

Open Research Data

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

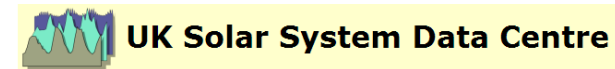


Who are we and why do we care about data?

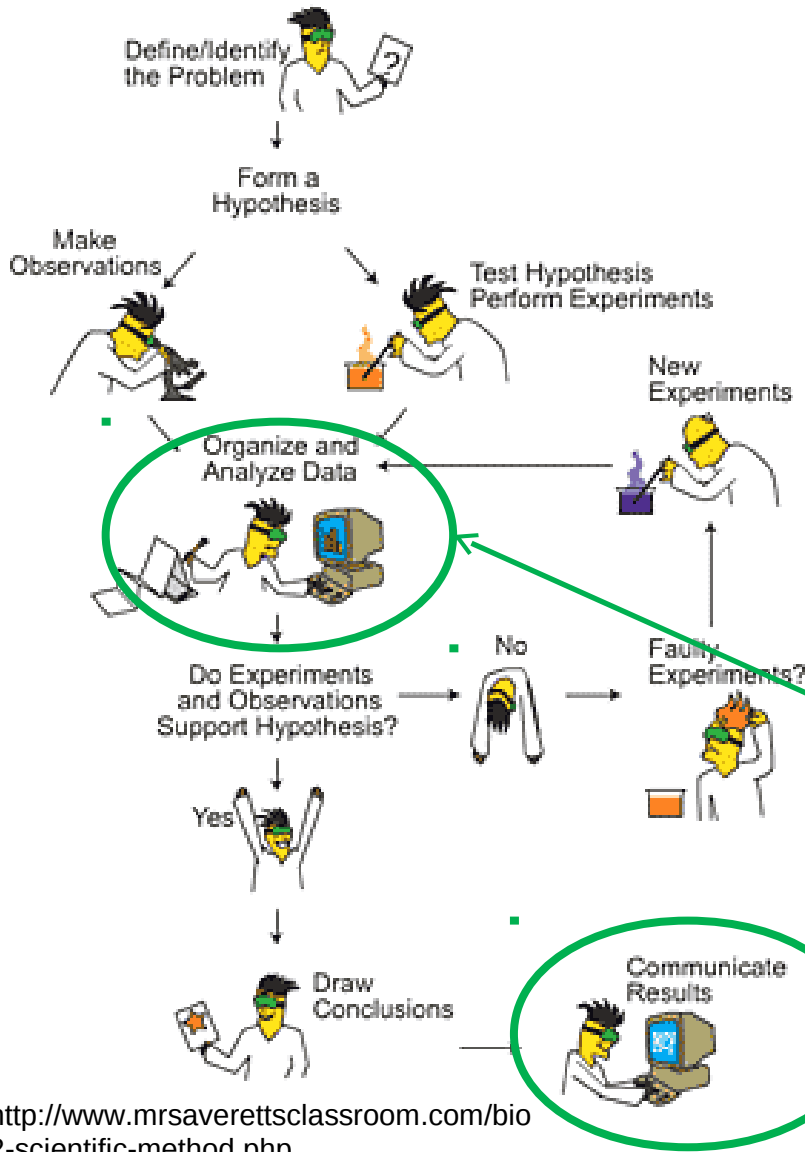
The UK's Natural Environment Research Council (NERC) funds six data centres which between them have responsibility for the long-term management of NERC's environmental data holdings.

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



The Scientific Method



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.

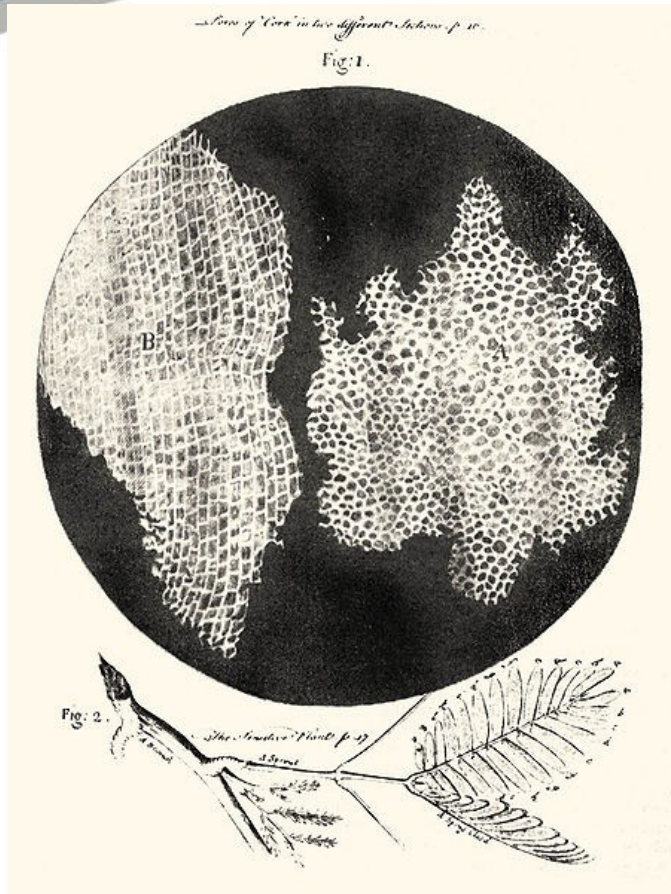
Unfortunately observational data is not reproducible (unless you have a time machine!)

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.

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

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	203 42	3		
	n b	153 30	3		
	a d	d = (12.13)	323 51	3		
	N d	377 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		
May 3.	a f	335 23	3		
	a g	g = (12.13)	215 17	3	115.5	4
	a h	h = (12.13)	193 29	3		
	g h	87 5	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	335.2	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?

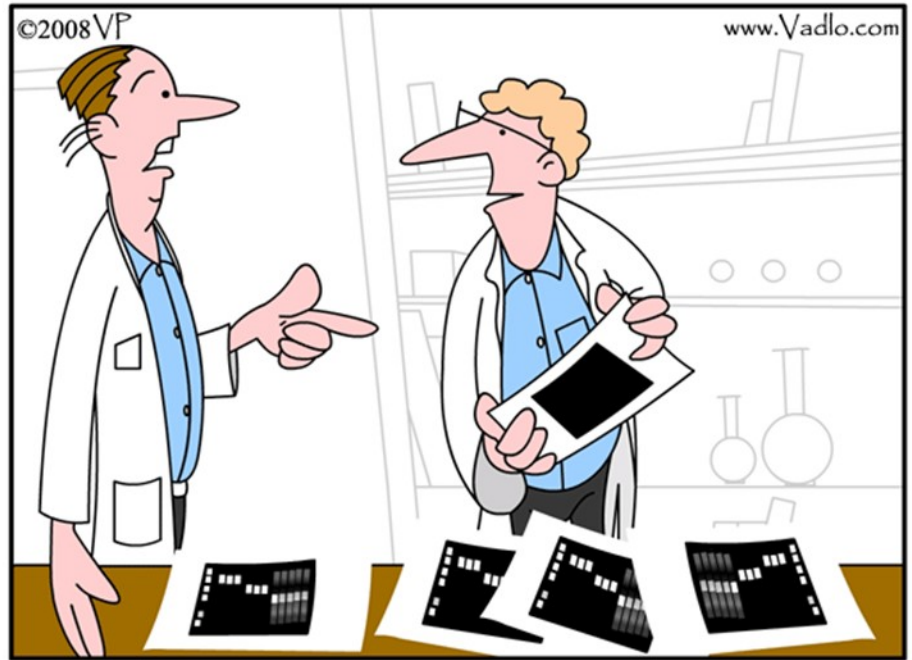
- **Pressure** from (UK) **government** to make data from publicly funded research available for free.
 - **Scientists** want **attribution** and **credit** for their work
 - **Public** want to know what the scientists are doing
 - Good for the **economy** if new industries can be built on scientific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



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

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!



AAAARGH!

What's wrong?

Hey, I remember this! I tried it over a year ago but it gave me negative data. Since I couldn't publish, I just filed it away and never told anyone.


What's wrong? I've been testing this stupid protocol for over a YEAR now AND IT DOESN'T WORK! I've wasted my entire grant and I have nothing to show for it!

Let me see that.

(c) The Upturned Microscope

**NEGATIVE DATA IS STILL DATA.
PUBLISH IT.
(For everyone's sake)**

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. At spatial inhomogeneity, which can be taken advantage of to improve the av 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 methc also been shown to produce the same statistics as real rain fields.

In this paper, we will demonstrate a typical discrete cascade model, h 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 divi

Home > Earth Sciences > General & Introductory Earth Sciences > Geoscience Data Journal > Vol 1 Issue 1 > Abstract

JOURNAL TOOLS

- Get New Content Alerts
- Get RSS feed
- Save to My Profile
- Recommend to Your Librarian

JOURNAL MENU

- Journal Home
- FIND ISSUES
 - Current Issue
 - All Issues
- FIND ARTICLES
 - Early View
- FOR CONTRIBUTORS
 - Author Guidelines
 - Submit an Article
- ABOUT THIS JOURNAL
 - Society Information
 - Overview
 - Editorial Board
 - Permissions
- SPECIAL FEATURES
 - Data Center FAQs
 - Open Access License and Copyright
 - Author FAQs
 - Article Publication Charges
 - Wiley Open Access
 - Institutional and Funder Payments
 - Guidelines for Reviewers
 - Guidelines for Repositories
 - L F Richardson Prize
 - Royal Meteorological Society Journal Awards

RMetS Geoscience Data Journal Open Access

Data Paper

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

S. A. Callaghan¹, 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

© 2013 The Authors. *Geoscience Data Journal* published by John Wiley & Sons Ltd and Royal Meteorological Society.

