www.kcoyle.net/meta_purpose.html

Metadata can often act as a surrogate for the real thing, in this case the planet.

Metadata for my scarf



- Descriptive: “teal blue”, “scarf”
- Dimensions: 200cm long, 20cm wide
- Location: “Around my neck”/”Hanging on the door of my wardrobe”
- Identifier: KOI (knitted object identifier)

Information needed to recreate it:

- The raw material: King Cole Haze Glitter DK, colourway 124 - Ocean, with dyelot 67233
- Needle size: 4mm
- Algorithm used to create it: 18 stitch feather and fan stitch with 2 stitch garter stitch border at the edges
- Number of stitches cast on: 54
- Tension (how tightly I knit in this pattern): 28 rows and 27 stitches for a 10cm by 10cm square

I can’t make my scarf Open Access, but I can make the metadata about it open – enabling other users to create it for themselves.

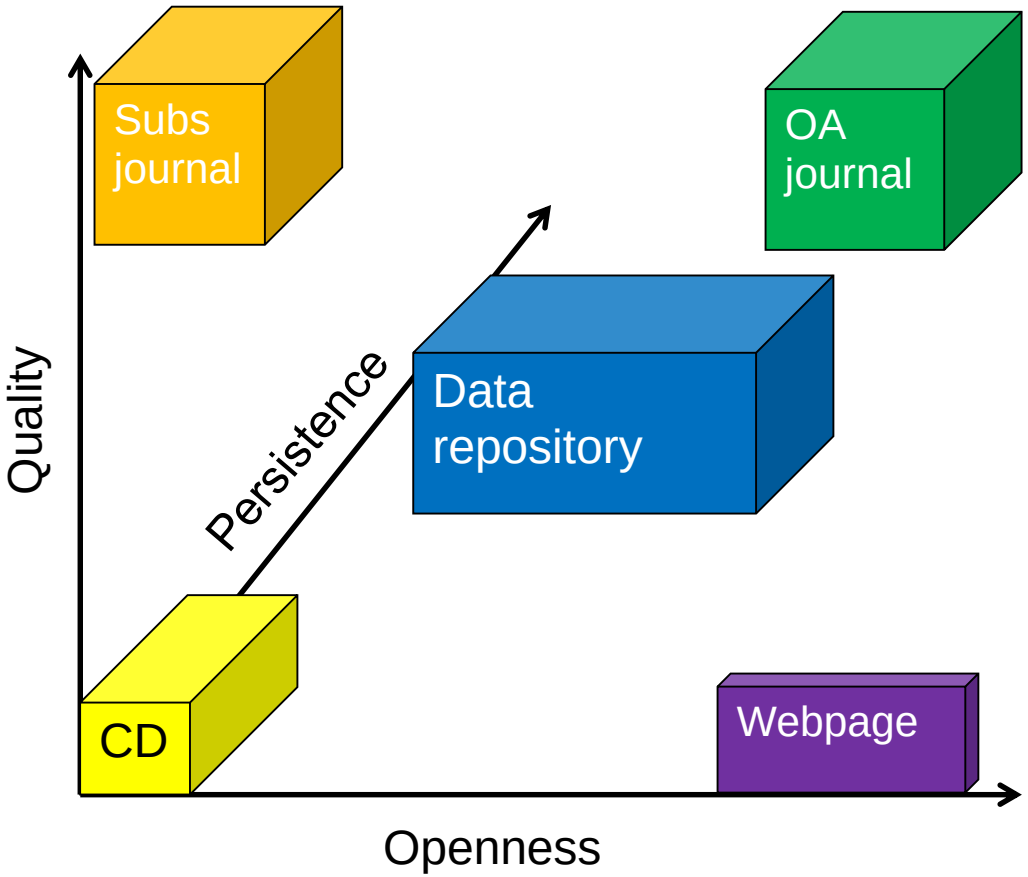
Dataset views and suggested uses

How to publish data/make data open

- Stick it up on a webpage somewhere
  - Issues with stability, persistence, discoverability…
  - Maintenance of the website
- Put it in the cloud
  - Issues with stability, persistence, discoverability…
- Attach it to a journal paper and store it as supplementary materials
  - Journals not too keen on archiving lots of supplementary data, especially if it's large volume.
- Put it in a disciplinary/institutional repository
- Write a data article about it and publish it in a data journal

