Science & Technology Facilities Council

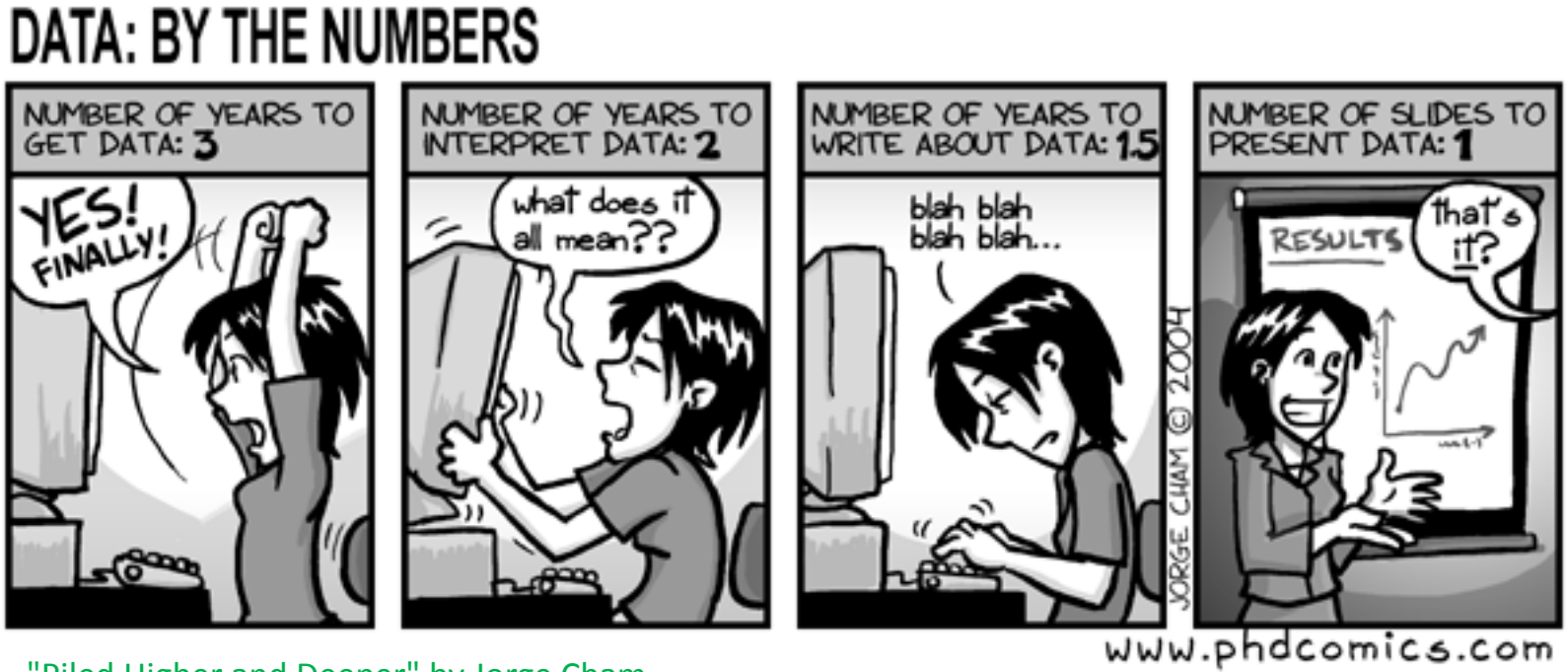
Why Good Data Curation is Essential for Doing Good Science

