What does your data mean?

Tools for understanding and making metadata for data

Jeremy Kenyon
jkenyon@uidaho.edu

Research Librarian,
UI Library

Summary

- Understand what documentation is
- Understand the composition of metadata
- Be able to select the appropriate standard(s) for your research project
- Select a tool or technology to create metadata

Types of Documentation

- Readmes
- Laboratory/Research Notebooks
- Data Dictionaries
- Structured Metadata
- Self-describing data formats

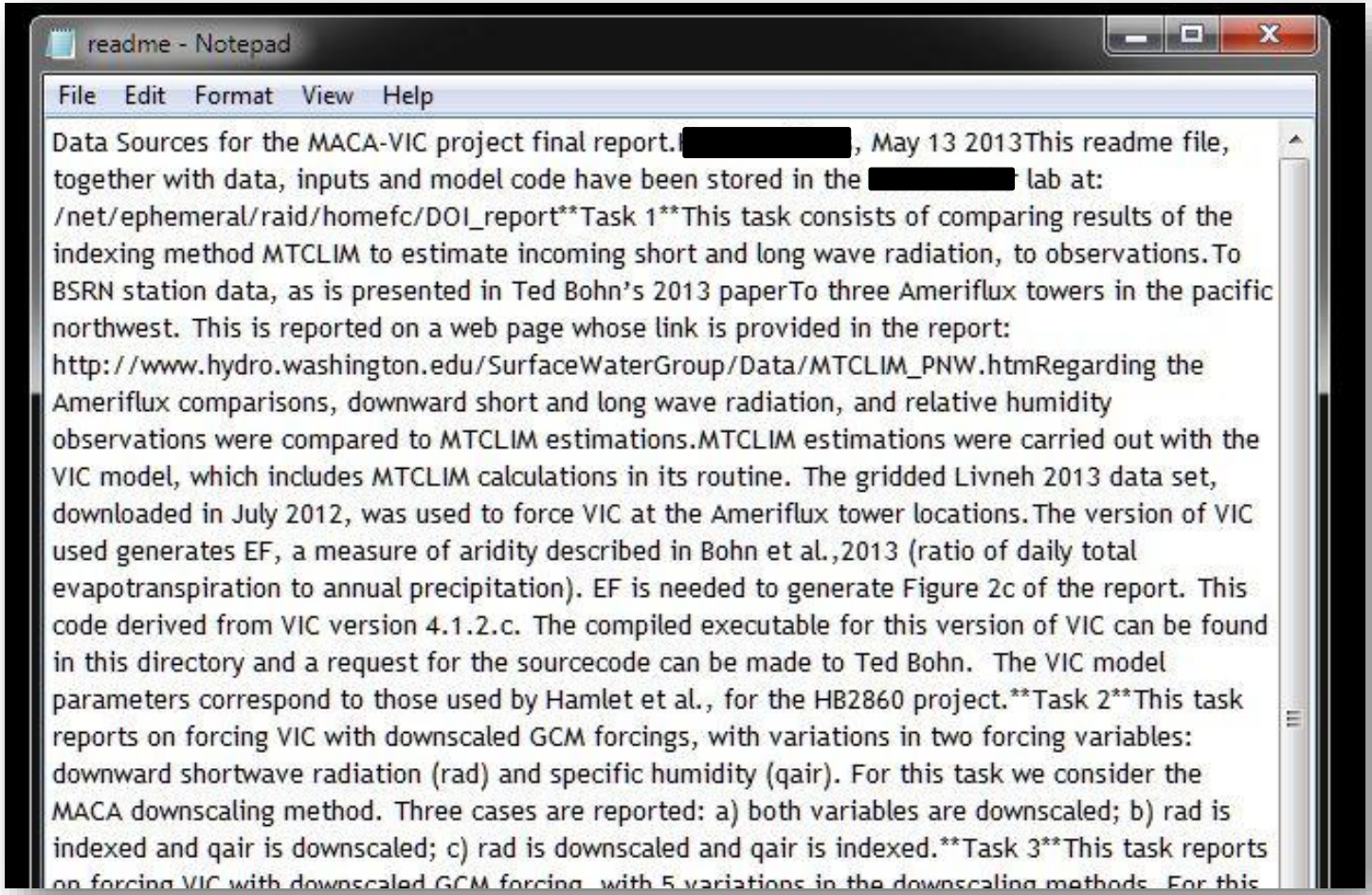
Readme Files

A readme (or read me) file contains information about other files in a directory or archive and is very commonly distributed with computer software.

