GeoCSV: Tabular Text Formatting for Geoscience Data

M. Stults[1], T. Ahern[1], B. Arko[2], S. Carbotte[2], E. Davis[3], D. Ertz[4], M. Gurnis[5], J. McWhirter[7], C. Meertens[4], M. Ramamurthy[3], C. Trabant[1], M. Turner[5], D. Valentine[6], I. Zaslavsky[6]

[1] IRIS Data Management Center, Seattle, Washington
[2] Lamont-Doherty Earth Observatory, Columbia University
[3] Unidata - Boulder, Colorado
[4] UNAVCO - Boulder, Colorado
[5] Seismology Laboratory - Caltech
[6] CUAHSI - San Diego Super Computer
[7] GeodeSystems- RAMADDA- Boulder, Colorado

Abstract

The GeoCSV design was developed within the GeoWS project as a way to provide a baseline of compatibility between tabular text data sets from various sub-domains in geoscience. Funded through NSF's EarthCube initiative, the GeoWS project aims to develop common web service interfaces for data access across hydrology, geodesy, seismology, marine geophysics, atmospheric science and other areas.

The GeoCSV format is an essential part of delivering data via simple web services for discovery and utilization by both humans and machines. As most geoscience disciplines have developed and use data formats specific for their needs, tabular text data can play a key role as a lowest common denominator useful for exchanging and integrating data across sub-domains.

The design starts with a core definition compatible with best practices described by the W3C - CSV on the Web Working Group (CSVW). Compatibility with CSVW is intended to ensure the broadest usability of data expressed as GeoCSV. An optional, simple, but limited metadata description mechanism was added to allow inclusion of important metadata with comma separated data, while staying with the definition of a "dialect" by CSVW. The format is designed both for creating new datasets and to annotate data sets already in a tabular text format such that they are compliant with GeoCSV.

Background

In 2014, IRIS and its partners started the EarthCube GeoWS project to determine common practices and develop guidelines for sharing diverse types of data using simple web services.

Key objectives of the GeoWS project are to ease the tasks of data discovery, access and usability. Part of the vision for data exchange includes providing simple services which deliver data in both human and machine readable forms.

It was determined that to reach the widest audience and provide the greatest chance for diverse applications to read respective datasets, that defining a mechanism for tabular data exchange would be a good place to start.

The GeoCSV guidelines were created using a well-established tabular format, CSV, and additionally allow optional, straight-forward extensions that could be used by applications as desired.

W3C - CSV on the Web Working Group (CSVW)

CSVW has scheduled 17 Dec 2015 for recommendations concerning:
- Metadata Vocabulary for Tabular data
- Model for Tabular Data and Metadata on the Web
- Generating JSON from Tabular Data on the Web
- Generating RDF from Tabular Data on the Web
—> see http://w3c.github.io/csvw/

Project Members
IRIS, Caltech Seismology, GeodeSystems, LDEO, SDSC - CUAHSI, UNAVCO, Unidata

Links
GeoWS project: http://earthcube.org/group/geows-geoscience-web-services
GeoWS technical home: http://geows.ds.iris.edu
GeoCVS document: http://geows.ds.iris.edu/documents/GeoCSV.pdf
W3C - CSVW: http://www.w3.org/2013/csvw/wiki/Main_Page