Alison Pamment

With thanks to Sarah Callaghan, David Hooper, Charlotte Pascoe

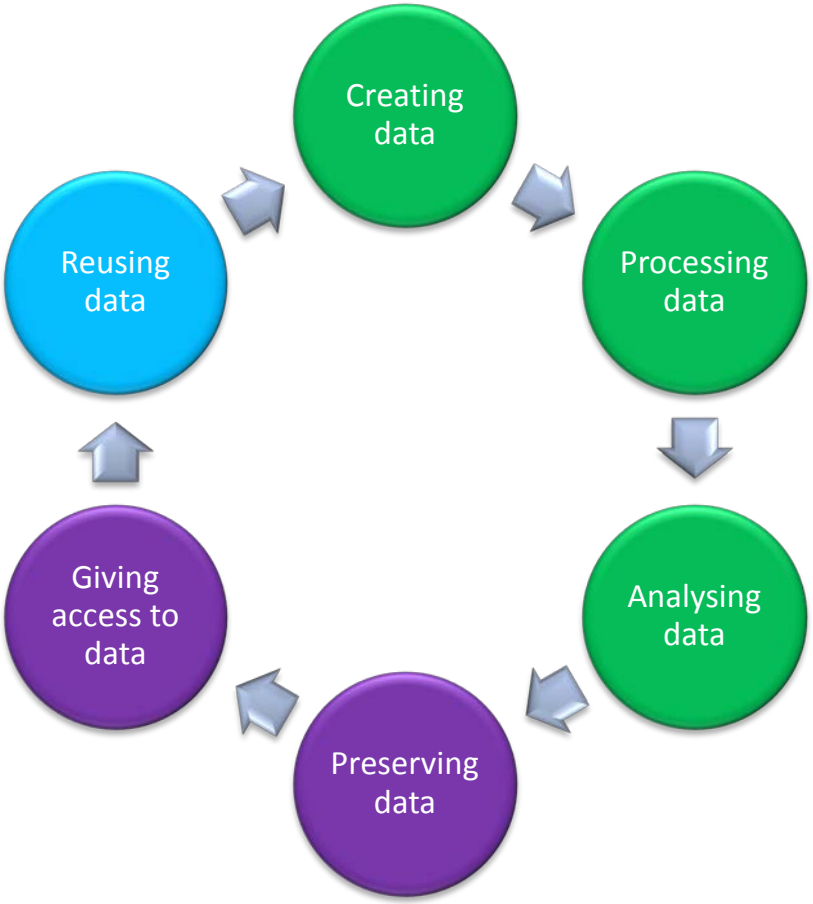
On behalf of the course team
(STFC/NERC:CEDA, NERC:NCAS CMS, NERC:NCAS Leeds)

Creating a dataset is hard work!



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

The research data lifecycle



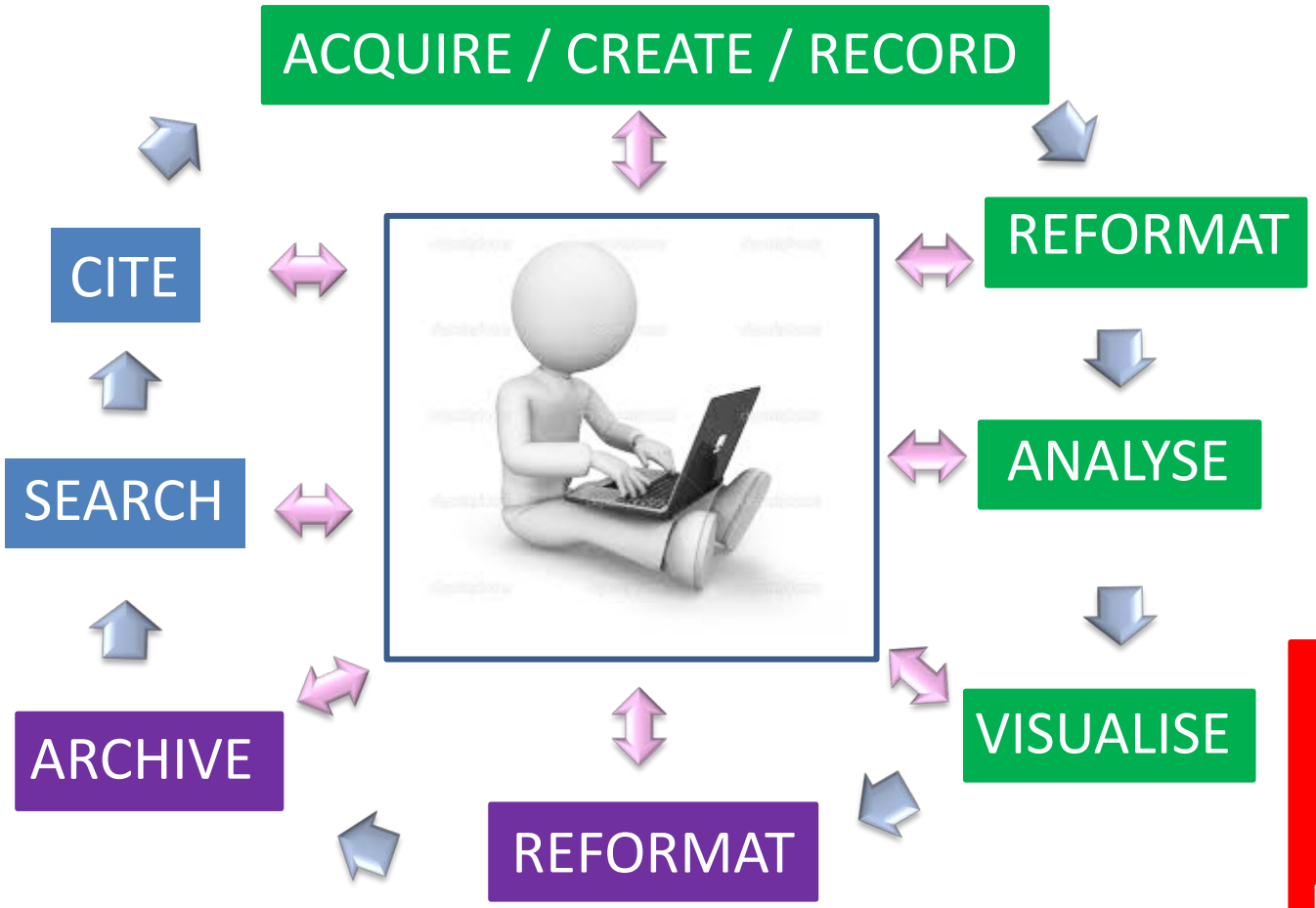
Researchers are used to creating, processing and analysing data.