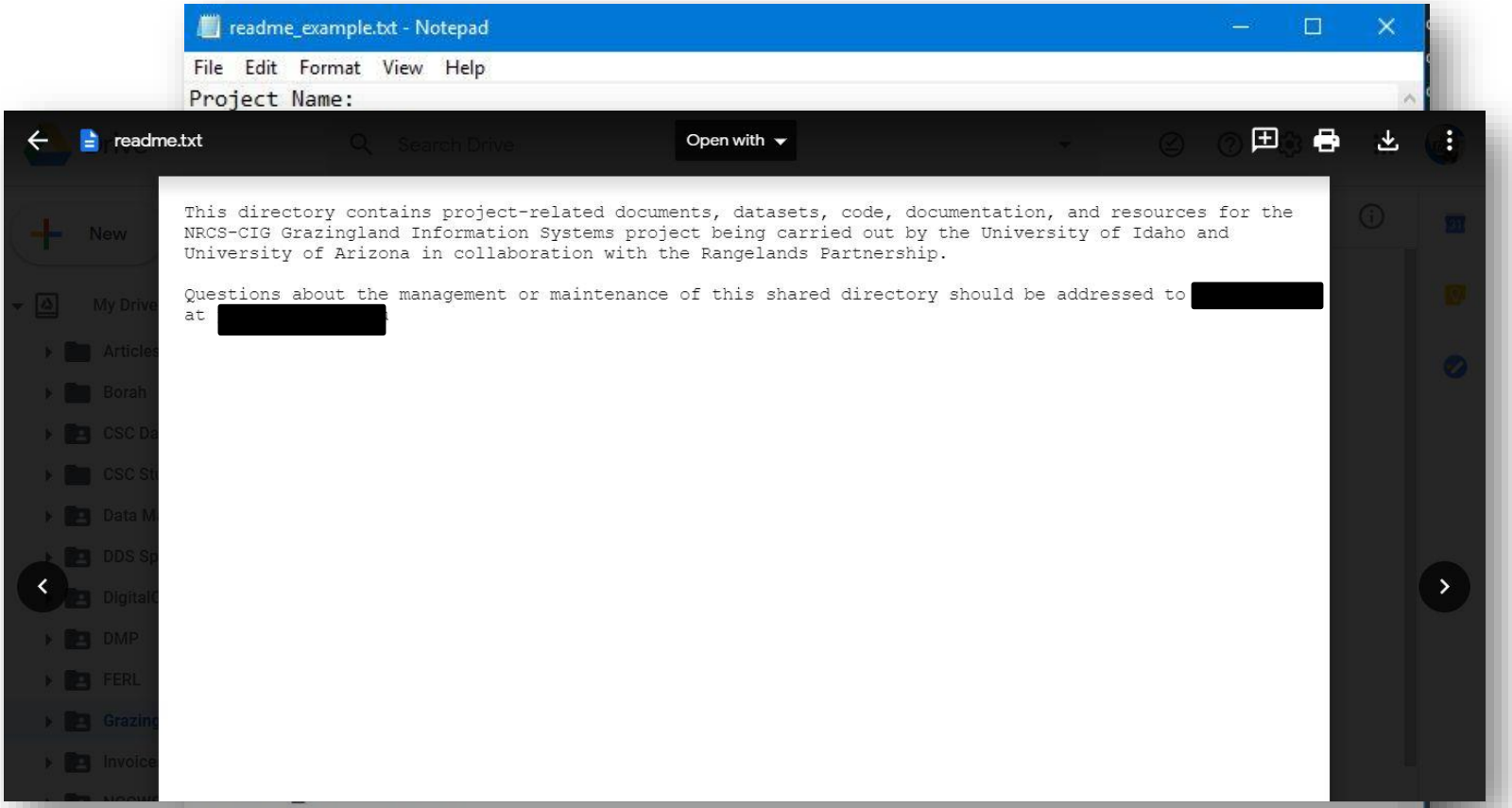
Wikipedia, 2014

- used to provide any amount of context and background
- typically unstructured
- simple and easy to create

