The BADC-CSV Format
Meeting user and metadata requirements

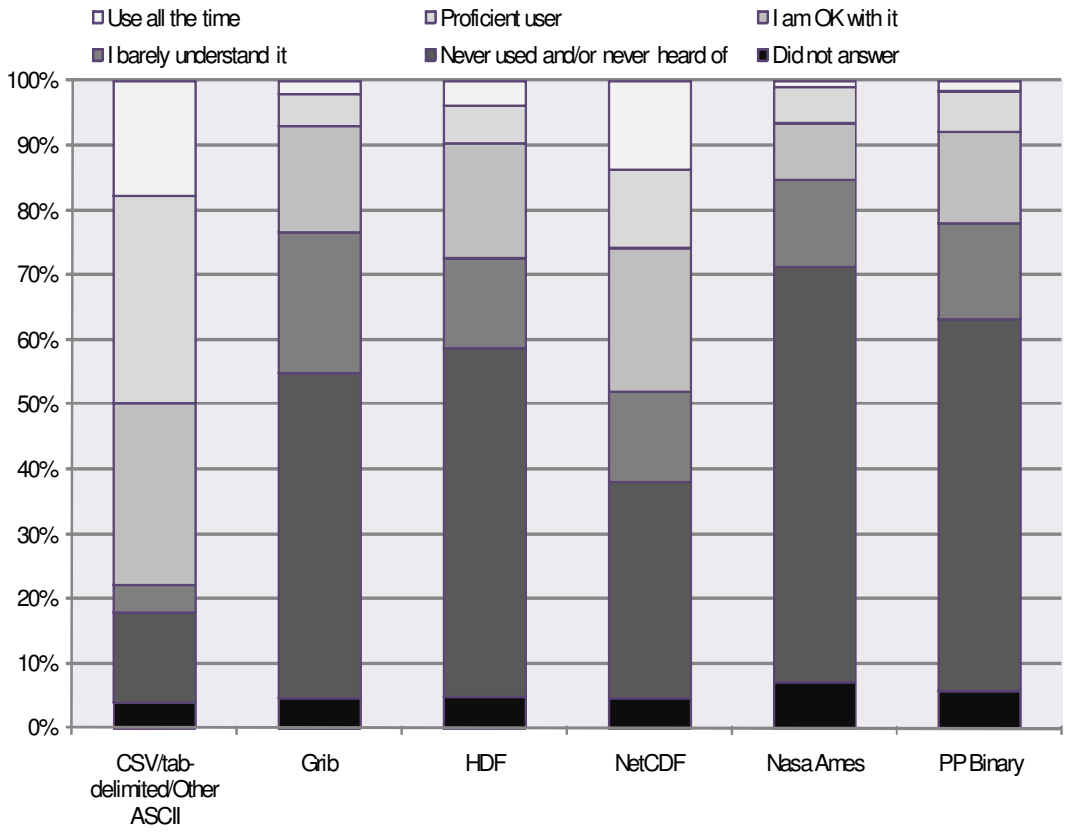
Graham A Parton, Samuel J Pepler
British Atmospheric Data Centre
graham.parton@stfc.ac.uk, sam.pepler@stfc.ac.uk

Abstract
The 2007 British Atmospheric Data Centre (BADC) Users Survey examined the skill base of the BADC's user community. Results indicated a large proportion of users who were familiar with data held in ASCII formats such as comma-separated variables (csv) and there was a high degree of familiarity with spreadsheet programmes (e.g. Excel) for data analysis purposes. These result, combined with the experiences of the BADC staff dealing with user enquiries and assisting data suppliers in preparing data for submission, and the metadata requirements of the BADC, highlighted the need for a new ASCII format to be generated. The BADC-CSV format adheres to metadata conventions covered by the NASA-Ames and netCDF formats, the CF and Dublin Core metadata conventions, the ISO19115 standard and the metadata requirements of the BADC and its sister data centres within the Natural Environment Research Council (NERC). The format meets end user and data supplier requirements by being a native format for spreadsheet software as well as other commonly used data production and analysis tools (e.g. IDL, MatLab). This paper presents the requirements for the format resulting from the 2007 user survey and data centre requirements, describes the structure of the format and demonstrates the format through short examples. Finally ongoing work to further develop the format is discussed.