This is an open access article under the terms of the [Creative Commons Attribution License](#), which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

Am score 10

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.

Abstract | Article | References | Cited By

Enhanced Article (HTML) | Get PDF (1849K)

Keywords:
site diversity; radio propagation; fade mitigation techniques

Jump to...

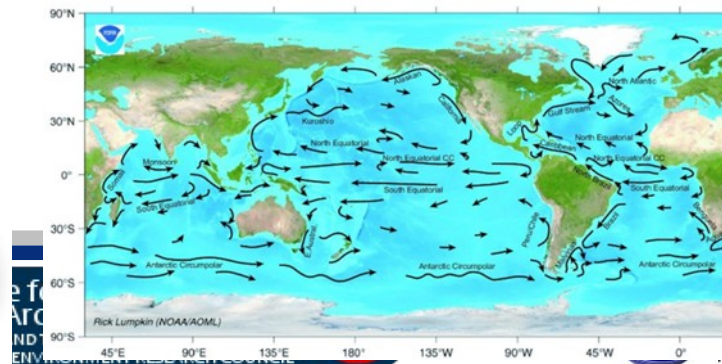
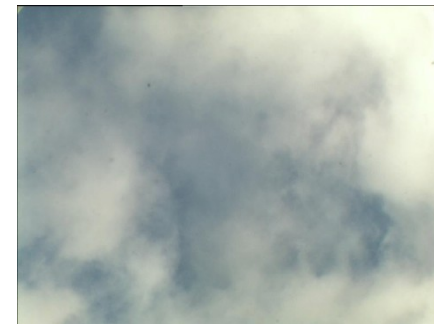
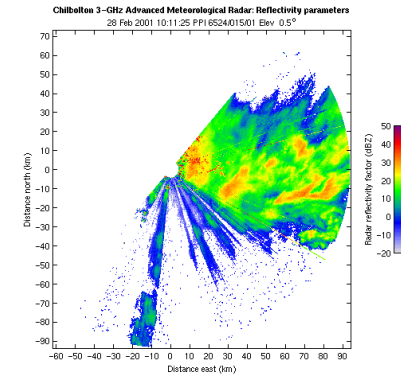
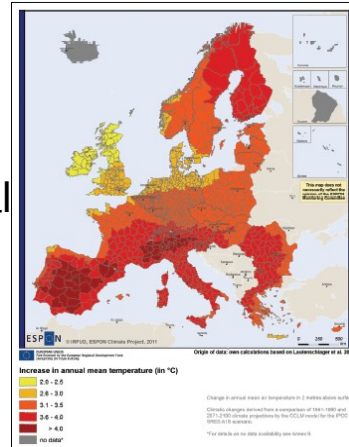
Abstract

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.

Be the first to know when a new issue of WEATHER is out!

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



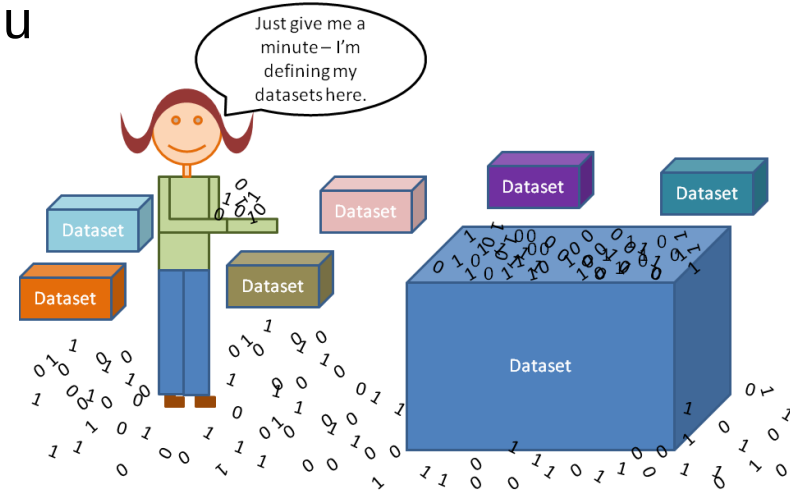
What is a Dataset?

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.

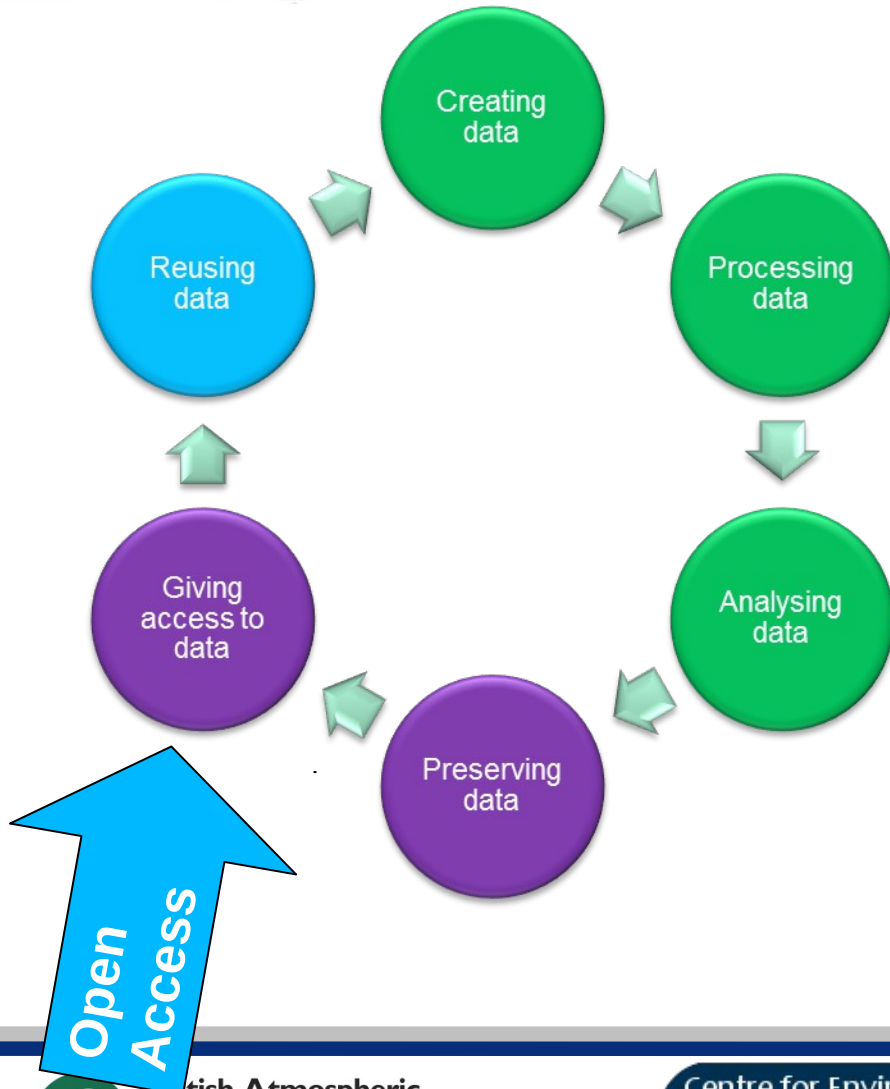


The research data lifecycle

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.

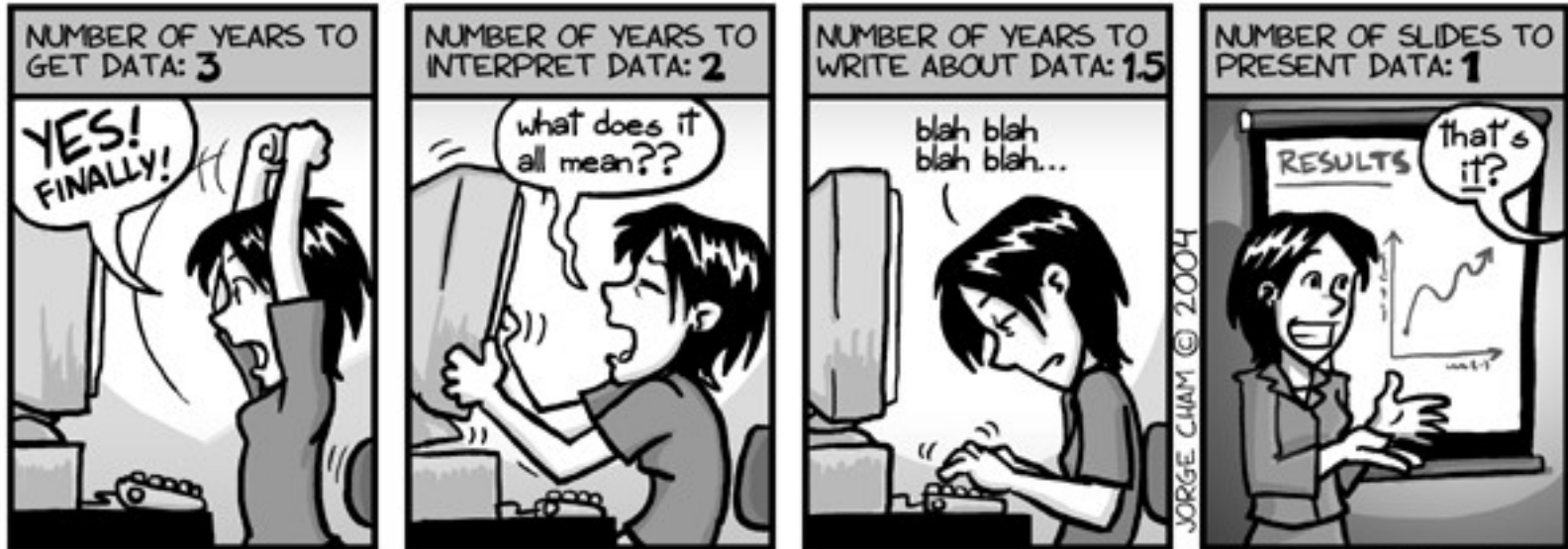


See <http://data-archive.ac.uk/create-manage/life-cycle> for more detail



Creating a dataset is hard work!