By David Fletcher
http://www.cloudtweaks.com/2011/05/the-lighter-side-of-the-cloud-data-transfer/

Open/Closed/Published/unpublished



We want to encourage researchers to make their data:

- Open
- Persistent
- Quality assured:
  - through scientific peer review
  - or repository-managed processes

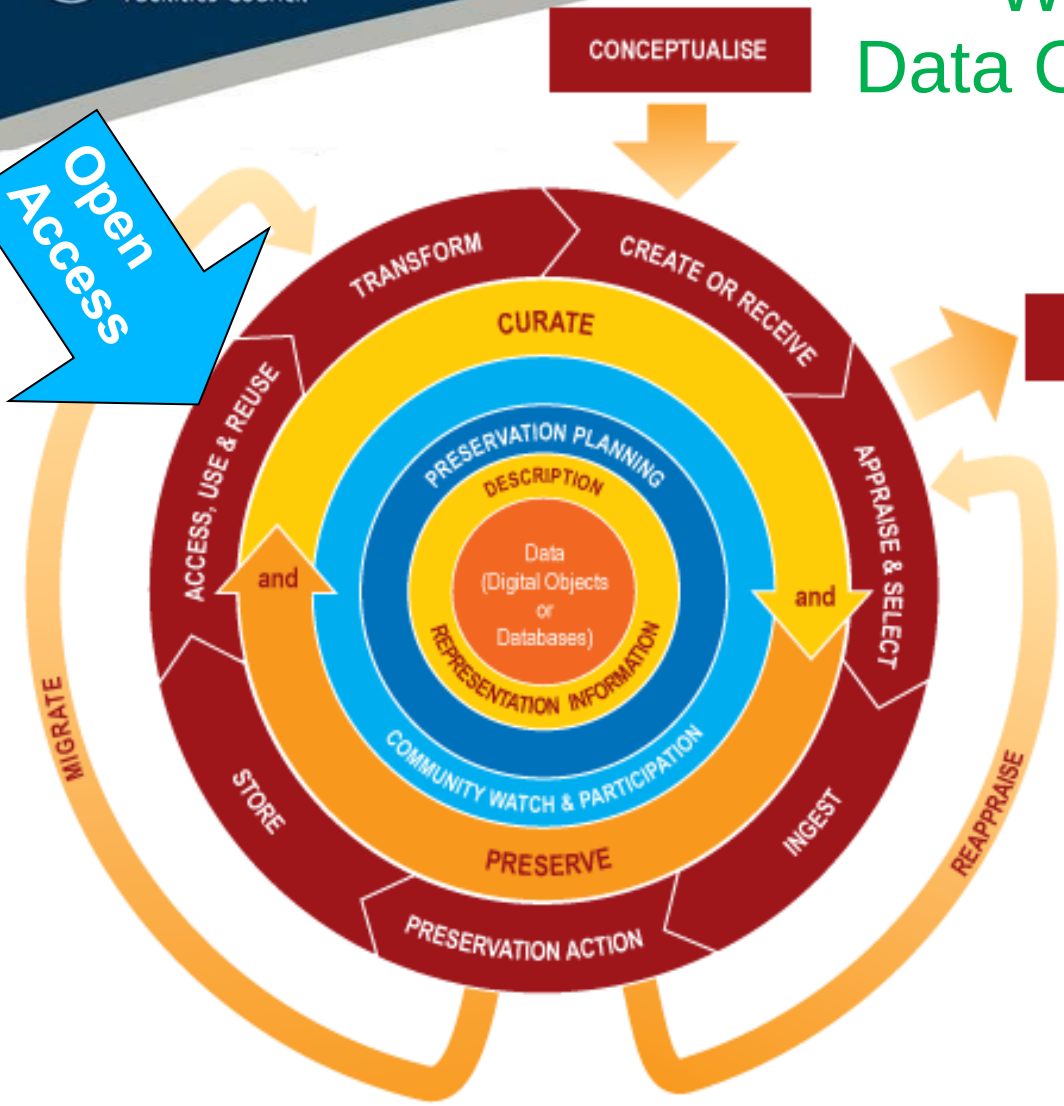
Unless there's a very good reason not to!