Keywords: Metadata standards, new format, netCDF, CF conventions, csv, ASCII, guidelines,

1 Introduction : Format History: An Alternative to NASA-Ames

In 2007 the British Atmospheric Data Centre (BADC) undertook a user survey to determine the skill base within the BADC user community. Results from the survey (figure 1) indicated that within the user community there was a high proportion of users able to handle ASCII files (such as csv data) and a high degree of familiarity with spreadsheet programmes such as Excel.



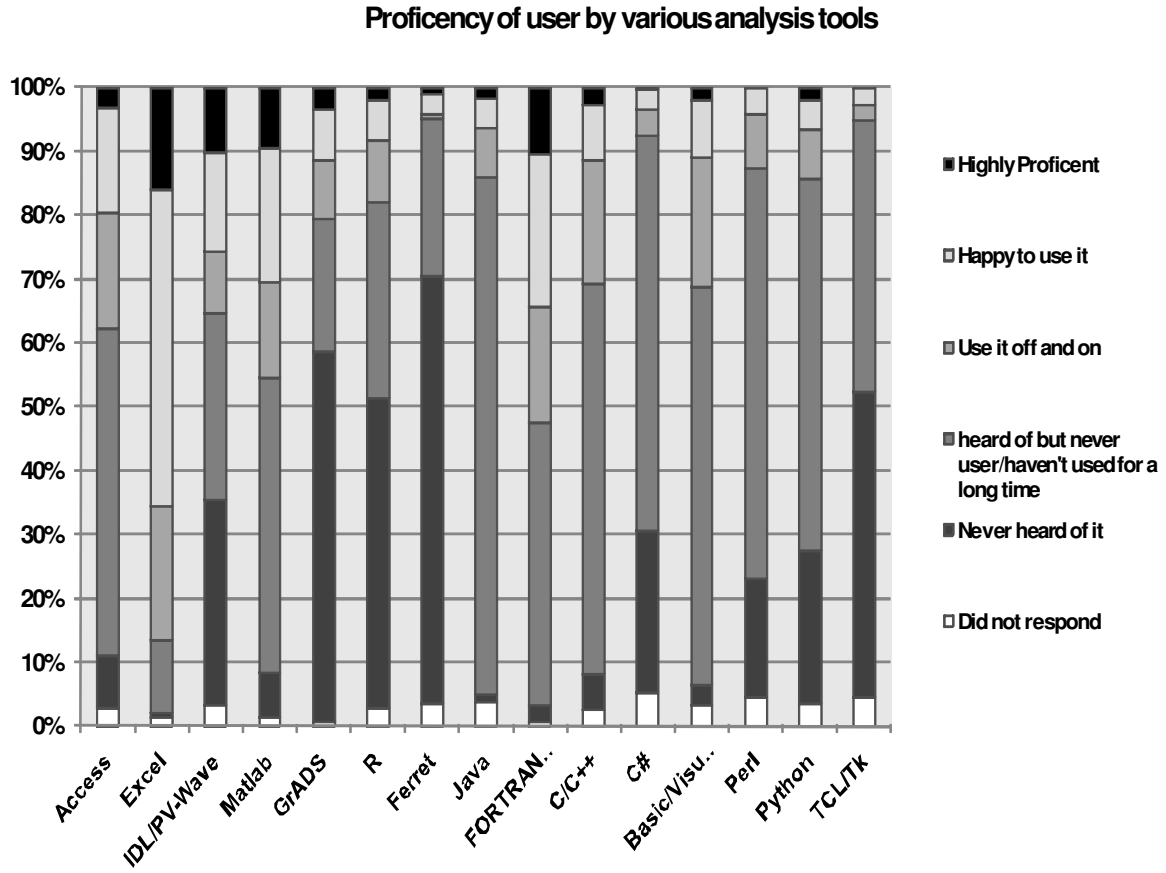


Fig. 1: results from the 2007 BADC User Survey. Top panel shows degree of familiarity of the BADC user community with various file types. Bottom panel shows proficiency of users with common analysis tools and scripting languages (PARTON, 2007).