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

Ways we interact with data



READ

WRITE

MOVE / COPY

Automating data interactions

- Wherever possible we use:
  ➡ common software tools
- which are designed to work with
  ➡ standard file formats
- which in turn comply with
  ➡ metadata conventions

There is (some) pain involved in learning these…

… but they make your life easier in the end

Increasing Data Impact

Good data and metadata formats...

- Help to guarantee unambiguous content
- Permit metadata harvesting from the data
- Ensure future users can open data files
  - How future proof is an Excel spread sheet?
- Enable data to be cited

which in turn…

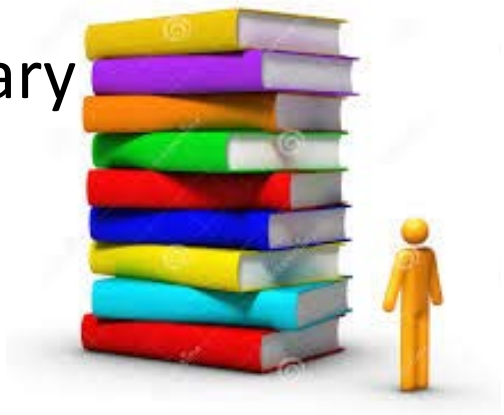
…increase data re-use…

…and data impact

DISCOVERY METADATA

Searching for data (1)

I want to find a library book on Python programming…



…I can search the library catalogue for “python”…

Searching for data (2)

"Monty Python at Work", Michael Palin. Publisher: Hern Books. TV Comedy.

"Learning Python", Mark Lutz. Publisher: O' Reilly. Computer Programming.

"Ball Pythons: Caring For Your New Pet (Reptile Care Guides)", Casey Watkins. Publisher: TokaySEO. Animal care.

Searching for data (2)

"Learning Python", Mark Lutz.
Publisher: O' Reilly.

2015, 382 pp, Computer Science, Shelf Mark 3L52, Dewey: 00532.44.3

Searching for data (4)

Nowadays we take all of this for granted – but it wouldn't be possible without **discovery** metadata