Publishing = making something public after some formal process which adds value for the consumer:

e.g. peer review and provides commitment to persistence

Science & Technology Facilities Council

What do data centres do?
Data Curation Lifecycle Model

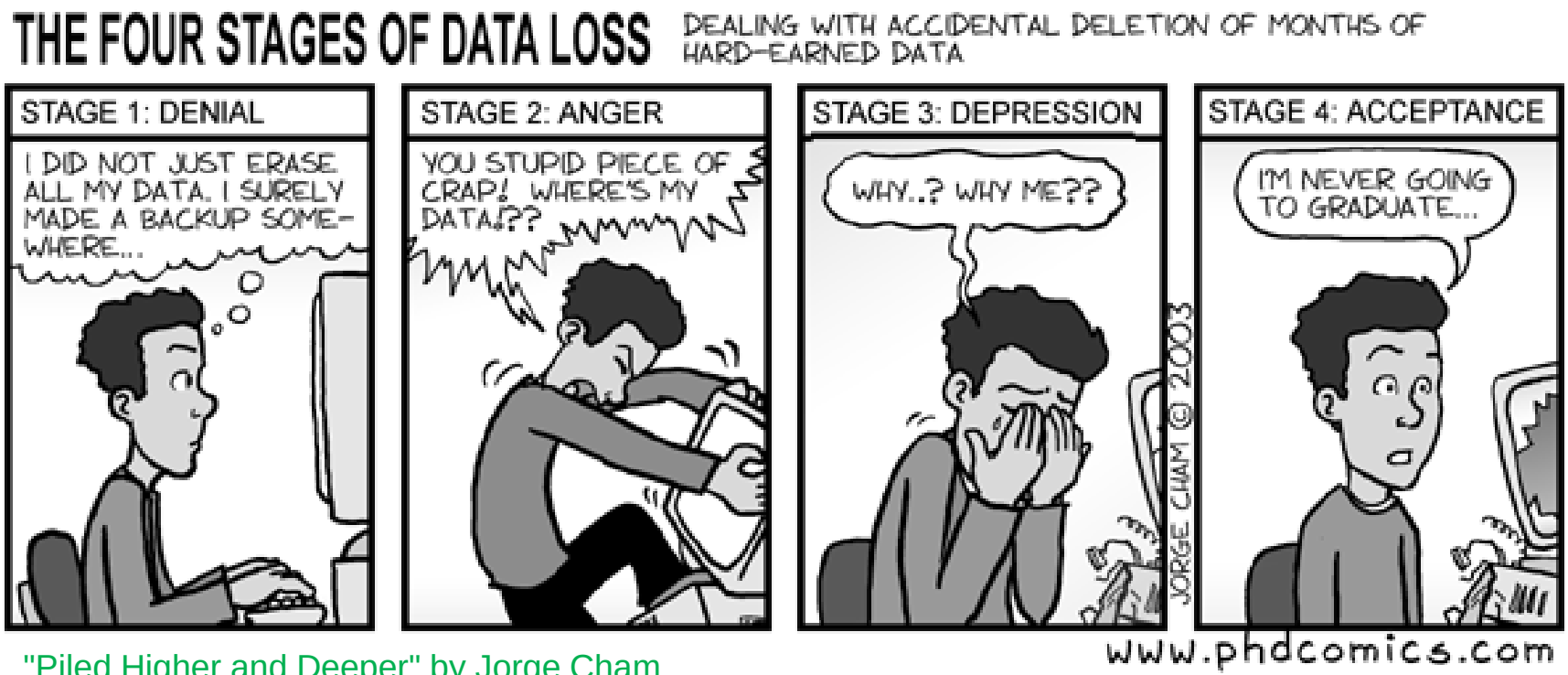


The Digital Curation Centre's Curation Lifecycle Model provides a graphical, high-level overview of the stages required for successful curation and preservation of data from initial conceptualisation or receipt through the iterative curation cycle.