Previously the BADC has used NASA-Ames format for ASCII data. The NASA-Ames format was devised primarily for aircraft observations, but can be adapted for many atmospheric observation data. However, correspondence with users has shown they find the NASA-Ames format to be complex and confusing. Users tend to strip the header off and import the text file into Excel. The metadata is generally not used in its machine readable form, but is simply read by the researcher. In addition, much effort is expended by the data centre's staff supporting data producers in the creation of NASA-Ames files. The format is seen by producers as complicated and it can't be done simply from spreadsheet packages like Excel. Additionally, the metadata fields offered by NASA-Ames are fixed and inflexible, with desirable metadata elements (from both the data producer and data centre perspectives) being limited to the comments section.

Model data, however, are stored at the BADC in the NetCDF format with CF metadata conventions. This provides a format framework with good flexible metadata which can be read by a number of analysis programs including FORTRAN, Matlab and IDL. It is however difficult for a researcher with little technical knowledge to use or to generate.

To solve these problems a new file format was developed to bring the advantages from the NetCDF file format with CF metadata conventions into a simple text file format. The approach was to use metadata conventions on top of comma separated values files (CSV) as produced by spreadsheet applications like Excel.

2 Format requirements

To ensure that the new ASCII format would meet the requirements of the data suppliers, data centres and end user the following criteria were set:

The format should be

- open source
- human readable
- recognisable by spreadsheet programmes (e.g. Excel, OpenOffice Calc)
- easy to generate within spreadsheet and other common data processing software and scripting languages (e.g. IDL, MatLab, Python)
- confirm to metadata conventions including CF, Dublin Core, NASA-Ames, I SO191115
- checkable by some libraries for levels of compliance

To meet these requirements a structured comma-separated-value format (referred hereafter as the BADC-CSV file format) was developed. The format would contain a designated metadata section followed by the data itself, where the elements within the metadata section could be drawn from a controlled list of metadata tags with the option of adding in additional elements where this is required/ desirable. Given the inherent flexibility in the metadata for any given file levels of compliance were also set.

3 Structure of the BADC-CSV format

The full structure of the BADC-CSV file format is given in the BADC format description document (PEPLER, 2007) available in the CEDA Document Repository.

The main points to note concerning the BADC-CSV format are that:

- The file follows CSV nomenclature, i.e. :
    - A line is a single CSV record ending in a line feed (i.e. \r\n)
    - An entry is a single comma delimited field

- A BADC-CSV file contain a series of records, each data record has 3 sections:
    - File type identifier
    - Metadata
    - Data

File type identifier

The first metadata line in the record should be the Conventions line. This aids recognising the record type. This is given as shown below to conform to the CF conventions and is the only prescibed metadata field that is capitalised. This follows the CF metadata conventions).

Conventions,G,BADC-CSV,<BADC-CSV format version number>

Metadata section

The all metadata entries are of the format:

<label>, <ref>, [<value>, <value>, …]

<label> is a metadata tag which may be an item form the list of controlled metadata items in the appendix to give greater conformity to various metadata standards, or may be one generated by the user.

<ref> is the reference to indicate where the metadata applies within the data record. A “G” indicates that the metadata applies globally to the entire record, while a data reference can be given to indicate that the metadata refers to just one of the data types within the record. This allows reference to variables and the data file in the same intuitive manner as NetCDF.

<value>, … is the set of one or more comma separated values associated with the metadata line. To aid readability it is permissible to have repeat metadata tags to allow values to be split over more than one line. Where values are not known, but required the word “unknown” should be used to indicate that these values have been addressed as fully as possible.

