GeoCSV - tabular text formatting for geoscience data

Version: 2.0.4 (2015-07-21)

Purpose: Specify a common system of annotations and rules for data in tabular text data formats in support of a specific style described in this document called "**GeoCSV**".

An important factor is readability for both humans and machines. Simplicity is considered key for adoption and use. This specification is primarily targeted at data delivered as data streams from GeoWS web services. Ideally, existing structured text data would need very minimal modification, perhaps a few additional GeoCSV comment lines, to be compliant.

At the highest level, the format described here is composed of these types of lines: "comment" lines, one "header" line, and "data" lines. The "header and "data" lines are expected to be 100% compatible with the recommendation of the CSV on the Web Working Group (CSVW). The most applicable recommendations are described in section 7, Best Practice CSV.

Requirements and Assumptions:

The general form of a GeoCSV document comprised of "comment" lines, a "header" line, and "data" lines is:

```
# dataset: GeoCSV 2.0
# known_keyword2: value 2
# known_keyword3: value 3
header_field1,header_field2,header_field3
row1data1,row1data2,row1data3
row2data1,row2data2,row2data3
...
```

1. **Readability** by both humans and computers is very important.
2. **UTF-8** is the text encoding.
3. A "**line**" in a dataset is a text ending with carriage return line feed (i.e. CRLF), or line feed (LF).
   a. Lines starting with # are treated as comments, unless a GeoWS keyword is present.
   b. Lines starting with # can occur anywhere in the data stream.
   c. Lines without leading # are treated as delimited data.
4. **A Header line and Data lines** are lines not starting with the comment character #, for those lines,
   a. The **header line** should be a line containing field (i.e. csvw:column) names.
   b. All header and data lines should follow the CSVW recommendation concerning whitespace and when to quote.

5. **Comment lines** are lines with # as the first character. Comment lines which include **known keywords** described below should use the following rules for whitespace and padding:

   # keyword : value

   or as a list of values*:*

   # keyword : value1, value2, value3

   A line beginning with a '#', followed by zero or more whitespace characters, followed by the keyword itself, followed by zero or more whitespace characters, followed by a ':', followed by the keyword value. A POSIX extended regular expression to identify a keyword and value:

   '^#\s*(keyword)\s*:(value)[\r\n]+'

6. **Keyword values** - Values that are singular or in a list may be padded with whitespace that is not considered part of the value, all whitespace before and after a value should be trimmed by a reader. These pairs of headers are equivalent:

   # keyword : value \n

   #keyword:value\n

   or

   # keyword : value1 , value2 of apples , value3

   # keyword:value1,value2 of apples,value3

7. **Known keywords** are:
   a. **dataset**: denotes the start of a data set.
   b. **field_unit**: units for each column of data
   c. **field_type**: types for each column, one of 'string','integer','float','datetime'
   d. **field_long_name**: long descriptive field names, ala CF
   e. **field_standard_name**: long descriptive field names from a vocabulary, ala CF
   f. **field_missing**: values used to denote missing values in the data
   g. **delimiter**: single character delimiter for data values
   h. **attribution**: identify attribution information, probably a URL
   i. **standard_name_cv**: identify controlled vocabulary for field_standard_name
   j. from CF: **title, history, institution, source, comment, references**
8. **# dataset: *value*** should always be present and should be the first line of a dataset. ***Value* denotes the current container type and version (two levels), the current *value* should be GeoCSV 2.0** This keyword must also be used to denote multiple datasets, changes in number of columns, column headers, keyword value changes, etc.
9. There are no global fields. Each dataset (as defined by a new occurrence of **# dataset: GeoCSV 2.0**) must be self contained, meaning that it must have respective comment lines, a header line and data lines.
10. Values for field names and attributes (all field_* keywords) should use CF (Climate and Forecast Conventions) attribute names and definitions whenever appropriate and possible.

11. Keyword values for all field attributes (all field_* keywords) are optional and should be left empty if unknown.
12. The default delimiter is a comma. Other delimiters must be defined using the delimiter keyword.
13. Non-obvious, but common delimiters such as space and horizontal tab should be specified using the following backslash escape sequences:
    a. \s - space character (ASCII 0x20)
    b. \t - horizontal tab (ASCII 0x09)
    c. \\ - backslash (ASCII 0x5C)
14. Fields of type **'datetime'** must be in an ISO 8601 format. The form of 'YYYY-MM-DDThh:mm:ss.sss[Z]' is strongly recommended, with the time portion optional for date-only specification and optional time zone designation per ISO 8601.
15. As a minor extension to CF field names for latitude and longitude, the producer and consumer should recognize any field names that begin with "lat" or "lon" (case-insensitively) respectively as latitude and longitude. In addition " lat" or " lon" anywhere in the field name, e.g. Geodetic Longitude, will also be recognized. Example field names: lon, long, longitude, Longitude, LON, LONG, LONGITUDE, Geodetic Longitude, lonnad27, lonnad83 are all accepted as longitude. Latitude has the same rules respectively.
16. Delimiters within delimiters (Note: from sub-section "7.4 Lines" in section "7. Best Practice CSV") - Values that contain commas (or delimiter in use), line endings, or double quotes should be escaped by having the entire value wrapped in double quotes. There should not be whitespace before or after the double quotes. Within these escaped cells, any double quotes should be escaped with two double quotes.