Who

How

Why

When

What

USAGE METADATA

The first 10 (of 240) lines from the file **sw010203**
(taken from the NERC MST Radar Facility archives)

```
4.31 155.3  3.92 136.1  5.15 140.2  4.23 137.1  4.75 150.2  4.71 137.9
4.35 146.5  4.52 138.0  4.83 153.7  5.40 145.8  4.63 141.0  4.90 137.3
4.31 143.3  4.58 157.0  4.94 141.7  4.65 143.1  4.63 143.0  4.88 149.5
5.42 148.5  4.92 140.4  4.04 146.7  3.92 151.5  5.02 135.3  5.06 151.6
4.65 152.3  4.31 168.8  3.79 145.3  5.92 152.9  5.02 145.8  4.77 161.6
4.79 144.1  4.60 147.5  5.33 150.1  4.81 141.0  6.02 146.9  4.38 149.0
4.42 142.5  4.58 133.4  4.35 150.5  4.96 149.8  5.56 143.4  5.08 148.5
5.19 141.6  4.40 142.4  4.10 152.6  5.02 134.0  4.94 142.9  5.27 144.4
5.38 141.5  5.88 144.8  6.00 140.1  4.75 158.3  5.08 148.1  5.46 163.5
4.27 150.8  4.69 138.8  5.71 144.0  5.21 138.8  5.00 132.4  5.06 144.4
```

What is known about this file?

sw indicates that the file contains "surface" wind data (i.e. speed and direction) from the location Frongoch

010203 represents the date in YYMMDD format

1st February 2003
(British convention)

2nd January 2003
(North American convention)

3rd February 2001
(Swedish convention)

The first 10 (of 240) lines from the file **sw010203**
(taken from the NERC MST Radar Facility archives)

```
4.31 155.3  3.92 136.1  5.15 140.2  4.23 137.1  4.75 150.2  4.71 137.9
4.35 146.5  4.52 138.0  4.83 153.7  5.40 145.8  4.63 141.0  4.90 137.3
4.31 143.3  4.58 157.0  4.94 141.7  4.65 143.1  4.63 143.0  4.88 149.5
5.42 148.5  4.92 140.4  4.04 146.7  3.92 151.5  5.02 135.3  5.06 151.6
4.65 152.3  4.31 168.8  3.79 145.3  5.92 152.9  5.02 145.8  4.77 161.6
4.79 144.1  4.60 147.5  5.33 150.1  4.81 141.0  6.02 146.9  4.38 149.0
4.42 142.5  4.58 133.4  4.35 150.5  4.96 149.8  5.56 143.4  5.08 148.5
5.19 141.6  4.40 142.4  4.10 152.6  5.02 134.0  4.94 142.9  5.27 144.4
5.38 141.5  5.88 144.8  6.00 140.1  4.75 158.3  5.08 148.1  5.46 163.5
4.27 150.8  4.69 138.8  5.71 144.0  5.21 138.8  5.00 132.4  5.06 144.4
```

What can we guess?

- Values are clearly arranged in pairs
  - 1st value of pair (e.g. 4.31) must represent speed - probably in units of m s^{-1}
  - 2nd value of pair (e.g. 155.3) must represent direction - probably in units of ⌖ from North (but meteorological or vector convention?)
- 240 lines, each with 6 columns, each with a pair of values => 1440 pairs of values
- There are 1440 minutes in a day => 1 minute sampling

The first 10 (of 240) lines from the file **sw010203**
(taken from the NERC MST Radar Facility archives)

```
4.31 155.3  3.92 136.1  5.15 140.2  4.23 137.1  4.75 150.2  4.71 137.9
4.35 146.5  4.52 138.0  4.83 153.7  5.40 145.8  4.63 141.0  4.90 137.3
4.31 143.3  4.58 157.0  4.94 141.7  4.65 143.1  4.63 143.0  4.88 149.5
5.42 148.5  4.92 140.4  4.04 146.7  3.92 151.5  5.02 135.3  5.06 151.6
4.65 152.3  4.31 168.8  3.79 145.3  5.92 152.9  5.02 145.8  4.77 161.6
4.79 144.1  4.60 147.5  5.33 150.1  4.81 141.0  6.02 146.9  4.38 149.0
4.42 142.5  4.58 133.4  4.35 150.5  4.96 149.8  5.56 143.4  5.08 148.5
5.19 141.6  4.40 142.4  4.10 152.6  5.02 134.0  4.94 142.9  5.27 144.4
5.38 141.5  5.88 144.8  6.00 140.1  4.75 158.3  5.08 148.1  5.46 163.5
4.27 150.8  4.69 138.8  5.71 144.0  5.21 138.8  5.00 132.4  5.06 144.4
```

In which order should we read the data?