Readme Files



Readme Files



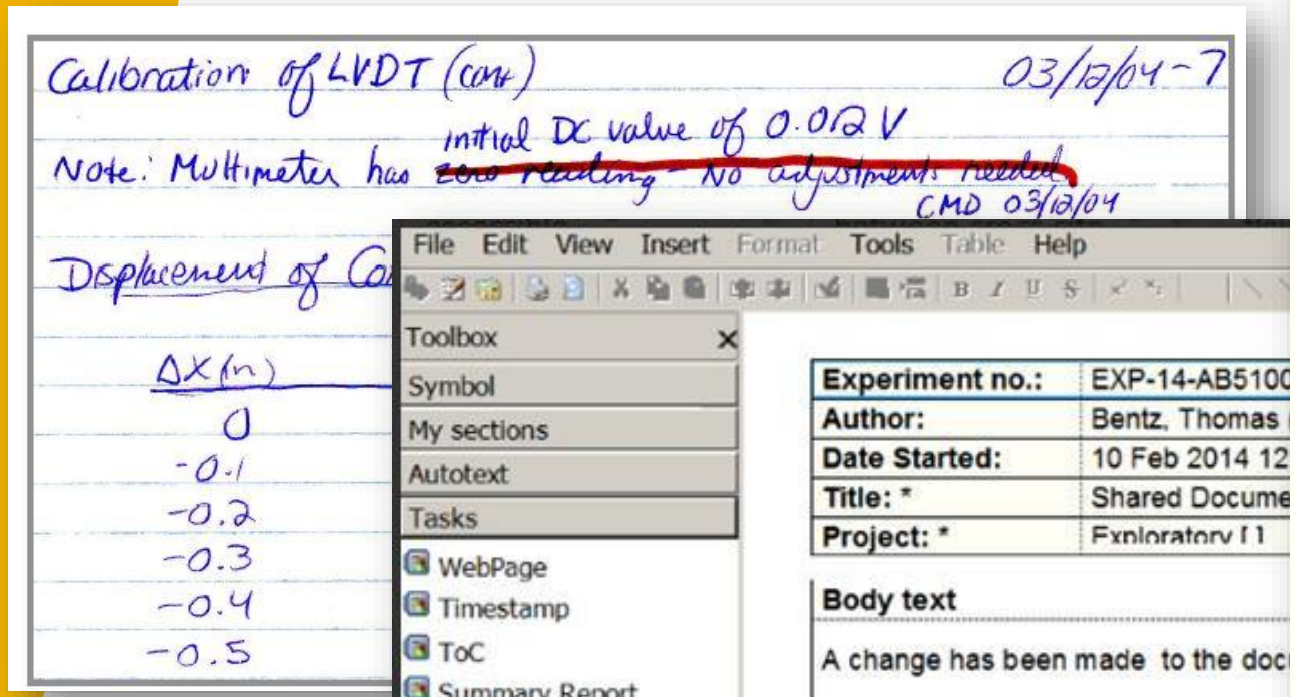
Research Notebooks

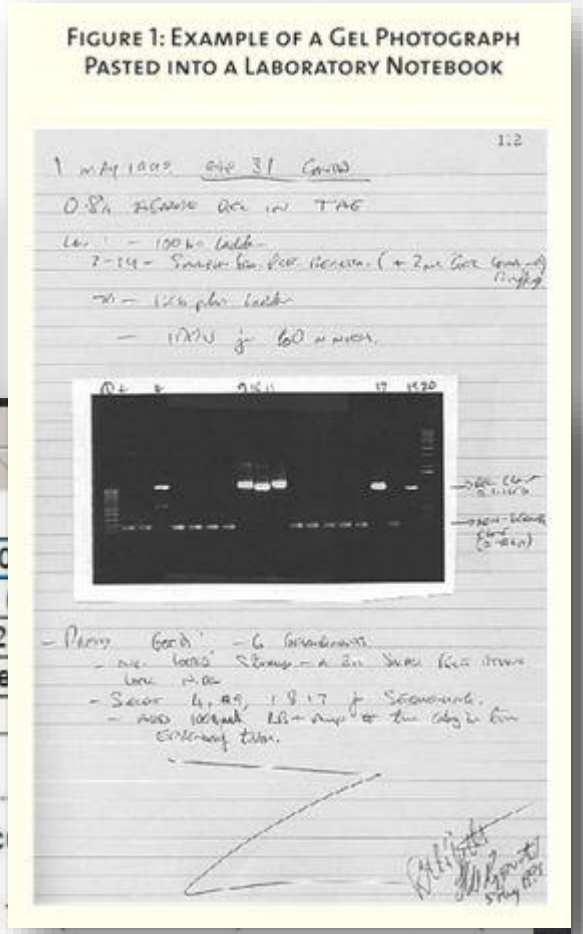
provides a reliable reference for writing up materials and methods and results for a study. It is a legally valid record that