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

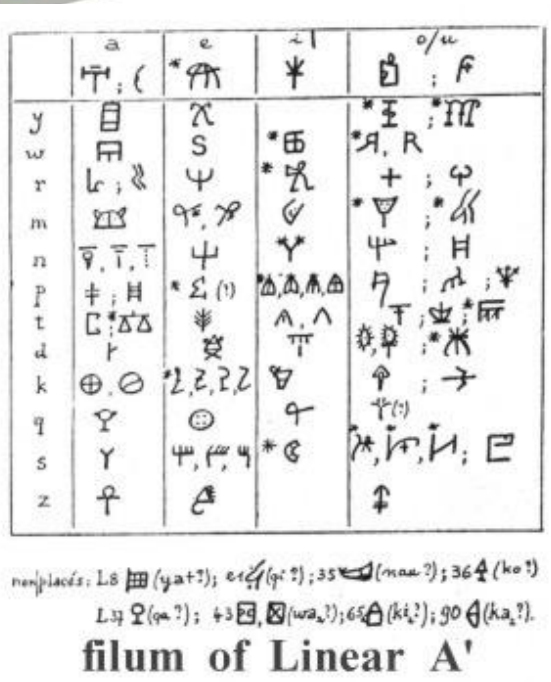
Science & Technology Facilities Council

Why should I bother putting my data into a repository?



"Piled Higher and Deeper" by Jorge Cham
www.phdcomics.com

It's ok, I'll just do regular backups



Phaistos Disk, 1700BC

These documents have been preserved for thousands of years!
But they've both been translated many times, with different meanings each time.

Data **Preservation is not enough,** we need **Active Curation to preserve Information**

Open is not enough!

"When required to make the data available by my program manager, my collaborators, and ultimately by law, I will grudgingly do so by placing the raw data on an FTP site, named with UUIDs like 4e283d36-61c4-11df-9a26-edddf420622d. I will under no circumstances make any attempt to provide analysis source code, documentation for formats, or any metadata with the raw data. When requested (and ONLY when requested), I will provide an Excel spreadsheet linking the names to data sets with published results. This spreadsheet will likely be wrong -- but since no one will be able to analyze the data, that won't matter."

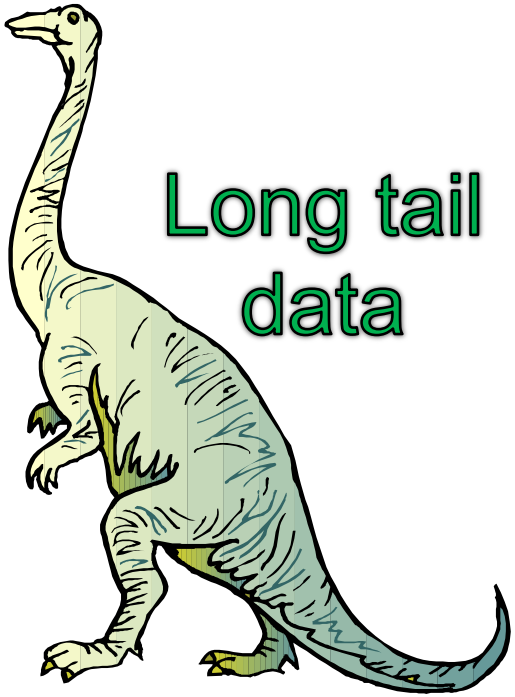
- http://ivory.idyll.org/blog/data-management.html