Column by column and then row by row or *vice versa*?

Try both ways and plot time series of the speed and direction data
There should be no sharp discontinuities in speed or direction

Vector (i.e. towards which the wind is blowing) or meteorological direction?

Compare with synoptic pressure maps or MST radar data

The first 10 (of 240) lines from the file **sw010203**
(taken from the NERC MST Radar Facility archives)

```
4.31 155.3  3.92 136.1  5.15 140.2  4.23 137.1  4.75 150.2  4.71 137.9
4.35 146.5  4.52 138.0  4.83 153.7  5.40 145.8  4.63 141.0  4.90 137.3
4.31 143.3  4.58 157.0  4.94 141.7  4.65 143.1  4.63 143.0  4.88 149.5
5.42 148.5  4.92 140.4  4.04 146.7  3.92 151.5  5.02 135.3  5.06 151.6
4.65 152.3  4.31 168.8  3.79 145.3  5.92 152.9  5.02 145.8  4.77 161.6
4.79 144.1  4.60 147.5  5.33 150.1  4.81 141.0  6.02 146.9  4.38 149.0
4.42 142.5  4.58 133.4  4.35 150.5  4.96 149.8  5.56 143.4  5.08 148.5
5.19 141.6  4.40 142.4  4.10 152.6  5.02 134.0  4.94 142.9  5.27 144.4
5.38 141.5  5.88 144.8  6.00 140.1  4.75 158.3  5.08 148.1  5.46 163.5
4.27 150.8  4.69 138.8  5.71 144.0  5.21 138.8  5.00 132.4  5.06 144.4
```

It is often possible to "decode" ASCII files in this way, it is much more difficult for binary.

No-one will be prepared to make this effort unless they have a strong need for the data.

The data will becomes useless if the file name is changed - the date information is not recorded anywhere else.

Even if the data can be read, they may be of little scientific value unless something is known about: the type of instrument used, where it was located & how it was operated.

The partial contents of file nerc-mstrf-wind-sensors_capel-dewi_20080114_wxt510.nc

```
global attributes:
          :verbose_metadata = "Free text description" ;
          :file_version_number = 1s ;
          :data_year = 2008s ;
          :data_month = 1s ;
          :data_day = 14s ;
dimensions:
          time = 1440 ;
variables:
          float longitude() ;
                    longitude:units = "degrees_east" ;
                    longitude:axis = "X"
          float latitude() ;
                    latitude:units = "degrees_north" ;
                    latitude:axis = "Y" ;
          float altitude() ;
                    altitude:units = "m" ;
                    altitude:axis = "Z" ;
          int time(time) ;
                    time:units = "seconds since 2008-01-14 00:00:00 +00:00" ;
                    time:axis = "T" ;
          float mean_wind_speed(time) ;
                    mean_wind_speed:units = "m s-1" ;
                    mean_wind_speed:coordinates = "latitude longitude altitude" ;
                    mean_wind_speed:cell_methods = "time: minimum (interval: 3 s)" ;
                    mean_wind_speed:missing_value = 99.9f ;
          short mean_wind_direction(time) ;
                    mean_wind_direction:units = "degree" ;
                    mean_wind_direction:coordinates = "latitude longitude altitude" ;
                    mean_wind_direction:cell_methods = "time: minimum (interval: 3 s)";
                    mean_wind_direction:missing_value = 999s ;
```

File Formats

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

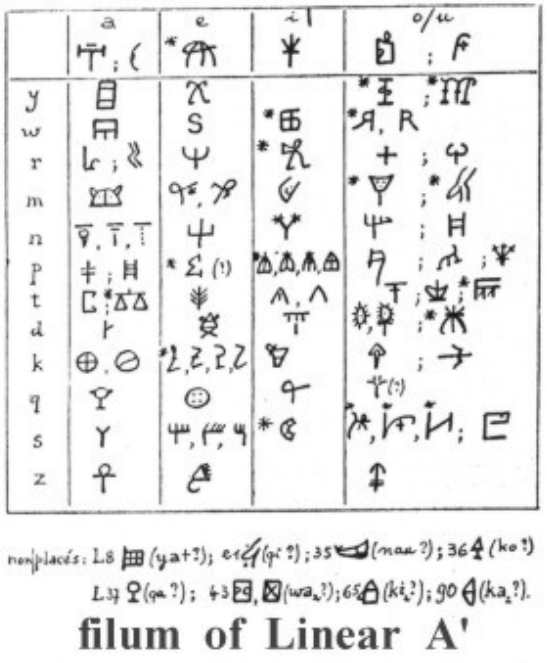
e.g. bas-2b-o3_halley_20040101.na

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.