DATA: BY THE NUMBERS



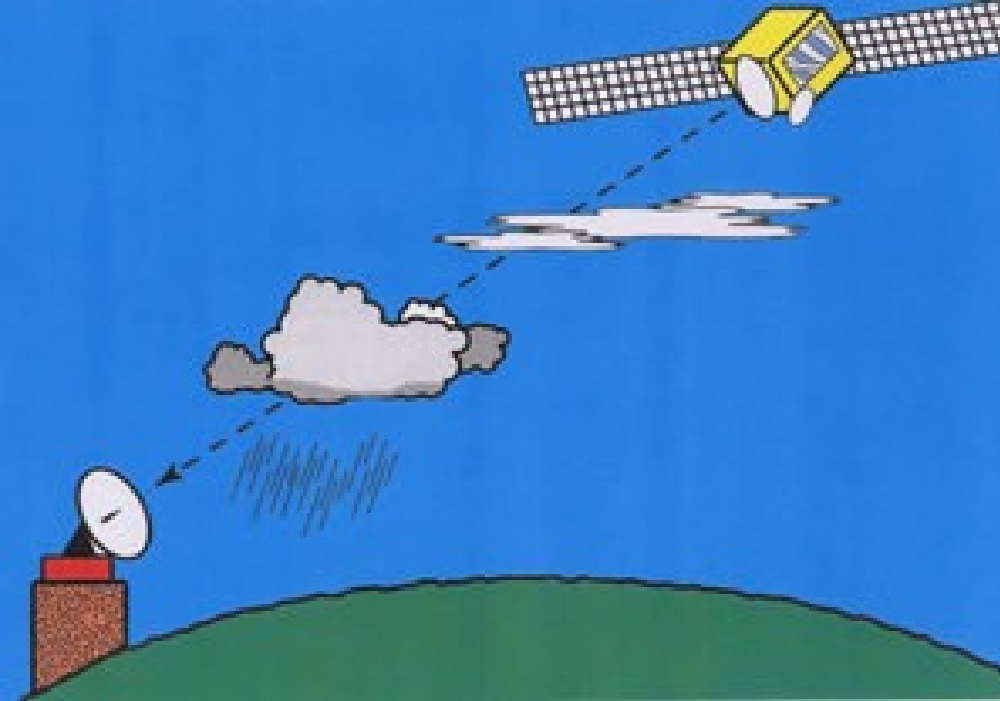
www.phdcomics.com

"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



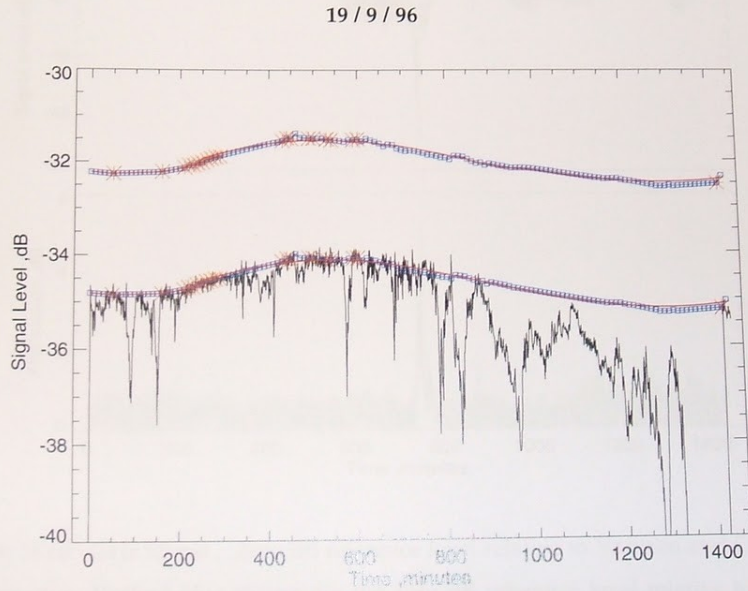
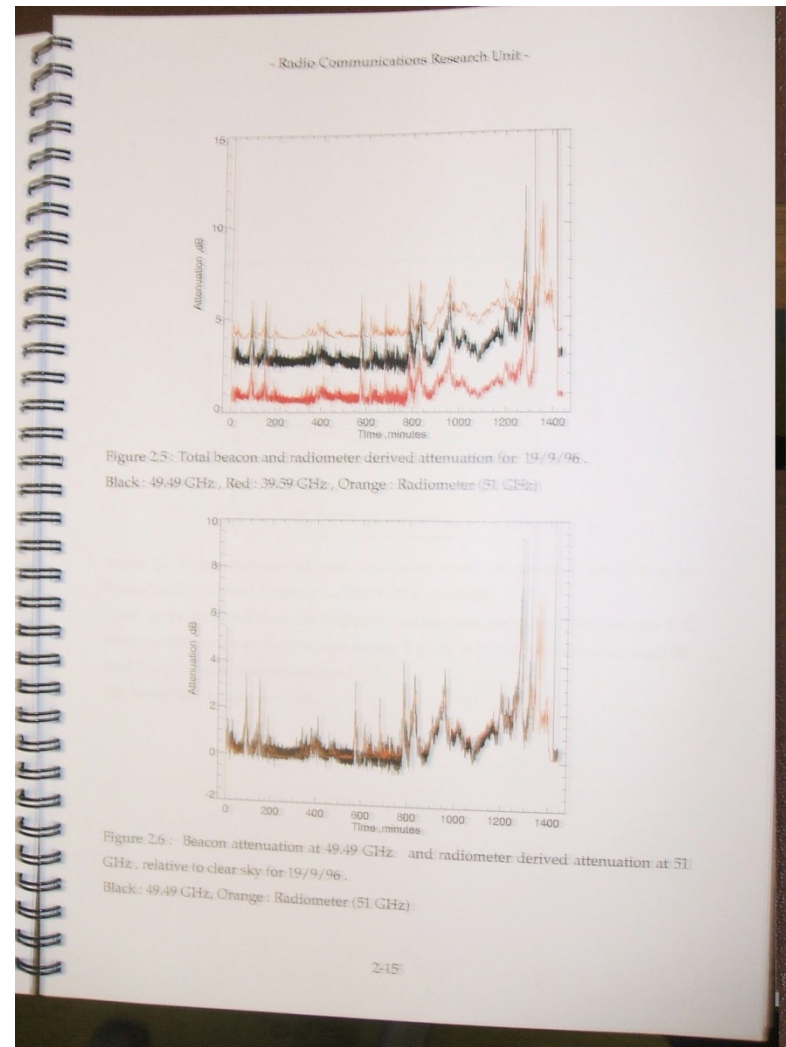


Figure 2.3 : 49.49 GHz Signal and Zero dB reference level relative to Vacuum and Clear sky.(Red : Fourier series , Blue :Linear interpolation , * : Signal level values at clear sky time points).

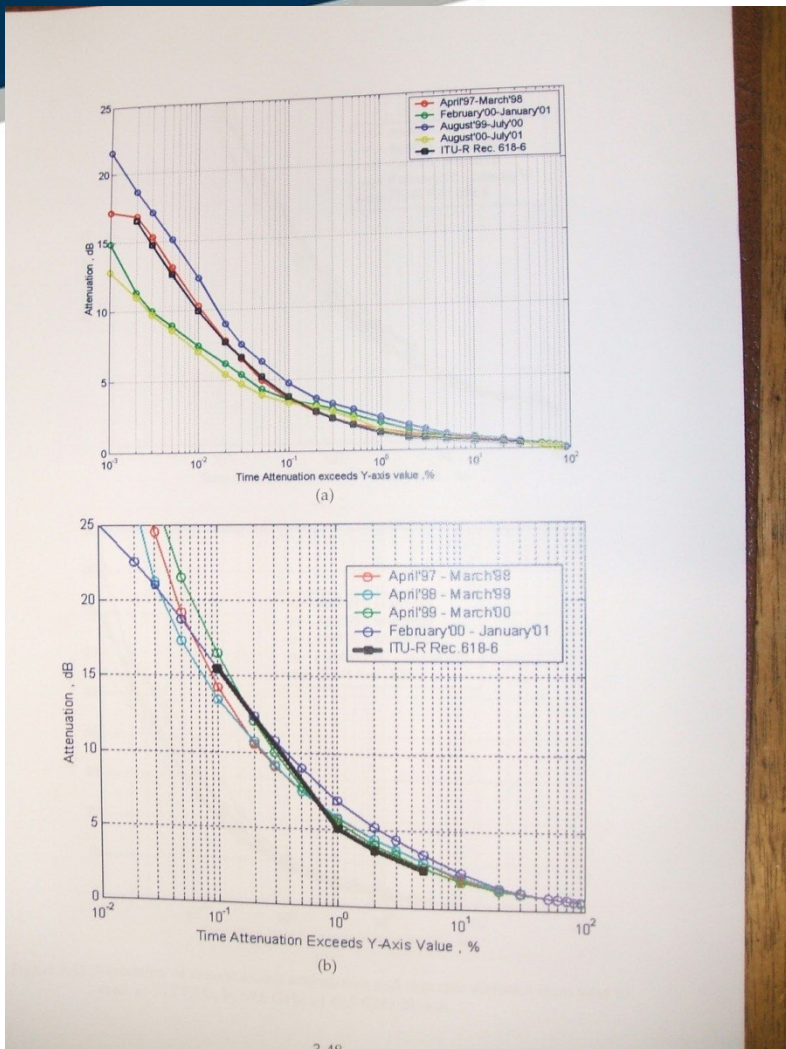
One day's worth of raw data from one of the receivers
My job was to take this...



