Data Providers

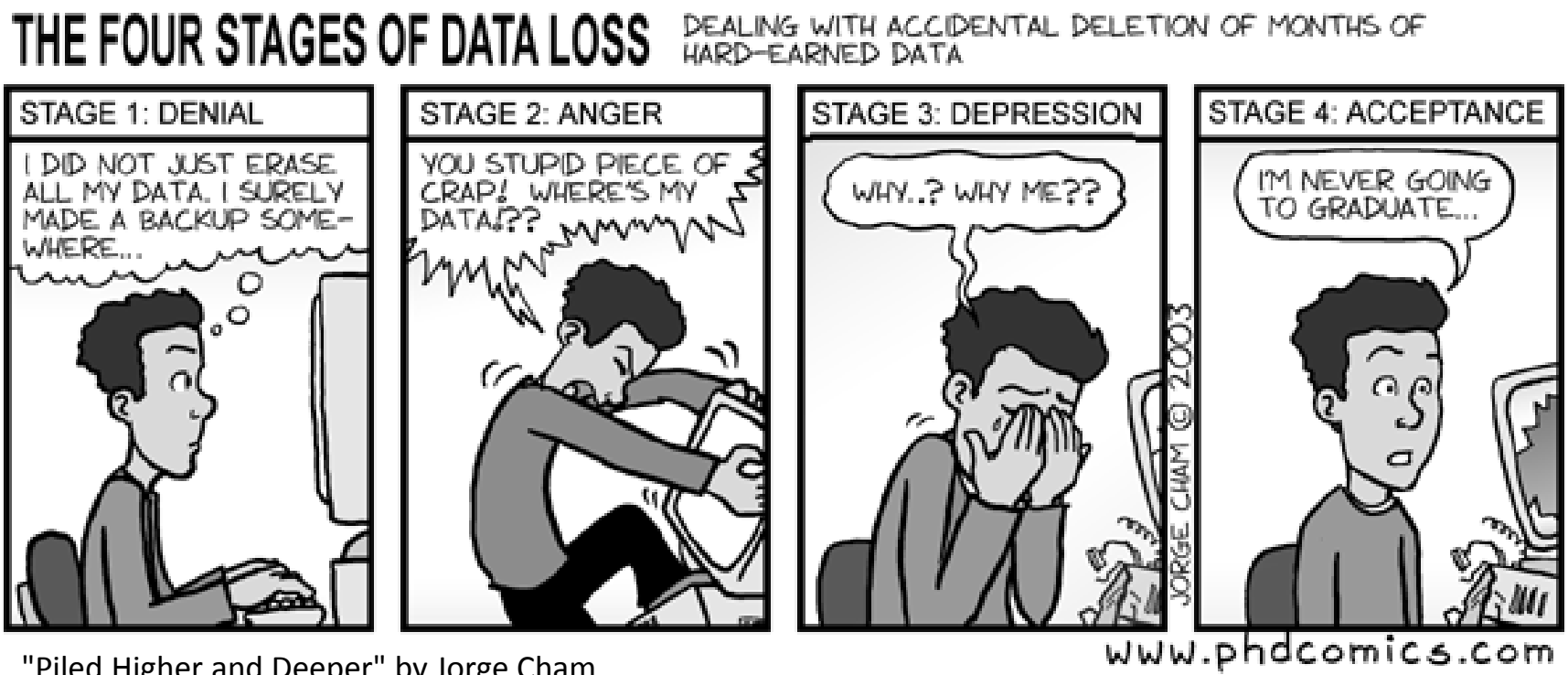
The NERC Data Policy

- Introduces a formal requirement for all applications for NERC funding to include **outline data management plans (ODMP)**, which will be evaluated as part of the standard NERC grant assessment process.

- All successful applicants for funding will be required to produce a **full data management plan (DMP)**, in conjunction with the relevant NERC data centre.

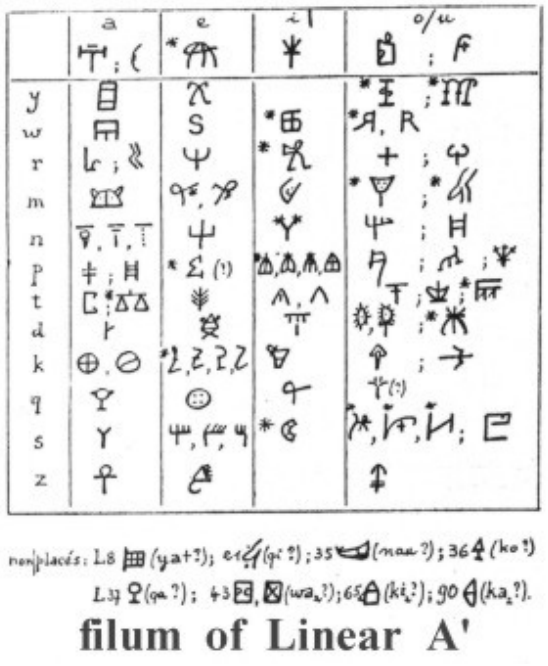
http://www.nerc.ac.uk/research/sites/data/documents/datapolicy-guidance.pdf

Why archive data anyway?