BUT WHY GO TO ALL THIS TROUBLE?

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

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"

BAD LANDING PAGES

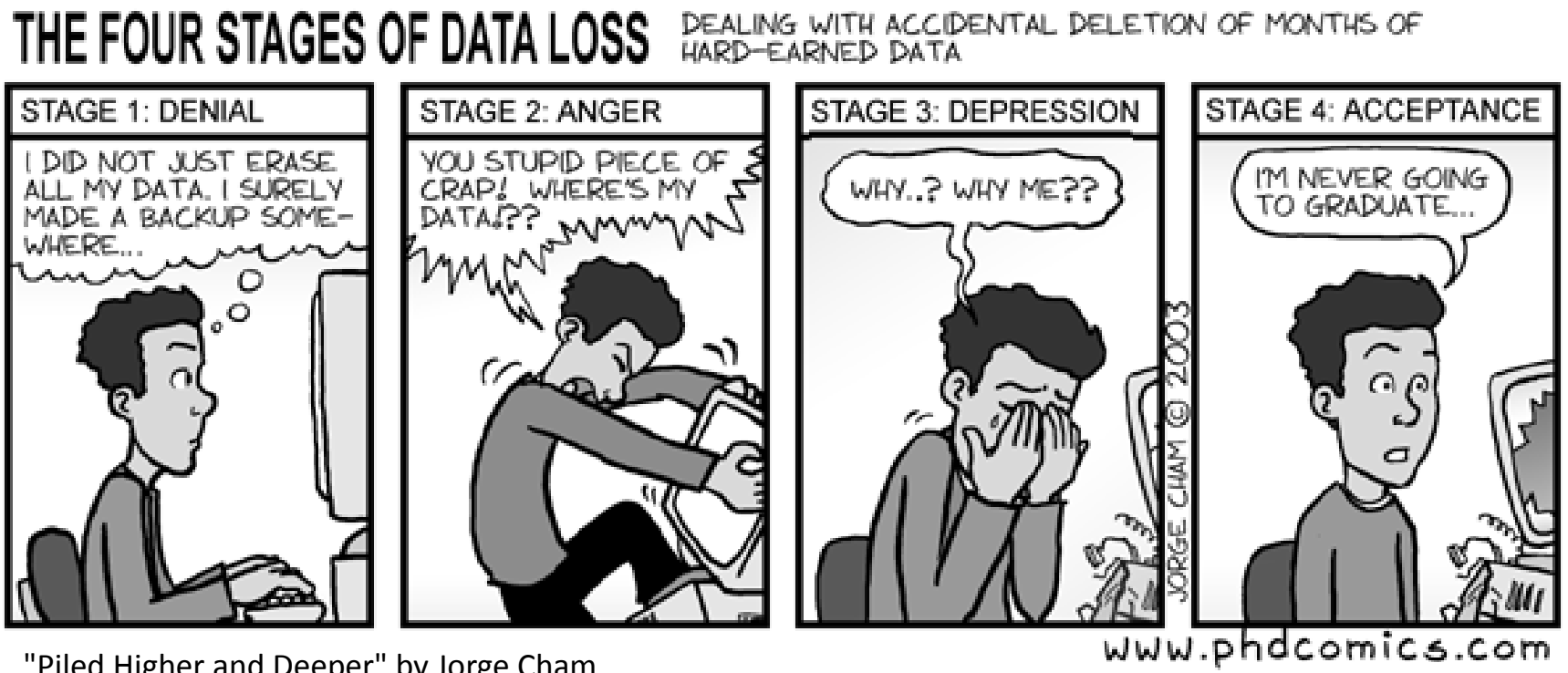
A DOI should point to a *html representation* of some *record* which describes a *data object* – i.e. a **landing page**.

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

…increase data re-use…

…and data impact

Why archive data anyway?



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

Thank you!