...turn it into this....



Analysing data



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

Each month of preprocessed 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.

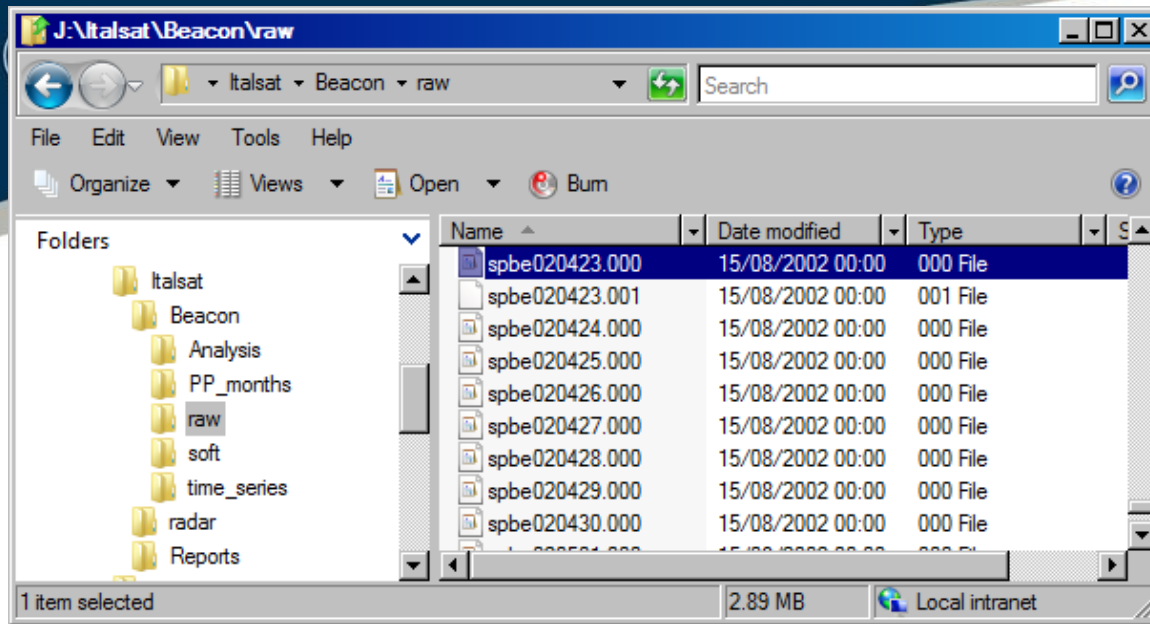
...with the final result being this.



Preserving data (the wrong way!)



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

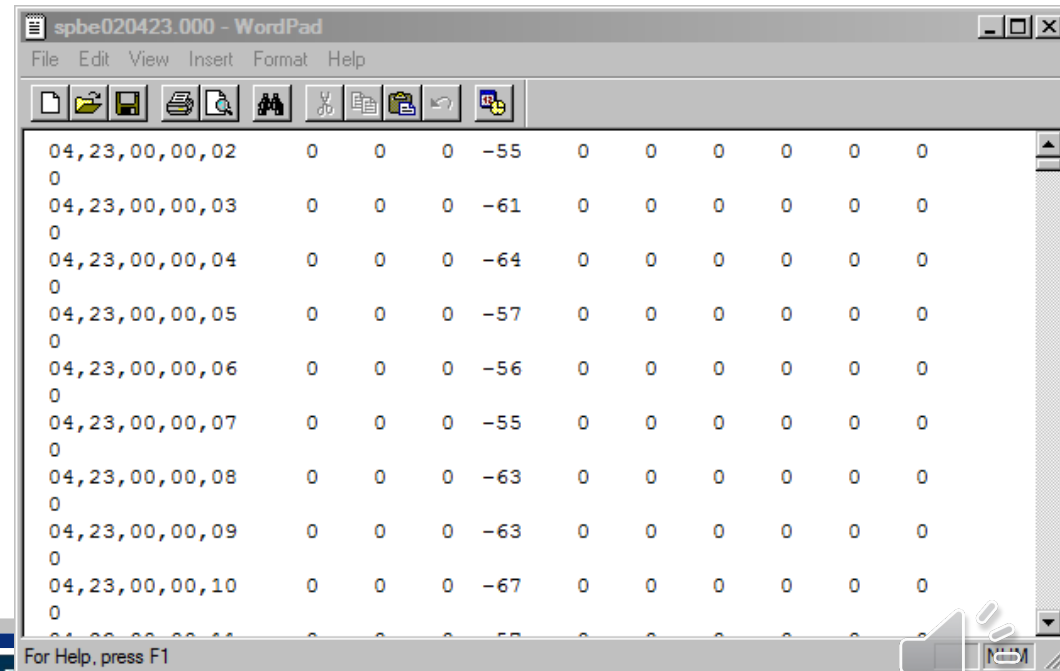


What the processed data set looks like on disk

What the raw data files looked like.

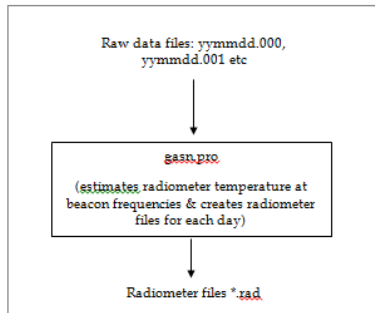
(I do have some Word documents somewhere which describe what all this is...)