preserves your rights or those of an employer or academic investigator to your discoveries.

D. Caprette, 2014

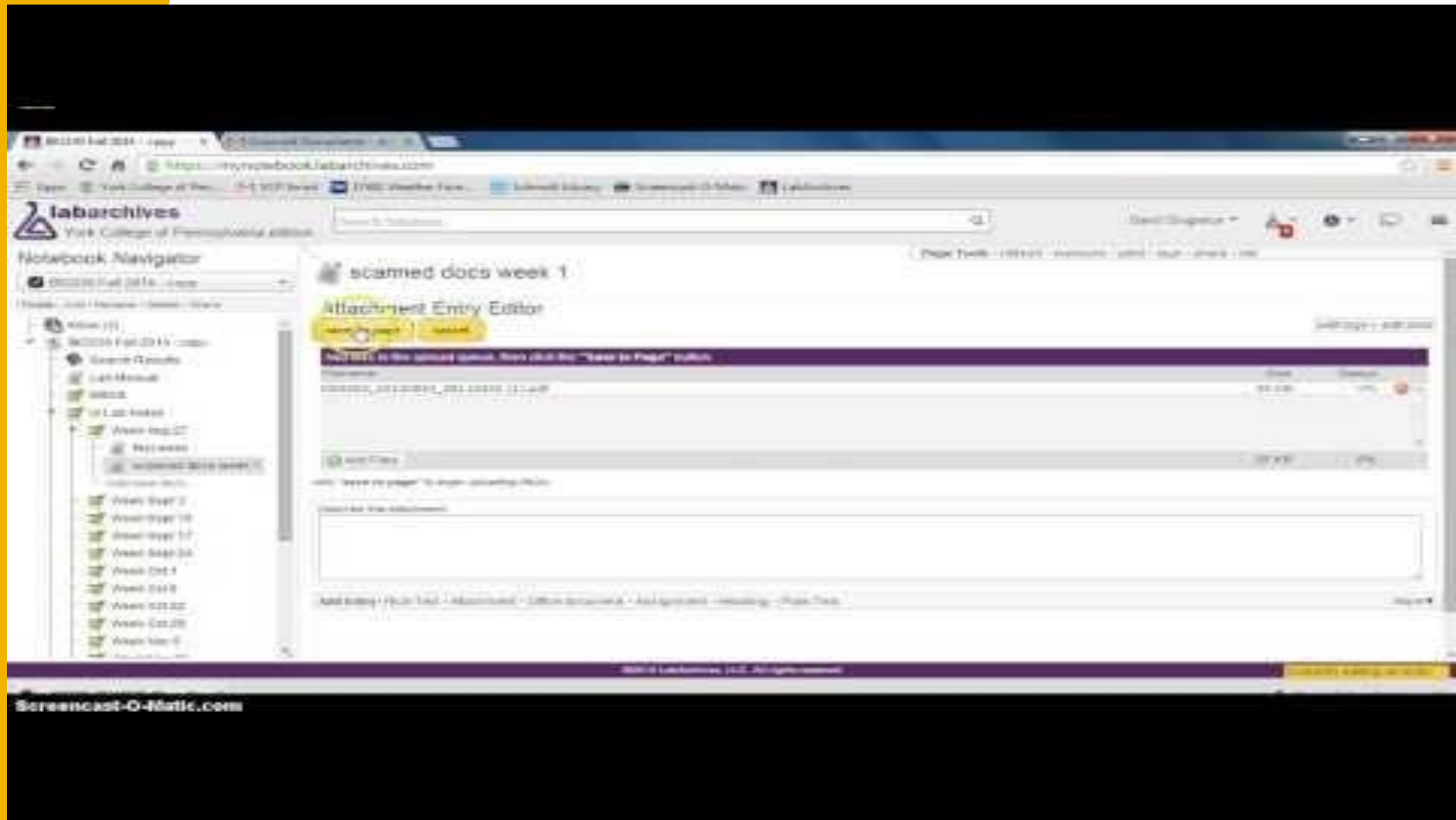
- used to provide methodological context
- unstructured
- simple and easy, can be either in print or electronic