Data section

The data section consists of a record with a single "data" entry, followed by a line of the data references and then the data records. The end of the data records is indicated by an "end data" entry. The end data entry is included to flag partial files. Both "data" and "end data" are to be given in lower case.

data
<references>
<data lines>
end data

The conventions allow for various observation types to be intuitively encoded. The simplest example is that of a collection of point values, while the format can be readily applies to ragged profiles and similar observational data collections. The format, however, is not recommended for more complex data storage such as model data. The following sections present two sample files which further demonstrate how the format should be used.

4 Example file 1 – simple point collection.

In the first example file shown below the three sections are easy to identify. The "Conventions" line clearly shows the file type identifier section, followed by the 10 lines of the metadata section before the start of the data section. The "G" in the conventions line and the first 4 metadata lines show that these elements apply globally to the entire record, while the "1", "2" and "3" given after the "variable_name" metadata tag link these metadata elements to the 3 columns of data. The first (with values 0.8, 1.1… 4.9) are days since 2007-03-14, while the second column (2.4… 5.7) is air temperature and the third is met station air temperature.

```
Conventions,G,BADC-CSV,1
title,G,My data file
creator,G,Prof W E Ather,Reading
contributor,G,Sam Pepler,BADC
creator,G, A. Pdra
variable_name,1,time, days since 2007-03-14
variable_name,2,air temperature
variable_name,3,met station air temperature
creator,3,unknown,Met Office
coordinate_variable,1,x
location_name,G,Rutherford Appleton Lab
data
1,2,3
0.8,2.4,2.3
1.1,3.4,3.3
2.4,3.5,3.3
3.7,6.7,6.4
4.9,5.7,5.8
end data
```

5 Example file 2 – Beyond 1-D data

While not currently in the implemented version of the format, the BADC-CSV format may be extended in the future to make it suitable for 2-D type data. Such data could be indicated by the inclusion of the following metadata element:

data_dimensions,<ref>,<number of columns>,<number of rows>

This can be applied to just one, a set or all variables by use of the <ref> element.

Within the data section, instead of having one line listing all references immediately prior to the start of the data lines, each reference is listed in turn with the appropriate "n x m" data block as shown in the example below

```
Conventions,G,BADC-CSV,2
title,G,A radiosonde profile
creator,G,Prof S. Ailingby, BADC
contributor,G,Graham Parton,BADC
variable_name,pres,pressure
variable_name,lat,latitude
variable_name,lon,longitude
variable_name,temp,air tempeature
variable_name,temp_ab,absolute air
temperature
data_dimensions,G,3,2
coordinate_variable,pres,y
location_name,G,Rutherford Appleton Lab
data
pres
1000,950,850
700,300,100
lon
-4,-3,-2
4,12,13
lat
52.4,52.5,52.4
52.6,53.5,53.3
temp
16,13,12
8,-10,-20
temp_ab
289,286,285
281,263,253
end data
```

In addition to enabling data stored in 2-D arrays it is possible to add records of different types within a file. This is permitted as each record is self describing, being contained entirely between the "Conventions" and "end data" lines of the record.

6 Compliance

When badc-csv files are submitted for archiving at the BADC files are checked to ensure that they are correctly formatted. All BADC-CSV files must adhere to the following levels of compliance:

- CSV: The file should conform to Excel dialect CSV file format.
- Structure: Data and Metadata sections exist
- Valid metadata: Metadata has right number of values and refers to legal objects.

The controlled metadata list (see appendix of the format description document for details) allows further checks to be made on the files. Some metadata elements are compulsory for all BADC-CSV file, while the remaining elements are desirable. The three levels of compliance are:

- Basic: Parameter names for all columns exist. This provides a file with the same information numbers and column headings. The basic structure of the file is correct. This level requires valid metadata.
- Complete: Mandatory metadata exists. Metadata should exist for some items. Requires basic compliance.
- Standardised: Metadata values for appropriate is from standard list. Requires complete compliance.