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

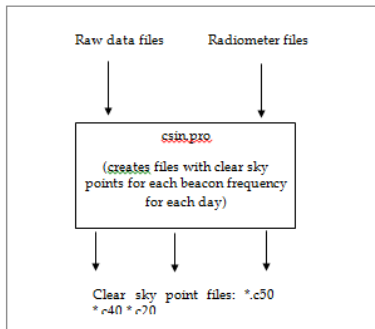


ITALSAT pre-processing flowchart

Step 1

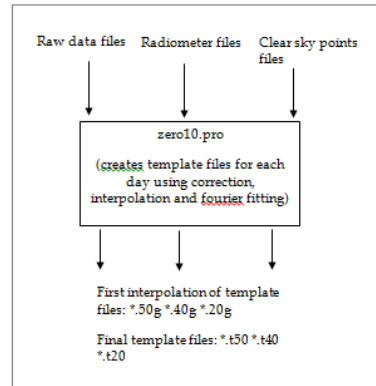


Step 2

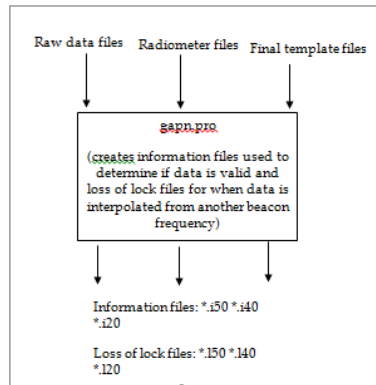


1

Step 3



Step 4



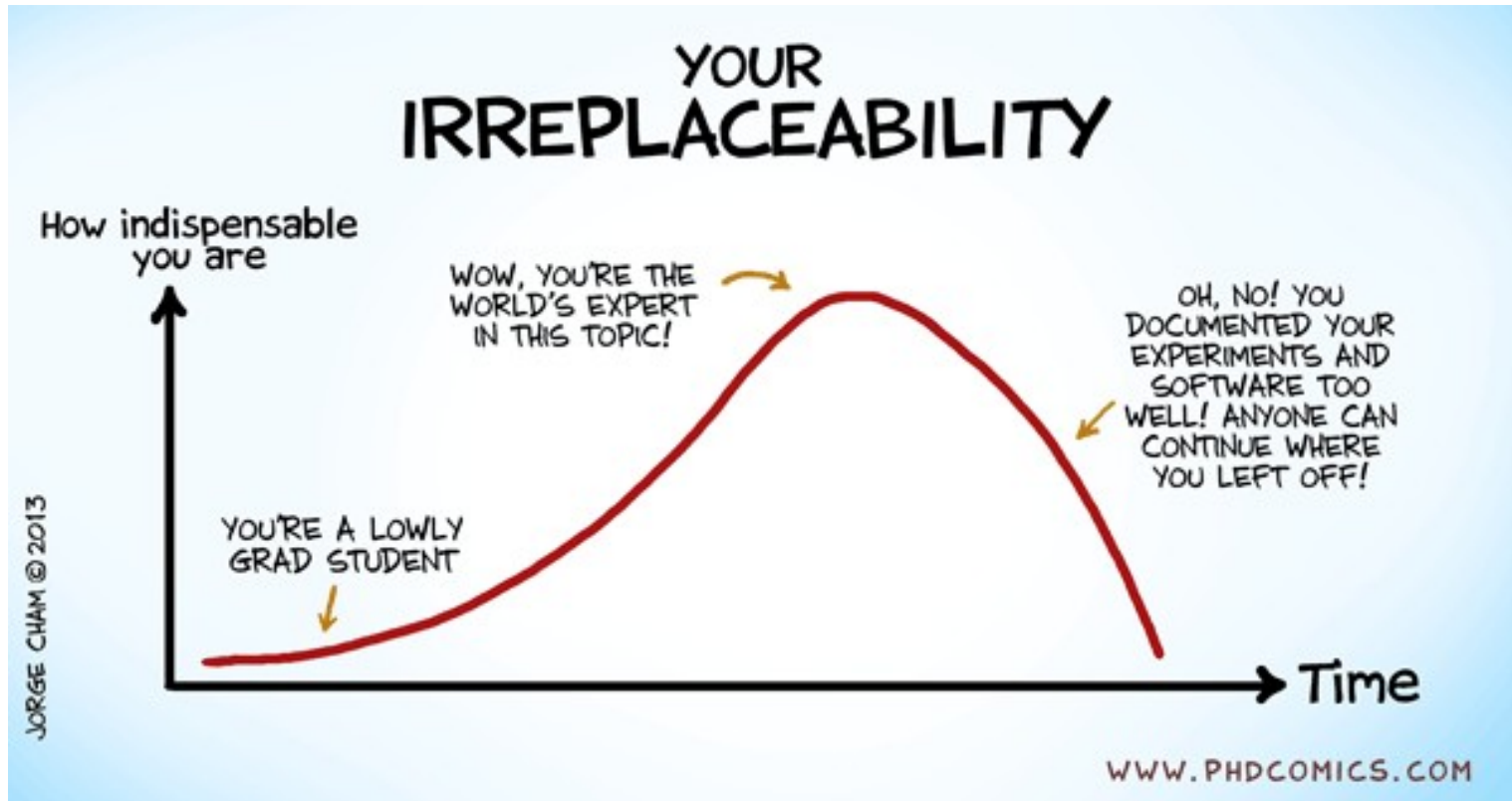
2

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

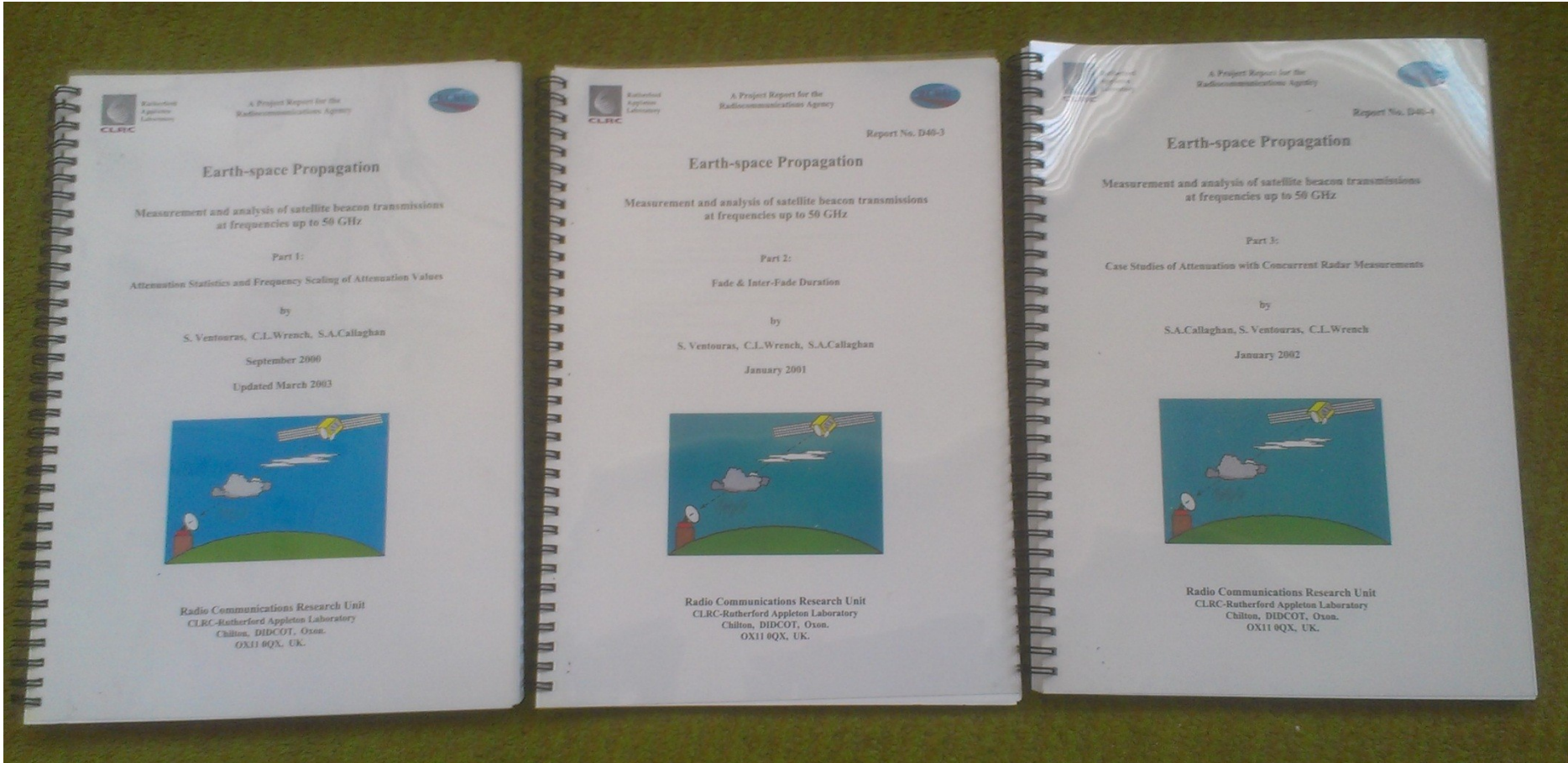


Documentation can sometimes produce mixed feelings



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

Publications – grey literature



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

Where's the data?

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. Wrench¹

Received 9 February 2005; revised 9 December 2005; accepted 15 February 2006

[1] Long-term statistics of tropospheric attenuation measurements made in the south of England at 49.5, 39.6, and 18.7 GHz; coincident rainfall at the receiving ground station. A method to remove beacon signals and to establish the reference level of total attenuation has been presented in detail. The total attenuation has been estimated to be $\sim \pm 0.5$ dB. A new method for estimating statistics has been proposed and validated against measurements at 18.7, 39.6, and 49.5 GHz. For both locations, the predictions compared with the established International Telecommunication Union recommendation method. A significant month-long increase in the attenuation and rainfall statistics and should be taken into account in the design and use of future slant path systems. The statistics are subject to diurnal variations; however, for both locations, they seem to follow a particular pattern.

Citation: Ventouras, S., and C. L. Wrench (2006), Long-term statistics of tropospheric attenuation from the ITALSAT satellite experiment in the United Kingdom, *Radio Science*, 41, RS2007, doi:10.1029/2005RS003252.

RS2007

VENTOURAS AND WRENCH: TROPOSPHERIC ATTENUATION

RS2007

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

Outage, %	Total Attenuation, dB								
	49.5 GHz			39.6 GHz			18.7 GHz		
	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

^aFor 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:

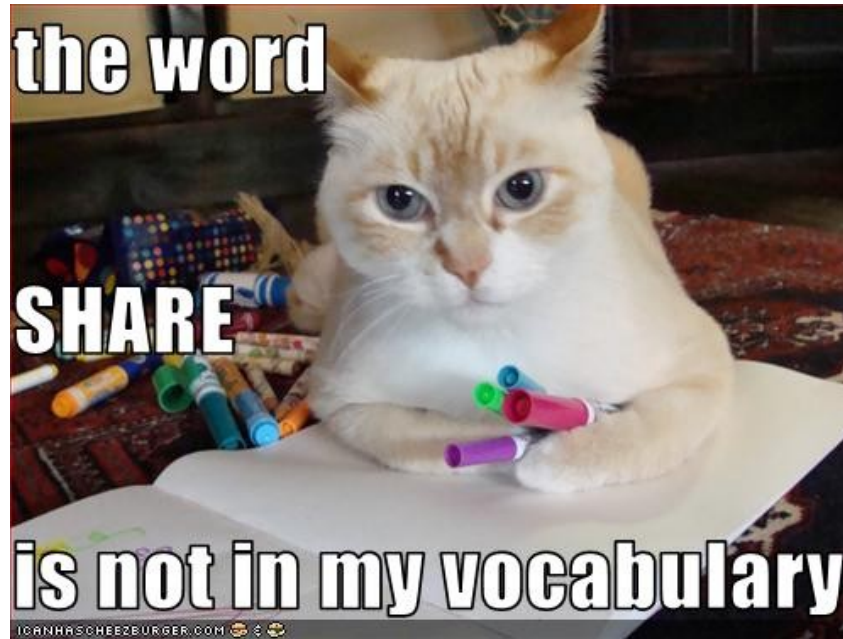


Composite image from Flickr user bnilsen and Matt Stempeck (NOI), shared under Creative Commons license

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.)