https://flic.kr/p/awnCQu

Science & Technology Facilities Council

BIG DATA

Versus

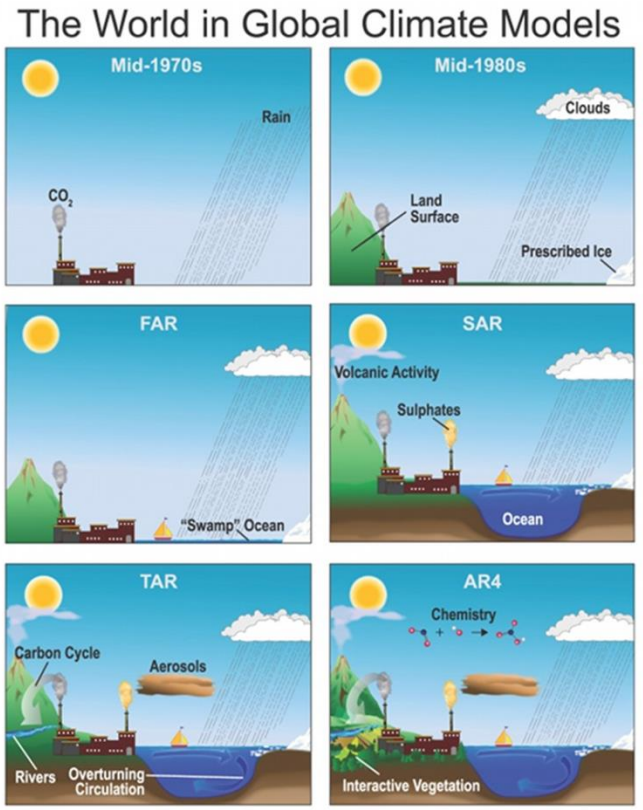


Long tail data

Example Big Data: CMIP5

CMIP5: Fifth Coupled Model Intercomparison Project

- Global community activity under the World Meteorological Organisation (WMO) via the World Climate Research Programme (WCRP)
- Aim:
  - to address outstanding scientific questions that arose as part of the 4th Assessment Report process,
  - improve understanding of climate, and
  - to provide estimates of future climate change that will be useful to those considering its possible consequences.



Many distinct experiments, with very different characteristics, which influence the configuration of the models, (what they can do, and how they should be interpreted).

CMIP5 numbers!

Simulations:

~90,000 years

~60 experiments

~20 modelling centres (from around the world) using

~30 major(*) model configurations

~2 million output “atomic” datasets

~10's of petabytes of output

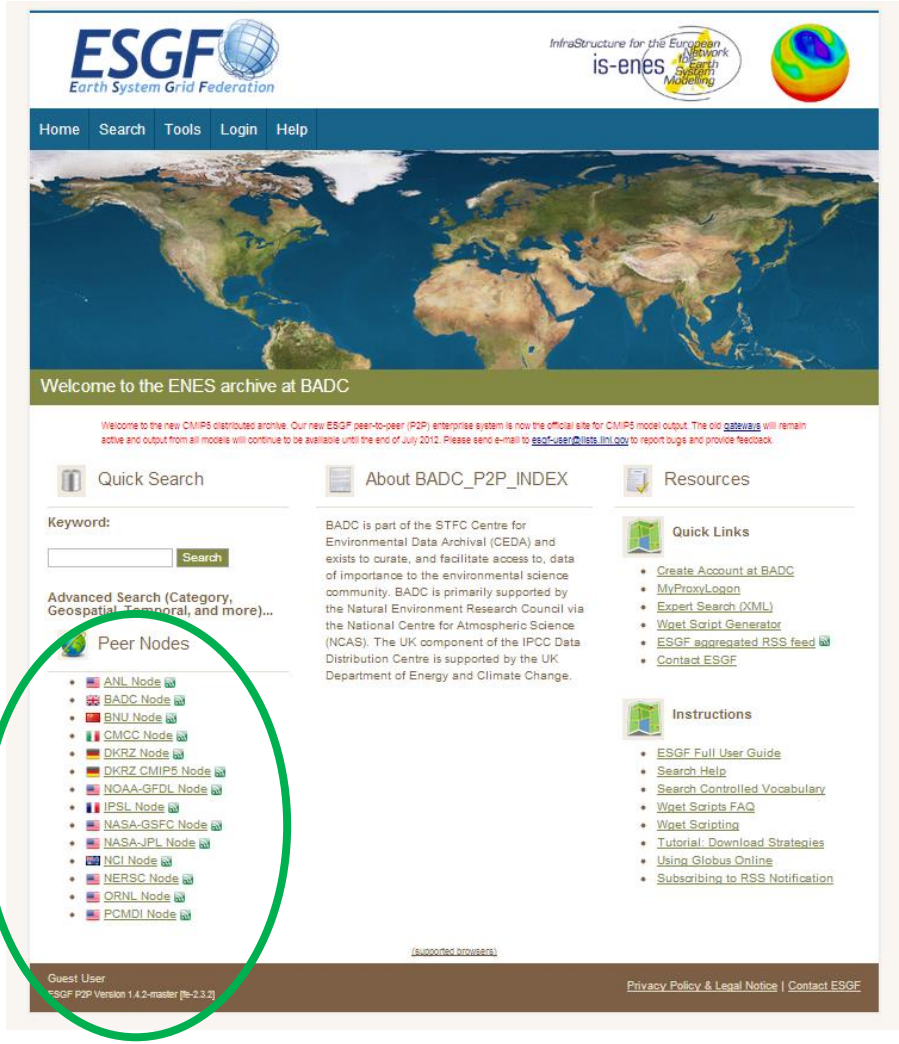
~2 petabytes of CMIP5 requested output

~1 petabyte of CMIP5 “replicated” output

Which are replicated at a number of sites (including ours)

Major international collaboration!

Funded by EU FP7 projects (IS-ENES, Metafor) and US (ESG) and other national sources (e.g. NERC for the UK)



Summary of the CMIP5 example

The Climate problem needs:

- Major physical e-infrastructure (networks, supercomputers)
- Comprehensive information architectures covering the whole information life cycle, including annotation (particularly of quality)

  … and hard work populating these information objects, particularly with provenance detail.
- Sophisticated tools to produce and consume the data and information objects
- State of the art access control techniques

Major distributed systems are social challenges as much as technical challenges.

CMIP5 is Big Data, with lots of different participants and lots of different technologies.

It also has a community willing to work together to standardise and automate data and metadata production and curation, and with the willingness to support the effort needed for openness.

https://flic.kr/p/g1EHPR

Big Data:

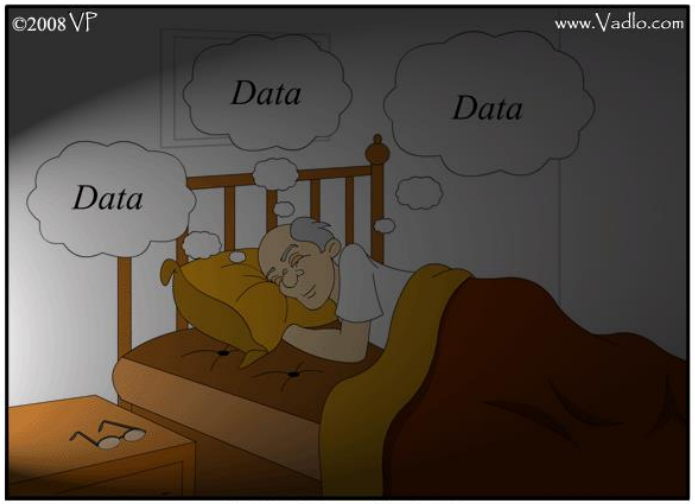
- Industrialised and standardised data and metadata production
- Large groups of people involved
- Methods for making the data open, attribution and credit for data creation established

Long Tail Data:

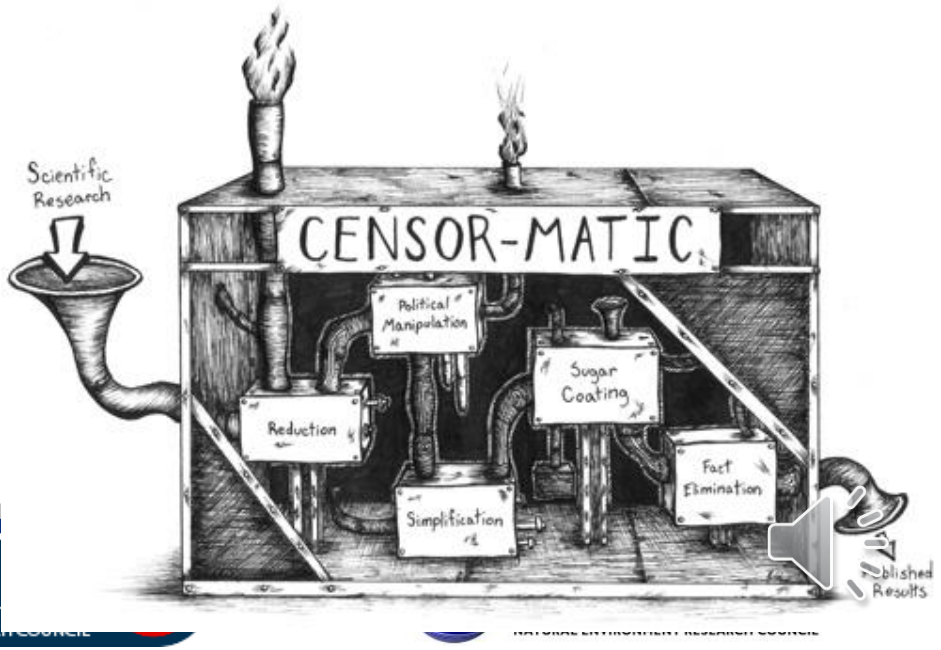
- Bespoke data and metadata creation methods
- Small groups/lone researchers
- No generally accepted methods for attribution and credit for data creation. Often data is closed due to lack of effort to open it

Summary and maybe conclusions?

- Data is important, and becoming more so for a wider range of the population
- Conclusions and knowledge are only as good as the data they're based on
- Science is supposed to be reproducible and verifiable
- It's up to us as scientists to care for the data we've got and ensure that the story of what we did to the data is transparent
  - So we and others can use the data again
  - And so people will trust our results



The Not-So-Secret life of a PI.



"Publishing research without data is simply advertising, not science" - Graham Steel

http://blog.okfn.org/2013/09/03/publishing-research-without-data-is-simply-advertising-not-science/

Thanks!

Any questions?

sarah.callaghan@stfc.ac.uk

@sorcha_ni

http://citingbytes.blogspot.co.uk/

Presentation funded by the European Commission as part of the project OpenAIREplus (FP7-INFRA-2011-2, Grant Agreement no. 283595)

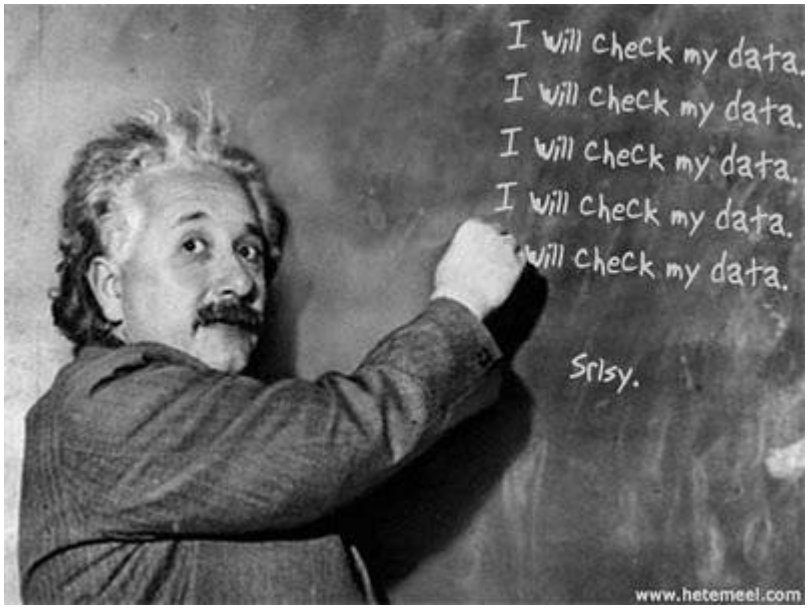


Image credit: Borepatch http://borepatch.blogspot.com/2010/06/its-not-what-you-dont-know-that-hurts.html