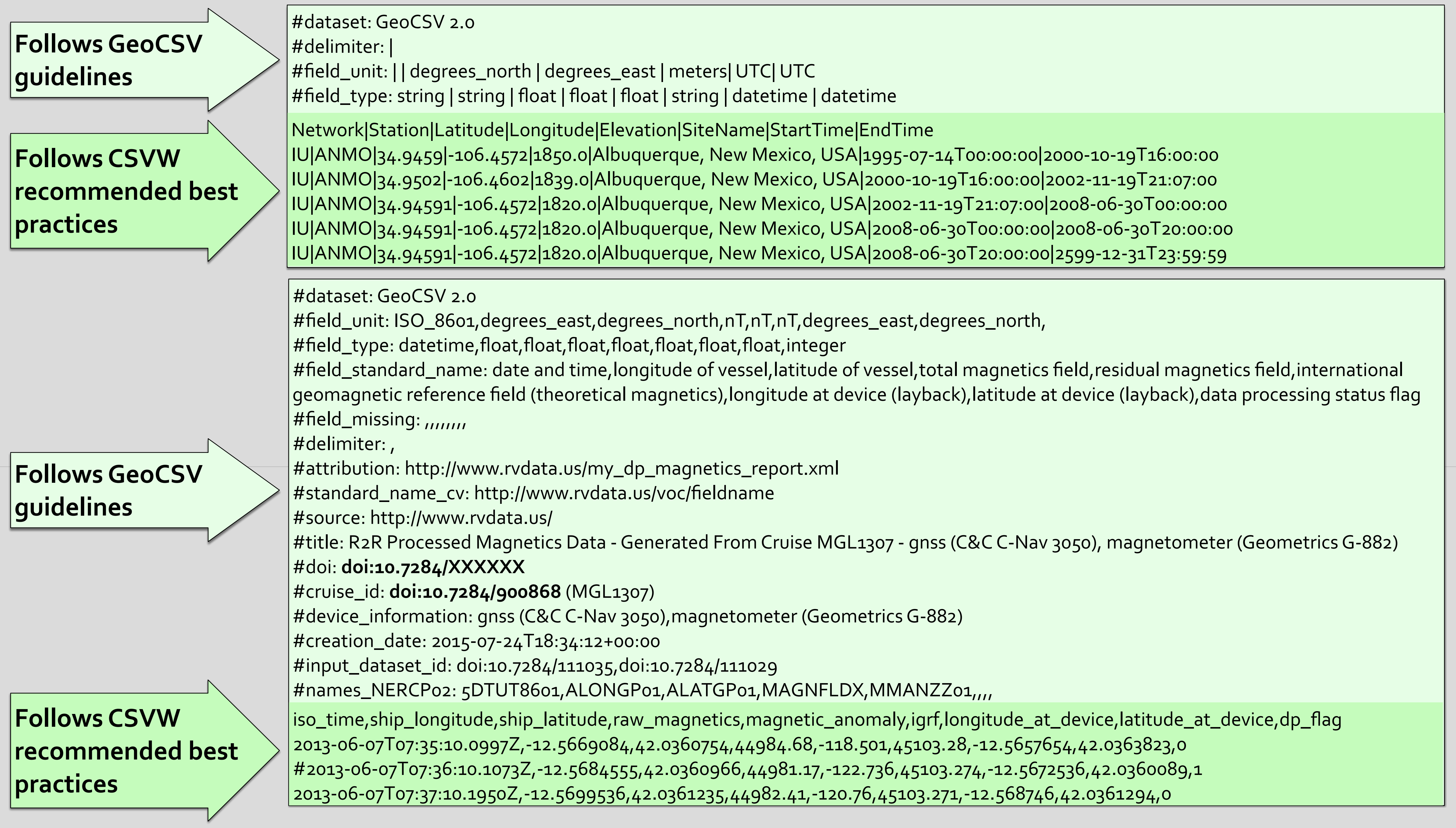
GeoCSV Purpose

Provide tabular data in both human and machine readable form
—> A GeoCSV dataset is made of two parts, an optional metadata part and a comma separated values (CSV) part
—> The metadata part is at the beginning of a dataset and each line starts with a # (hash)
—> The CSV part of the data set should conform to W3C - CSVW recommendations for best practice CSV
——> see http://w3c.github.io/csvw/publishing-snapshots/REC-syntax/Overview.html#syntax

Provide a common system of rules and annotations to enable self-defining datasets
—> Use simple keyword convention to define descriptive information
—> Keywords are contained only on lines that start with #
—> If multiple data sets are desired, the first line of each CSV set should always start with #dataset:

