Read and Write Data

Data file formats and metadata – why?

Why use a standard file format?

An example of data sharing that really happened…

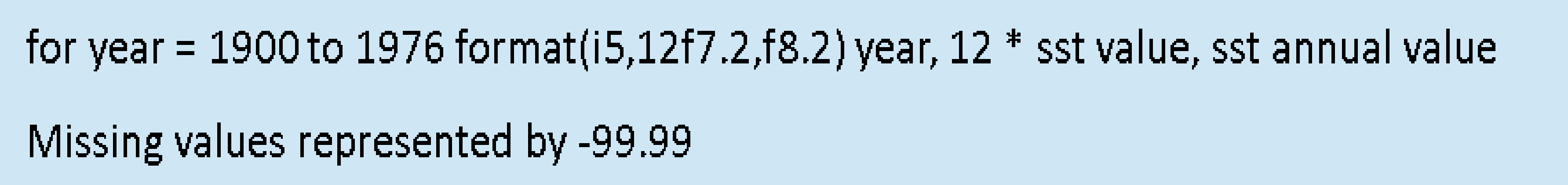
You ask a colleague for some Sea Surface Temperature data.

> *He sends a file named "SST.bin" and says "these are monthly sst anomalies from our latest climate model run covering the period 1900-1976 ".*

Why use a standard file format?

You do not know how to read the file.

He says "use a Fortran program containing the lines":



```
for year = 1900 to 1976 format(i5,12f7.2,f8.2) year, 12 * sst value, sst annual value

Missing values represented by -99.99
```

So, you try to follow the instructions…

You write the program, but get "strange" data values

*Your colleague tells you that he is using a **Little Endian** computer. Yours is a **Big Endian** computer.*

*Also, you **do not** know if both computers use the same amount of memory to store floating point numbers.*

So, you persevere…

You finally manage to read the data. But now you do not know the location of each grid point. You need the exact (lat, lon) coordinates.

> *He sends you a ASCII file (text file) with the coordinates arranged as a single column of numbers*

But more questions

You have a hard time matching the 1D column with your 2D SST array.

After that, you also may need to ask for other details about the data, e.g. which units of measurement were used for SST, which climatological values were used to calculate the anomalies, and so on…

The moral(s) of the story

- Scientists should be able to spend their time doing science.

- Every time a collaborator/user has to decode some data and ask about some metadata they are duplicating effort.

Why use standard file formats?

- We can share code.
- Even better, we can use third-party software.
- We (maybe) don't have to write any code!
- Our data will work with existing tools and services.
- Data Centres will want our data in their archives.

Why provide metadata?

- To clearly describe how we generated the data.
- To explain the context of our research.

- To unambiguously label the scientific phenomena we measured/simulated.

- To accurate geo-locate our data.
- To describe any temporal and statistical processing that has taken place.

So, which formats?

(CF-)NetCDF is usually recommended – binary.

In cases where a text format is required you might want to use:

NASA Ames
BADC-CSV (used at CEDA)