7 Application

The BADC–CSV format is already being used by the BADC to format incoming data from the Met Office's "MetDB" system – providing daily updates of land and ship based SYNOP (meteorological) messages to the its user community. The format enables users to quickly prepare programmes to read in, manipulate and produce publication standard plots of the data within a few hours. An example of the data is given in figure 2 below, showing a plot generated from two input BADC-CSV formatted files.

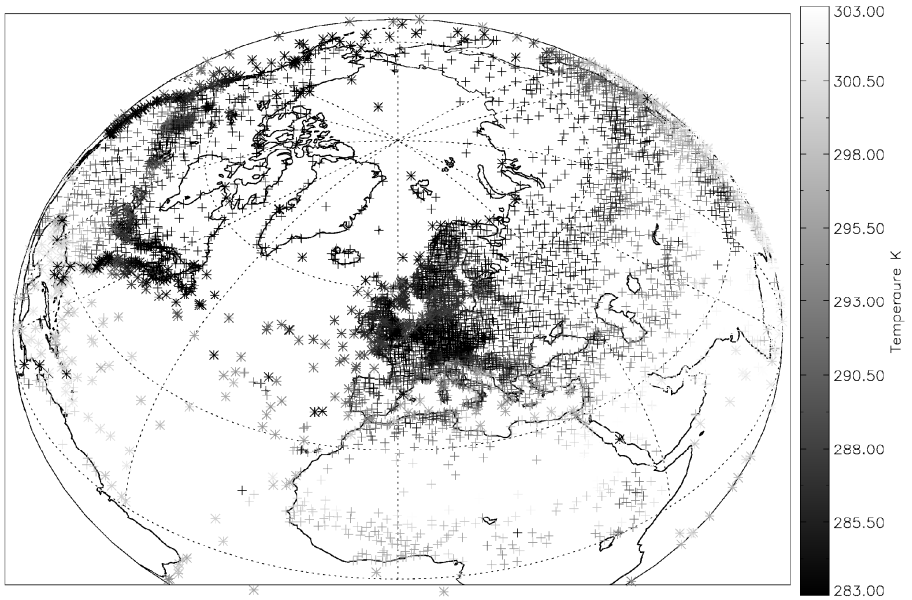


Fig. 2: Surface temperatures from land (cross) and ship(star) SYNOP messages, 06UT 26/09/2009, generated in IDL from BADC-CSV formatted data.

8 Future work

The BADC will be setting up a standards committee for the BADC-CSV format to ensure that all future developments of the format remain compliant of and take into account any changes in the metadata standards that it was designed to conform to, e.g. CF and ISO19115.

8 Conclusions

The NASA-Ames format no longer meets the requirements of the BADC and NEODC data centres for an ASCII based format within their archives. The need to develop a new format has arisen from the limitations of the NASA-Ames format to have controlled and well-structured meta-data elements beyond those it explicitly specifies and the desire for an easy to generate – easy to use human readable format from data providers and end users. The resulting csv based format conforms to existing metadata standards, such as CF-conventions, but there is now a need to ensure a governance community is set upto control any future development of the format to ensure it remains concurrent with changes in metadata standards practices and user requirements.

At present the format is suitable for basic observational data such as point collections, time series and collections of ragged profiles. It was not developed for, nor is recommended for, more complex data formats such as multiple level model data, where other standard formats are more suited.

References

PARTON, G.A. (2007): BADC User Survey 2007.
http://badc.nerc.ac.uk/community/news/BADC_survey_2007.pdf 2007-01-24.

PEPLER, S. J. (2008) BADC-CSV file format. http://cedadocs.badc.rl.ac.uk/313/1/badc-csv-format.pdf.

Contact information

British Atmospheric Data Centre
R25 1.117
STFC Rutherford-Appleton Laboratory
HSIC
Didcot
Oxfordshire
OX11 0QX
graham.parton@stfc.ac.uk
+44 (0)1235 446432