"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

Step by Step Archiving with CEDA

1. NERC Proposal inc. ODMP

2. DVC (Checklist)

3. DMP

4. Prepare data and metadata

- Data in chosen file format as per DMP
- Create filenames
- Prepare metadata following convention

5. Submit data to data centre

- Upload data as per Data Centre instructions (e.g. ftp)
- Verify Data Catalogue entry
- Get credit for your data with a DOI!

Useful Tip: Remain in contact with CEDA at all times!

Outline DMP(2)

Data centre	Dataset Description	Delivery date to Data Centre	Reuse scenarios
BADC	FAAM core data - directly from FAAM.	<24hrs after collection. Uses standard data feed to BADC from FAAM	Research re-use (researcher looking at similar atmospheric conditions or analyses of all FAAM flights.
BADC	Non-core FAAM data from instrument scientists.	When available - but not later than 2013-09-30.	Research re-use. These data are likely to be used by the XXX project.
BADC	Radiosondes in support of project campaign. These will be from the 3 experiment sites.	After quality checks. Approximately 2 months after each sonde launch.	Research re-use.
BADC	Trajectories run for areas of interest.	As run. BADC used to distribute results to project team.	Project participants only.
BADC	WRF model runs in support of campaign.		Project participants only.

Data Preparation - Introduction

Good data and metadata formats...

- Ensure future users can open data files
  - How future proof is an Excel spread sheet?
- Permit metadata harvesting from the data
- Generic extraction/processing tools for the data
- Can guarantee un-ambiguous content

File Format Explained

CEDA holds documentation and tools about the following data and metadata formats:

BADC-CSV (.csv) Campaign research data

NASA Ames (.na) primarily for aircraft observations, but can be adapted for many atmospheric observation data.

HITRAN defined by the High-resolution Transmission Molecular Absorption (HITRAN) database, widely adopted by the spectroscopy community.

JCAMP-DX only suitable for spectra from spectroscopy experiments.

NetCDF (.nc) portable self-describing binary data format e.g. model data

HDF (.hdf) Hierarchical Data Format for sharing data in a distributed environment

PP (.pp) Met Office proprietary record-based binary format (e.g. Met Office model data)

GRIB (.grb) GRIdded Binary: binary format & the data is packed to increase storage efficiency. GRIB data is also self-describing (e.g. ECMWF data)

File Names Explained

CEDA File Naming Convention:

The chosen convention is as follows:

instrument_[location|platform]_YYYYMMDD[hh][mm][ss][_extra].ext

e.g. bas-2b-o3_halley_20040101.na

For non-standard data (e.g. model data, flight data), the above convention is tweaked to best fit the needs. For example, for model data, the instrument field in the filename should instead be used for a model code (indicating the type, version etc., of the model).
e.g. jules-v2-0_ceh-condor_20060501_ancillary.nc

List of defined Instruments and locations is available from the CEDA Website.

Metadata Preparation

- Metadata are **critical for future re-use of data**.

- At CEDA, all metadata are made available through the NERC Data Catalogue Service (DCS) which increases the **visibility and discoverability of datasets**, thereby increasing their impact.

- Metadata within individual files give detailed usage information

- The file metadata should follow conventions appropriate to the file format, e.g. CF-netCDF.

Depositing Data: Methods

- Over FTP (ftp.badc.rl.ac.uk) primarily (into the "incoming" directory)

- Over the web via the File Uploader service (includes a filename checker)

- Only authorised depositors may upload data to CEDA.

- Files can be checked to see if they have been transferred correctly (checksums)

What happens at ingest?

- Data format and filename are checked copied to the archive and data integrity is checked again before original is deleted.

- Information is recorded in log files.

- Some data ingestions are carried out manually by a CEDA staff member, others are automatic

- Arrival of new data in the archive is reflected by the update of the metadata in the **CEDA Data Catalogue** (and then in the NERC DCS).

A DOI for what sort of Data?

Dataset has to be:

- **Stable** (i.e. not going to be modified)
- **Complete** (i.e. not going to be updated)
- **Permanent** – by assigning a DOI we're committing to make the dataset available for posterity.
- **Good quality** – by assigning a DOI we're giving it our data centre stamp of approval, saying that it's complete and all the metadata is available.

When a dataset is cited that means:

- There will be bitwise fixity
- With no additions or deletions of files
- No changes to the directory structure in the dataset "bundle"

A DOI should point to a *html representation* of some *record* which describes a *data object* – i.e. a **landing page**.

Upgrades to versions of data formats will result in new editions of datasets.