Viewing ITALSAT radio pr: x

badc.nerc.ac.uk/view/badc.nerc.ac.uk_ATOM_ACTIVITY_f2984bd6-a664-11e1-ac44-00163e251233

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

Search for in **All**

ITALSAT radio propagation measurements at 50GHz in the United Kingdom

General Info

Title: ITALSAT radio propagation measurements at 50GHz in the United Kingdom
Type: Activity
Sub-Type: Deployment
Publication State: Citable
URI: http://badc.nerc.ac.uk/view/badc.nerc.ac.uk_ATOM_ACTIVITY_f2984bd6-a664-11e1-ac44-00163e251233

Summary
Measurements of tropospheric attenuation (excess and total) made at Sparsholt in Hampshire, UK using the ITALSAT satellite F1 beacon signal at 50 GHz.

Content
Tropospheric attenuation measurements made at Sparsholt in Hampshire, UK using the ITALSAT satellite F1 beacon signal at 49.5 GHz. ITALSAT F1 (owned and operated by the Italian Space Agency) was in geostationary orbit at 13 degrees east, and it could be seen from the receiving station at an elevation angle of 30 degrees. The received signal was vertically polarised and was sampled once a second. North-south tracking of the satellite with the beacon receiver maintained ~20dB of dynamic range throughout of the measurement period. The method applied to remove the nonatmospheric changes of the beacon signal and to establish the reference level from which to measure the excess and total attenuation is described in [Ventouras et al., Long-term statistics of tropospheric attenuation from the Ka/U band ITALSAT satellite experiment in the United Kingdom, Radio Sci., 41, RS2007, doi:10.1029/2005RS003252]. The accuracy of fade level retrieval is estimated to be +/-0.5dB

Author

Name	email
Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio Research, [S.Ventouras, S.A.Callaghan, C.L.Wrench]	spiros.ventouras@stfc.ac.uk

Online References

Relation	Title
Apply for access	Apply for Access: 50GHz data
Download	Data directory for ITALSAT 50GHz data
Documentation	doi:10.5285/597C906A-B09E-4822-8860-3B53EAF8C57F

Associated Data

Type	Title
Data Production Tool	ITALSAT 50GHz receiver
Activity	ITALSAT 50GHz Radio Propagation Experiment - UK
Observation Station	Sparsholt College, Hampshire, UK

Parent Activity
No data specified at present

Sub-Activities
None

Data Coverage

Spatial coverage		Temporal coverage	
Min X: -1.433	Max Y: 51.0667	Start Date: 1997-04-01	End Date: 2000-12-31
	Max X: -1.433		
	Min Y: 51.0667		

Spatial resolution
No spatial resolution information available.

Vertical extent
No vertical extent information available.

Links and Services
[XML](#) Downloadable XML version of this record

Not all information in this record may be rendered in this view. Please see the XML version for complete content

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

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

Get Data

[Home](#) [My BADC](#) [Data](#) [Search](#) [Community](#) [Help](#)

[Get Data](#) [Access Rules](#) [Submit Data](#) [Dataset Index](#)

Get Data

[Logout](#) [Help](#)

Username: *scallagh* **Download multiple files** **Depth:** 1

Current directory: / [badc](#) / [chilbolton](#) / [data](#) / [italsat](#)

Dataset: *Measurements from the Chilbolton Facility for Atmospheric and Radio Research (CFARR)* [Details](#)

- [italsat-20ghz-beacon](#)
- [italsat-40ghz-beacon](#)
- [italsat-50ghz-beacon](#)

[Home](#) [Contact](#) [Disclaimer](#) Last Modified: 05/17/2013 08:49:43

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!



<http://www.flickr.com/photos/nazlicetiner/6448303541/>



http://www.flickr.com/photos/maco_nix/5019885742/

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

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



<http://www.flickr.com/photos/lovefibre/3251690074/>



<http://www.flickr.com/photos/halfbisqued/8084145976/>



<http://www.flickr.com/photos/lucathegalga/2282305884/>



<http://www.flickr.com/photos/ujkakevin/2303531028/>



Metadata

It is generally agreed that we need methods to:

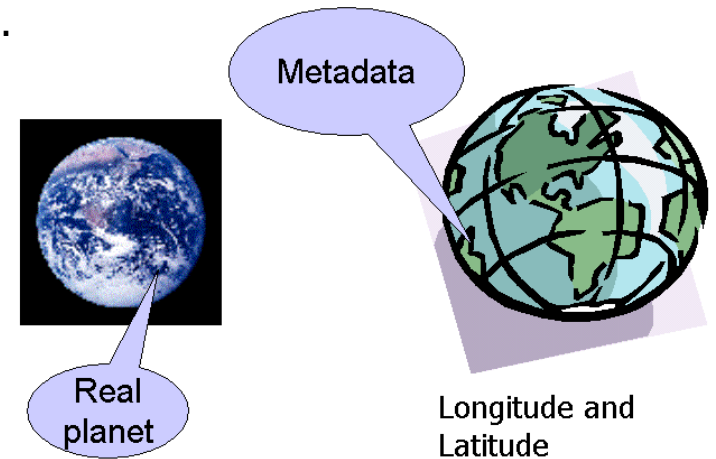
- define and document datasets of importance.
- augment and/or annotate data
- amalgamate, reprocess and reuse data

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

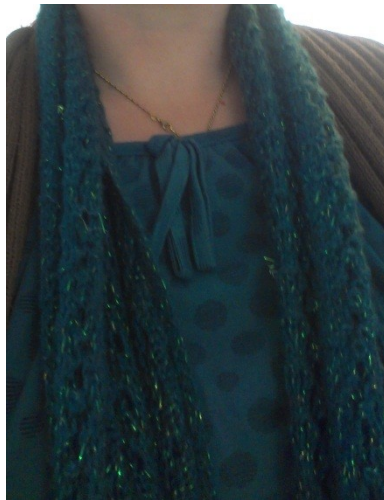
Metadata can often act as a surrogate for the real thing, in this case the planet.



- 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

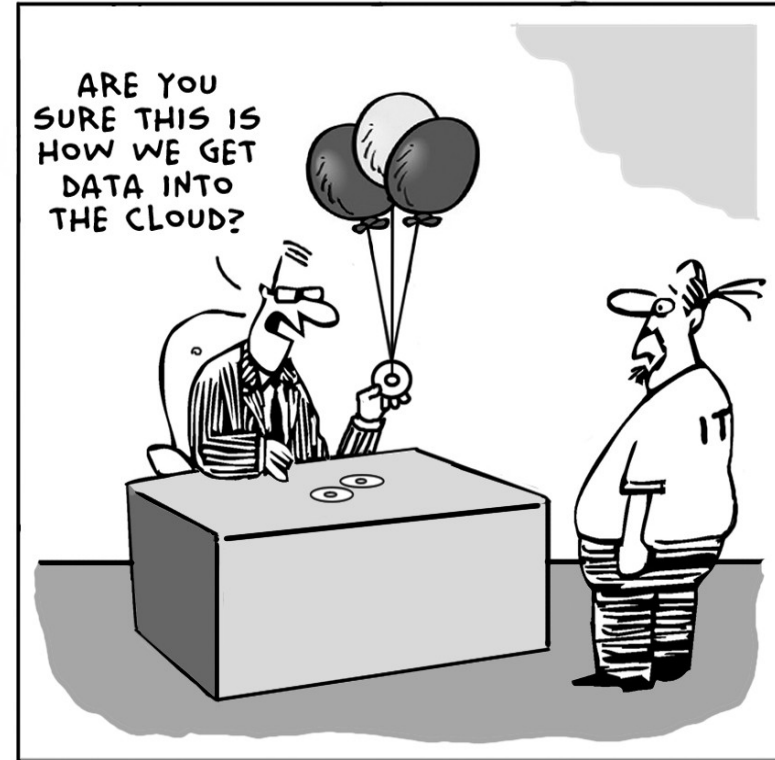


I can't make my scarf Open Access, but I can make the metadata about it open – enabling other users to create it for themselves.