Research Notebooks





Research Notebooks



Data Dictionaries

Give you space to define variables while keeping the actual dataset streamlined and computable.

Briney, 2015

- Can be structured, albeit idiosyncratically
- Fairly efficient and easy to create
- Simple and can be re-used later on

Data Dictionaries

Potential Fields to Include:

- Variable Name
- Variable Definition
- Variable Definition Source
- How measured
- Data units
- Data format
- Min/max values
- Coded values/defs
- Null values representation
- Precision of measurement
- Known issues
- Relationship to other variables
- Other notes

Example:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4322587/table/t4/
https://collectionbuilder.github.io/images/data_dictionary.pdf

Tools:

https://docs.google.com/spreadsheets/d/1Qx9WuCpwK15rfMU5hAD85k24jfUZePOf_gfHUrVkbzM/edit?usp=sharing

Formal Metadata

structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information source.

NISO, 2004

- used to provide context and background
- structured
- can be either simple or complex depending on the type
- required for exchanging information between systems

Metadata examples

Dublin Core – one of the simplest forms of describing digital information:

1. TITLE
2. CREATOR
3. SUBJECT
4. DESCRIPTION
5. PUBLISHER
6. CONTRIBUTORS
7. DATE
8. TYPE
9. FORMAT
10. IDENTIFIER
11. SOURCE
12. LANGUAGE
13. RELATION
14. COVERAGE
15. RIGHTS MANAGEMENT

Example:

https://www.lib.uidaho.edu/digital/crbp/items/crbp1239.html

Tools:

https://developers.exlibrisgroup.com/blog/creating-dublin-core-xml-files-using-excel/

https://docs.google.com/spreadsheets/d/1HyvtBqrMdnosbK0TnOX81LdPzm-1lk0_P-AN16fv2GI/edit#gid=0

Metadata examples

Geographic (spatial) metadata:

- FGDC
- ISO 19115

Example:

Raster: https://www.sciencebase.gov/catalog/item/58828bc6e4b0dc04318c5d23

Vector: https://www.sciencebase.gov/catalog/item/54494685e4b0f888a81bb4d9

Tools:

https://osf.io/y78b8/wiki/home/

Metadata examples

Social scientific metadata:

- DDI (Data Documentation Initiative)
- SDMX (Statistical Data and Metadata Exchange)

Example:

DDI:
https://www.datalumos.org/datalumos/project/101387/version/V2/view;jsessionid=E3CD8A299581CAE75ACF6C9897880F10

Tools:

DDI: https://ddialliance.org/resources/tools
https://www.colectica.com/software/colecticaforexcel/

SDMX:

- https://sdmx.org/?page_id=4500
- https://ec.europa.eu/eurostat/web/sdmx-infospace/sdmx-it-tools