Profile a common mechanism for additional information per dataset
—> the GeoCSV guidelines provide agreed upon "well-known" keywords
—> Example known keywords: dataset, field_unit, field_type, delimiter, title, etc.
—> Keywords follow CF (Climate and Forecast Conventions) whenever appropriate and possible

Allow optional keyword additions that may be adopted as standard
—> Example **DOI**: a keyword to establish a link to persistent/citable identifiers for related content



Follows GeoCSV guidelines

```
#dataset: GeoCSV 2.0
#delimiter: |
#field_unit: | | degrees_north | degrees_east | meters| UTC| UTC
#field_type: string | string | float | float | float | string | datetime | datetime
```

Follows CSVW recommended best practices

```
Network|Station|Latitude|Longitude|Elevation|SiteName|StartTime|EndTime
IU|ANMO|34.9459|-106.4572|1850.0|Albuquerque, New Mexico, USA|1995-07-14T00:00:00|2000-10-19T16:00:00
IU|ANMO|34.9502|-106.4602|1839.0|Albuquerque, New Mexico, USA|2000-10-19T16:00:00|2002-11-19T21:07:00
IU|ANMO|34.94591|-106.4572|1820.0|Albuquerque, New Mexico, USA|2002-11-19T21:07:00|2008-06-30T00:00:00
IU|ANMO|34.94591|-106.4572|1820.0|Albuquerque, New Mexico, USA|2008-06-30T00:00:00|2008-06-30T20:00:00
IU|ANMO|34.94591|-106.4572|1820.0|Albuquerque, New Mexico, USA|2008-06-30T20:00:00|2599-12-31T23:59:59
```

Follows GeoCSV guidelines

```
#dataset: GeoCSV 2.0
#field_unit: ISO_8601,degrees_east,degrees_north,nT,nT,nT,degrees_east,degrees_north,
#field_type: datetime,float,float,float,float,float,float,float,integer
#field_standard_name: date and time,longitude of vessel,latitude of vessel,total magnetics field,residual magnetics field,international
geomagnetic reference field (theoretical magnetics),longitude at device (layback),latitude at device (layback),data processing status flag
#field_missing: ,,,,,,,,
#delimiter: ,
#attribution: http://www.rvdata.us/my_dp_magnetics_report.xml
#standard_name_cv: http://www.rvdata.us/voc/fieldname
#source: http://www.rvdata.us/
#title: R2R Processed Magnetics Data - Generated From Cruise MGL1307 - gnss (C&C C-Nav 3050), magnetometer (Geometrics G-882)
#doi: doi:10.7284/XXXXXX
#cruise_id: doi:10.7284/900868 (MGL1307)
#device_information: gnss (C&C C-Nav 3050),magnetometer (Geometrics G-882)
#creation_date: 2015-07-24T18:34:12+00:00
#input_dataset_id: doi:10.7284/111035,doi:10.7284/111029
#names_NERCP02: 5DTUT8601,ALONGP01,ALATGP01,MAGNFLDX,MMANZZ01,,,,
```

Follows CSVW recommended best practices

```
iso_time,ship_longitude,ship_latitude,raw_magnetics,magnetic_anomaly,igrf,longitude_at_device,latitude_at_device,dp_flag
2013-06-07T07:35:10.0997Z,-12.5669084,42.0360754,44984.68,-118.501,45103.28,-12.5657654,42.0363823,0
#2013-06-07T07:36:10.1073Z,-12.5684555,42.0360966,44981.17,-122.736,45103.274,-12.5672536,42.0360089,1
2013-06-07T07:37:10.1950Z,-12.5699536,42.0361235,44982.41,-120.76,45103.271,-12.568746,42.0361294,0
```

http://ds.iris.edu/ Facilitate Collaborate Educate

2015 AGU Fall Meeting
IN11F-1809