Examples (following pages)

Long lines have been wrapped for readability, but are not wrapped in the real data. Also, keywords have been bolded for illustration and does not represent real data set formatting.

UNAVCO Example:

```
# dataset: GeoCSV 2.0
# field_unit: UTF-8, UTF-8, degrees_north, degrees_east, meters, UTC, UTC
# field_type: string, string, float, float, float, datetime, datetime
# attribution:
http://www.unavco.org/community/policies_forms/attribution/attribution.html
# GeodeticDatum: ITRF2008 epsg:1061
# Ellipsoid: GRS 1980 epsg:7019
# Ellipsoidal Coordinate System: EllipsoidalCS epsg:6423
# Axes: Geodetic longitude, Geodetic latitude, Ellipsoidal height. Orientations:
east, north, up.
# Units of Measure: decimal degrees, decimal degrees, meters
ID,station_name,latitude,longitude,ellip_height,session_start_time,session_stop_ti
me
ASBU,Astronaut
Butte,43.8206,-121.3685,1234,2011-08-18T00:00:00,2015-02-16T23:59:45
CIHL,Cinder Hill,43.7509,-121.1487,4567,2011-09-13T16:04:30,2015-02-16T23:59:45
CPCO,Central Pumice
Cone,43.7221,-121.2332,4321,2011-08-18T00:00:00,2012-03-05T12:02:30
CPCO,Central Pumice
Cone,43.7221,-121.2332,222,2012-06-14T00:00:00,2012-09-25T23:59:45
CPCO,Central Pumice
Cone,43.7221,-121.2332,999,2012-09-26T19:28:45,2013-06-10T22:11:15
```

IRIS Examples:

Seismic station metadata example:

```
# dataset: GeoCSV 2.0
# delimiter: |
# field_unit: ASCII | ASCII | degrees_north | degrees_east | meters| UTC| UTC
# field_type: string | string | float | float | float | string | datetime |
datetime
Network|Station|Latitude|Longitude|Elevation|SiteName|StartTime|EndTime
IU|ANMO|34.9459|-106.4572|1850.0|Albuquerque, New Mexico,
USA|1989-08-29T00:00:00|1995-07-14T00:00:00
IU|ANMO|34.9459|-106.4572|1850.0|Albuquerque, New Mexico,
USA|1995-07-14T00:00:00|2000-10-19T16:00:00
```

Minimal IRIS Station example:

```
# dataset: GeoCSV 2.0
# delimiter: |
Network|Station|Latitude|Longitude|Elevation|SiteName|StartTime|EndTime
IU|ANMO|34.9459|-106.4572|1850.0|Albuquerque, New Mexico,
USA|1989-08-29T00:00:00|1995-07-14T00:00:00
IU|ANMO|34.9459|-106.4572|1850.0|Albuquerque, New Mexico,
USA|1995-07-14T00:00:00|2000-10-19T16:00:00
```

Event (earthquake) parameter example:

```
# dataset: GeoCSV 2.0
# delimiter: |
```

```
EventID|Time|Latitude|Longitude|Depth/km|Author|Catalog|Contributor|ContributorID|
MagType|Magnitude|MagAuthor|EventLocationName
3954686|2010-03-01T06:27:32|38.251|69.4919|12.0|ISC|ISC|ISC|00301439|mb|4.3| NNC|
TAJIKISTAN
3954685|2010-03-01T06:25:56|37.26|138.91|9.0|JMA|ISC|ISC|15237974|mb|0.5|JMA| NEAR
WEST COAST OF HONSHU, JAPAN
```

R2R Example:

Shiptrack navigation and geophysical profiles for research cruises from the US academic fleet:

```
#dataset:GeoCSV 2.0
#names:
iso_time,ship_longitude,ship_latitude,raw_magnetics,magnetic_anomaly,igrf,device_l
ongitude,device_latitude,dp_flag
#field_unit:
ISO_8601,degrees_east,degrees_north,nT,nT,nT,degrees_east,degrees_north,unitless
#field_type: datetime,float,float,float,float,float,float,integer
#field_long_name:
date_and_time,longitude_of_vessel,latitude_of_vessel,total_magnetics_field,residua
l_magnetics_field,theoretical_magnetics,longitude_at_device_(layback),latitude_at_
device,(layback),data_status_flag
#attribution: http://www.rvdata.us/about/products
#delimiter: ,
#source: http://www.rvdata.us/
#title: Processed Magnetics Data from Research CruiseMGL1307
#cruise_id: MGL1307
#device_information: magnetometer (make: Geometrics model: G-882)
#creation_date: 2015-01-18T19:38:55+00:00
#input_data: doi:10.7284/111029
#names_NVSP02: DTUT8601,ALONGP01,ALATGP01,MAGNFLDX,MMANZZ01
iso_time,ship_longitude,ship_latitude,raw_magnetics,magnetic_anomaly,igrf,device_l
ongitude,device_latitude,dp_flag
2013-06-07T07:35:10.0997Z,-12.5669084,42.0360754,44984.68,-118.501,45103.28,-12.56
57654,42.0363823,0
2013-06-07T07:36:10.1073Z,-12.5684555,42.0360966,44981.17,-122.736,45103.274,-12.5
672536,42.0360089,0
2013-06-07T07:37:10.1950Z,-12.5699536,42.0361235,44982.41,-120.76,45103.271,-12.56
8746,42.0361294,0
```