Metadata examples

Disciplinary area	Metadata standard
General	Dublin Core (DC)
Life Sciences	Darwin Core
	Ecological Metadata Language (EML)
Humanities	Text Encoding Initiative (TEI)
Social Sciences	Data Documentation Initiative (DDI)
Geospatial	Content Standard for Digital Geospatial Metadata (CSDGM)
	ISO 19115-2:2009
Physical Sciences	NetCDF Climate and Forecast (CF) Metadata Conventions
	Crystallographic Information Framework (CIF)

Selecting the right standard

Identify constraints with your project resources

1. Anticipated repository requirements
2. Funding agency compliance
3. Team capabilities

Other Tools:

- Morpho (EML): http://knb.ecoinformatics.org/morphoportal.jsp
- DarwinCore Assistant (DwC): http://tools.gbif.org/dwca-assistant/

Other Tools:

https://www.dataone.org/software_tools_catalog

http://www.dcc.ac.uk/resources/metadata-standards

Self-describing formats

These formats contain extensive internal metadata, which provides user systems with all the information needed for both use and discovery. Station data, grids and rasters can be accommodated in these formats.

IODE, 2014

- used to provide data-level description and context
- highly structured
- complex formats, usually require extensive skill
- Often seen as NetCDF or HDF formats in environmental modeling

Self-describing formats

```
Dataset {
    Float64 lat[lat = 442];
    Float64 lon[lon = 523];
    Float32 time[time = 60];
    Grid {
     ARRAY:
        Float32 wind_speed[time = 60][lat
     MAPS:
        Float32 time[time = 60];
        Float64 lat[lat = 442];
        Float64 lon[lon = 523];
    } wind_speed;
} NWCSC_INTEGRATED_SCENARIOS_ALL_CLIMATE/r
---------------------------------------
lat[442]
49.41666333333333, 49.37499666666667, 49.3
49.12499666666667, 49.08333, 49.041663333
48.79166333333333, 48.74999666666667, 48.7
48.49999666666667, 48.45833, 48.416663333
48.16666333333333, 48.12499666666667, 48.0
47.87499666666667, 47.83333, 47.791663333
47.54166333333333, 47.49999666666667, 47.4
47.24999666666667, 47.20833, 47.166663333
46.91666333333333, 46.87499666666667, 46.8
46.62499666666667, 46.58333, 46.541663333
46.29166333333333, 46.24999666666667, 46.2
45.99999666666667, 45.95833, 45.916663333
45.66666333333333, 45.62499666666667, 45.5
45.37499666666667, 45.33333, 45.291663333
45.04166333333333, 44.99999666666667, 44.5
44.74999666666667, 44.70833, 44.666663333
44.41666333333333, 44.37499666666667, 44.3
```

```
dimensions:
   profile = 30 ;
   z = 42 ;
variables:
   float lon ;
       lon:standard_name = "longitude";
       lon:long_name = "station longitude";
       lon:units = "degrees_east";
   float lat ;
       lat:standard_name = "latitude";
       lat:long_name = "station latitude" ;
       lat:units = "degrees_north" ;
   char station_name(name_strlen) ;
       station_name:cf_role = "timeseries_id" ;
       station_name:long_name = "station name" ;
   int station_info;
       station_name:long_name = "some kind of sta
   float alt(profile , z) ;
       alt:standard_name = “altitude”;
       alt:long_name = "height above mean sea lev
       alt:units = "km" ;
       alt:axis = "Z" ;
       alt:positive = "up" ;
   double time(profile ) ;
       time:standard_name = "time";
       time:long_name = "time of measurement" ;
       time:units = "days since 1970-01-01 00:00:
```

References

1. Briney, Kristin. 2015. Data Management for Researchers: organize, maintain, and share your data for research success. Exeter, UK: Pelagic Publishing.
2. http://en.wikipedia.org/wiki/README
3. http://library.oceanteacher.org/OTMediawiki/index.php/Self-Describing_Formats
4. http://www.ruf.rice.edu/~bioslabs/tools/notebook/notebook.html
5. http://www.niso.org/publications/press/UnderstandingMetadata.pdf
6. http://www.iphandbook.org/handbook/ch08/p02/
7. http://institute.lanl.gov/ei/LADSS/_docs/Lab_Notebooks_6-11-09.pdf
8. http://www2.archivists.org/glossary/terms/m/metadata