Dataset views and suggested uses

How to publish data/make data open

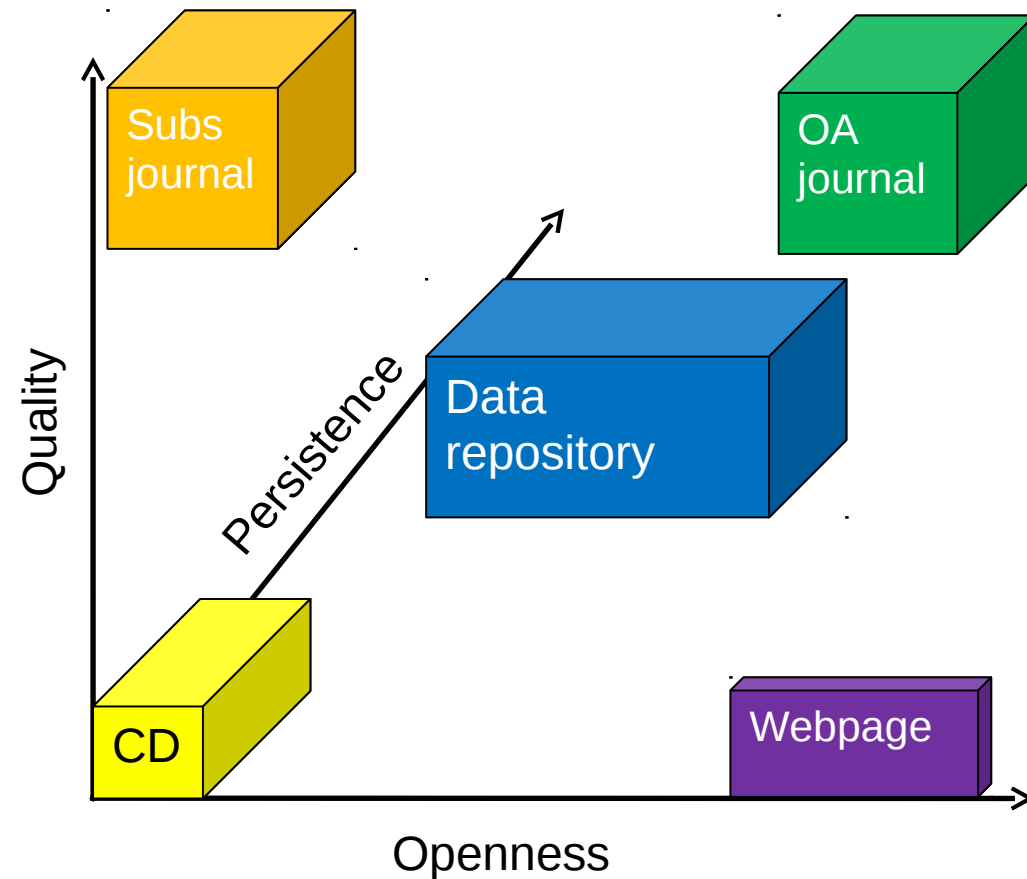
- Stick it up on a webpage somewhere
 - Issues with stability, persistence, discoverability...
 - Maintenance of the website
- Put it in the cloud
 - Issues with stability, persistence, discoverability...
- Attach it to a journal paper and store it as supplementary materials
 - Journals not too keen 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

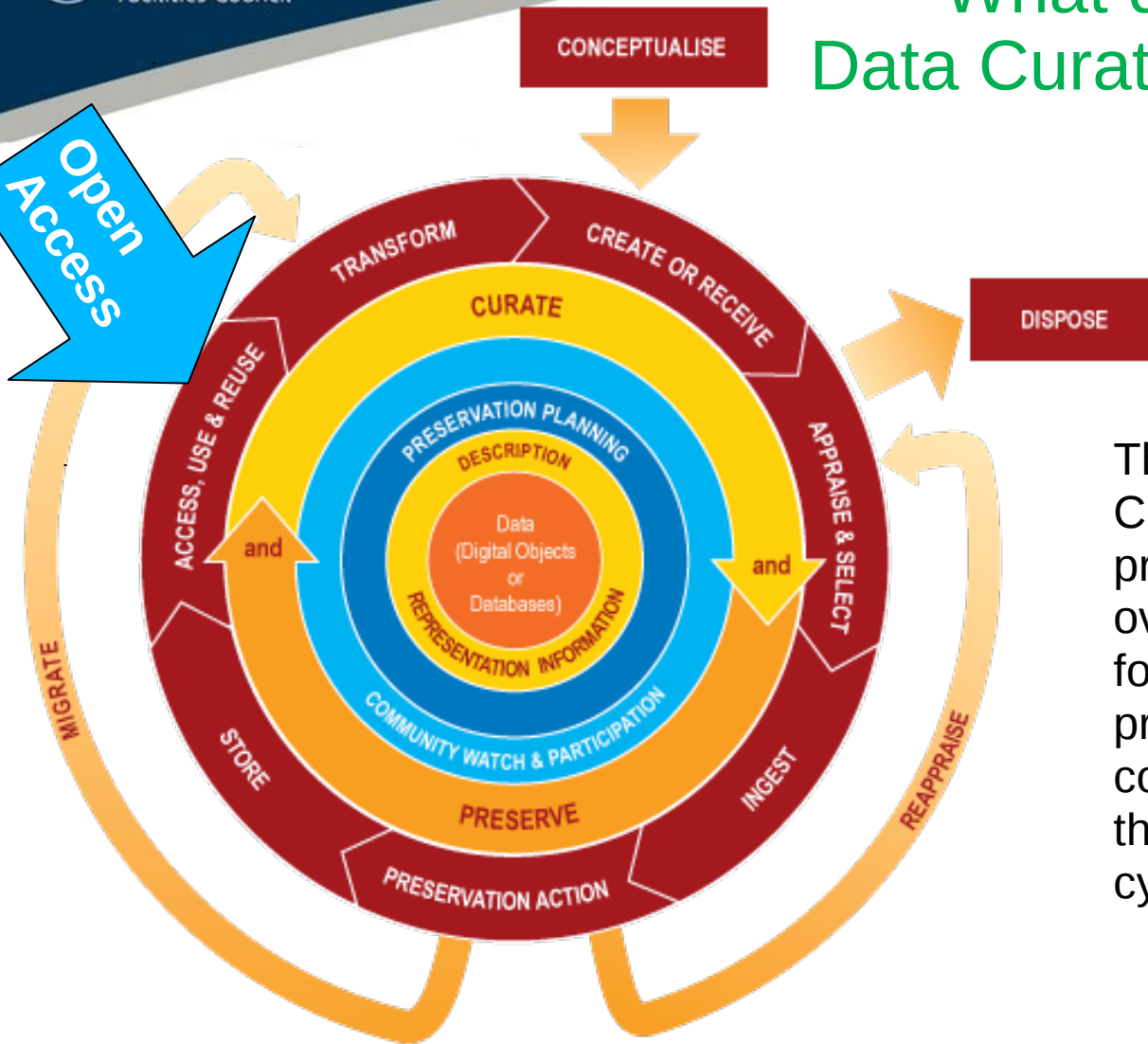
Unless there's a very good reason not to!

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

e.g. peer review **and** provides commitment to persistence



What do data centres do? Data Curation Lifecycle Model



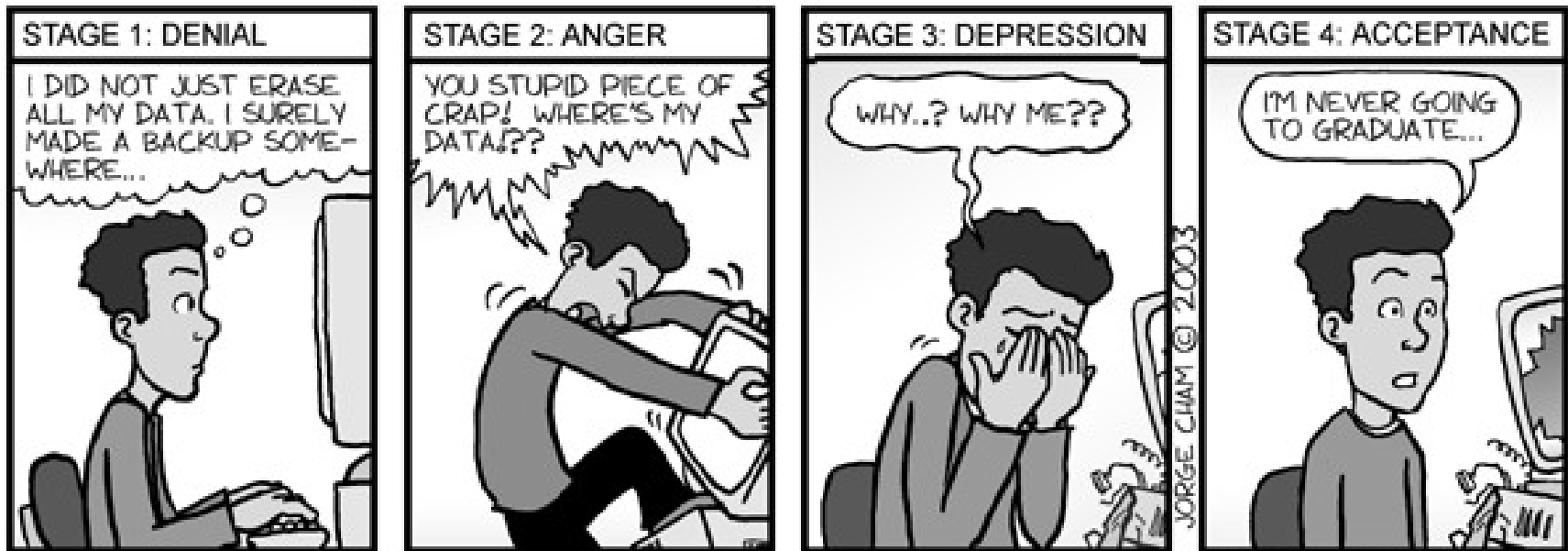
The Digital Curation Centre's Curation Lifecycle Model provides a graphical, high-level overview of the stages required for successful curation and preservation of data from initial conceptualisation or receipt through the iterative curation cycle.

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

Why should I bother putting my data into a repository?

THE FOUR STAGES OF DATA LOSS

DEALING WITH ACCIDENTAL DELETION OF MONTHS OF
HARD-EARNED DATA



www.phdcomics.com

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



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

	a	e	ι	ο/ω
	⊕; (*⊕	⊕	⊕; f
y	⊕	⊕	*⊕	*⊕; ⊕
w	⊕	S	*⊕	*⊕, R
r	⊕; ⊕	⊕	*⊕	+; ⊕
m	⊕	⊕; ⊕	⊕	*⊕; ⊕
n	⊕; ⊕, ⊕	⊕	⊕	⊕; H
p	⊕; H	*⊕ (⊕)	⊕, ⊕, ⊕	⊕; ⊕; ⊕
t	⊕; ⊕	⊕	⊕, ⊕	⊕; ⊕; ⊕
d	⊕	⊕	⊕	⊕; ⊕
k	⊕; ⊕	⊕, ⊕, ⊕	⊕	⊕; ⊕
q	⊕	⊕	⊕	⊕; ⊕
s	⊕	⊕, ⊕, ⊕	*⊕	*⊕, ⊕, ⊕; ⊕
z	⊕	⊕		⊕

non-placés: L.8 ⊕ (yat?); ει ⊕ (qi?); 35 ⊕ (mau?); 36 ⊕ (ko?)
L.3 ⊕ (qa?); 43 ⊕, ⊕ (wa?); 65 ⊕ (ki?); 90 ⊕ (ka?).

film of Linear A'



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.htm>



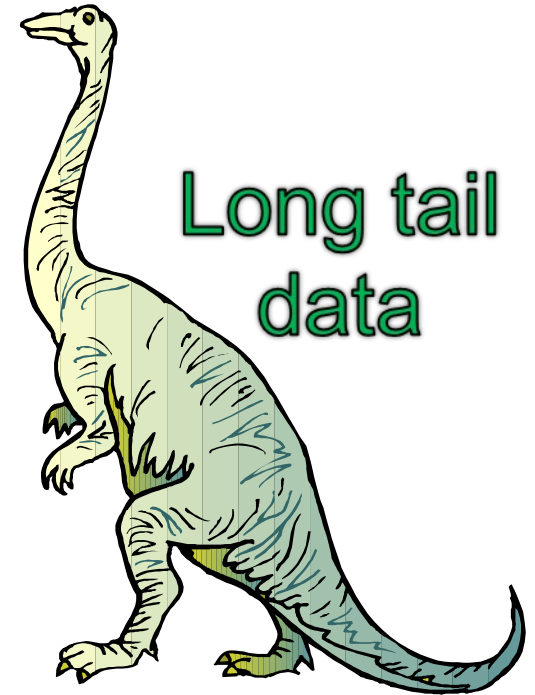
<https://flic.kr/p/awnCQu>



BIG
DATA

Versus

Long tail
data



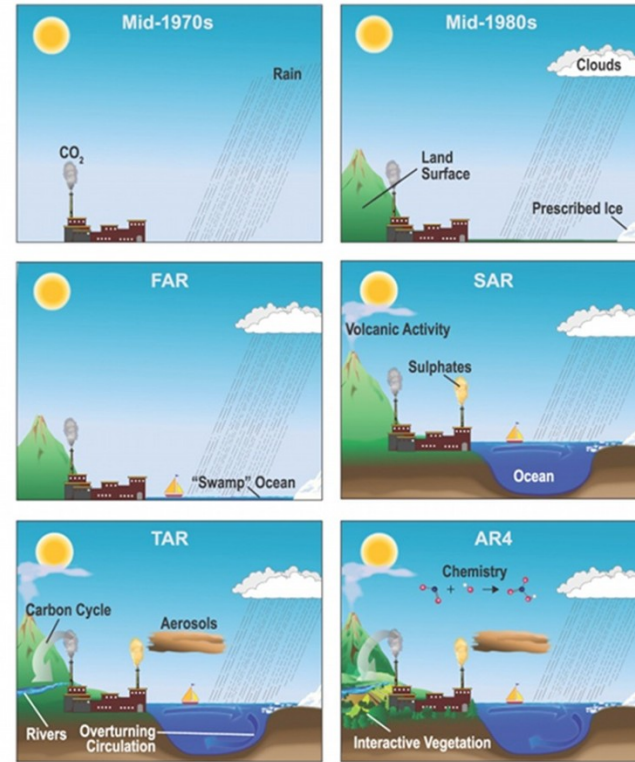
CMIP5: Fifth Coupled Model Intercomparison Project

- 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.

The World in Global Climate Models



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

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)

ESGF Earth System Grid Federation

Infrastructure for the European Research Network is-enes

Welcome to the ENES archive at BADC

Welcome to the new CMIP5 distributed archive. Our new ESGF peer-to-peer (P2P) enterprise system is now the official site for CMIP5 model output. The old [archive](#) will remain active and output from all models will continue to be available until the end of July 2012. Please send e-mail to esgf-user@lists.llnl.gov to report bugs and provide feedback.

Quick Search

Keyword: Search

Advanced Search (Category, Geospatial, Temporal, and more)...

Peer Nodes

- ANL Node
- BADC Node
- BINU Node
- CMCC Node
- DKRZ Node
- DKRZ CMIP5 Node
- NOAA-GFDL Node
- IPSL Node
- NASA-GSFC Node
- NASA-JPL Node
- NCI Node
- NERSC Node
- ORNL Node
- PCMDI Node

Guest User
ESGF P2P Version 1.4.2-master [0-2.3.2]

Privacy Policy & Legal Notice | Contact ESGF

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

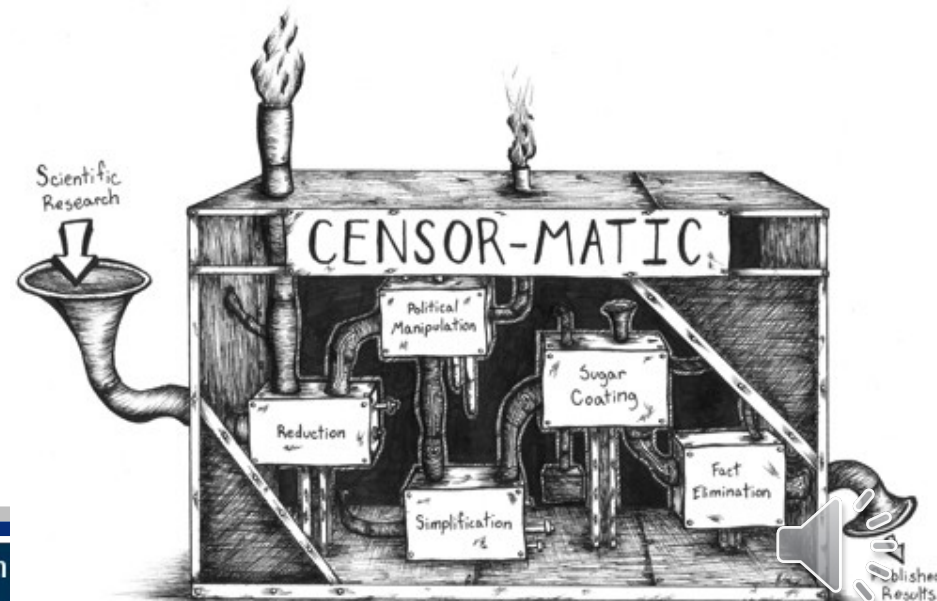


Summary and maybe conclusions?

- 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)

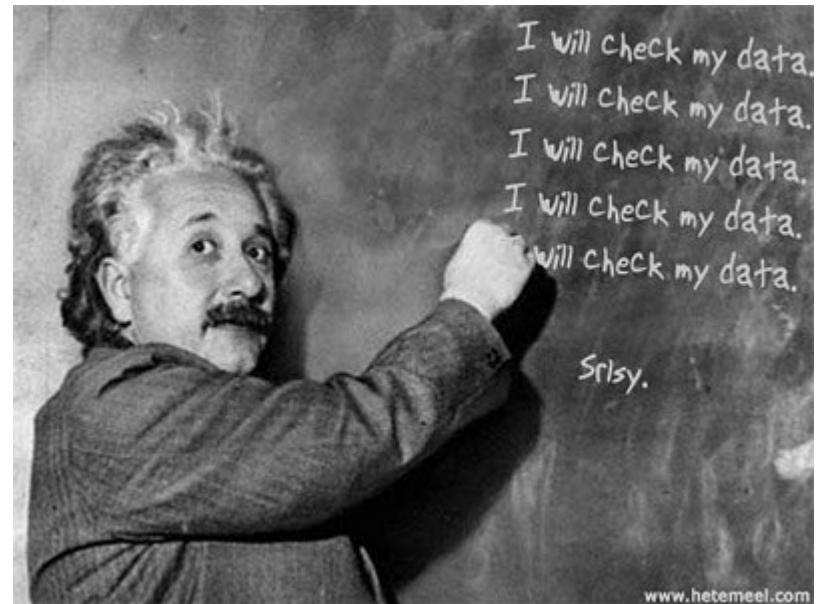


Image credit: Borepatch
<http://borepatch.blogspot.com/2010/06/its-not-what-you-dont-know-that-hurts.html>

