Guidelines for writing "readme" style metadata

by Wendy Kozlowski (wak57@cornell.edu). Last updated 05.30.2014
Adapted from Dryad's recommendations for authors (http://datadryad.org/depositing) now available at https://web.archive.org/web/20120413115438/http://www.datadryad.org/depositing and from The Knowledge Network for Biocomplexity's Introduction to Ecological Metadata Language (EML) (http://knb.ecoinformatics.org/eml_metadata_guide.html, now available at https://web.archive.org/web/20120424124714/http://knb.ecoinformatics.org/eml_metadata_guide.html)

Best practices:

Create one readme file for each data file, whenever possible. It is also appropriate to describe a "dataset" that has multiple, related, identically formatted files, or files that are logically grouped together for use (e.g. a collection of Matlab scripts). When appropriate, also describe the file structure that holds the related data files. See Example 2 for describing grouped or multiple files.

Name the readme so that it is easily associated with the data file(s) it describes.

Write your readme document as a plain text file, avoiding proprietary formats such as MS Word whenever possible.[1] Format the readme document so it is easy to understand (e.g. separate important pieces of information with blank lines, rather than having all the information in one long paragraph).

Format multiple readme files identically. Present the information in the same order, using the same terminology.

Use standardized date formats. Suggested format: W3C/ISO 8601 date standard, which specifies the international standard notation of YYYYMMDD or YYYYMMDDThhmmss.

Follow the scientific conventions of your discipline for taxonomic, geospatial and geologic names and keywords. Whenever possible, use terms from standardized taxonomies and vocabularies, a few of which are listed below.

Source	Content	URL
Getty Research Institute Vocabularies	geographic names, art & architecture, cultural objects, artist names	http://www.getty.edu/research/tools/vocabularies/
Integrated Taxonomic Information System	taxonomic information on plants, animals, fungi, microbes	http://www.itis.gov/
NASA Thesauri	engineering, physics, astronomy, astrophysics, planetary science, Earth sciences, biological sciences	http://www.sti.nasa.gov/sti-tools/#.UHhG0MXA98E
GCMD Keywords	Earth & climate sciences, instruments, sensors, services, data centers, etc.	http://gcmd.nasa.gov/learn/keywords.html
The Gene Ontology Vocabulary	gene product characteristics, gene product annotation	http://amigo.geneontology.org/cgi-bin/amigo/browse.cgi
USGS Thesauri	agriculture, forest, fisheries, Earth sciences, life sciences, engineering, planetary sciences, social sciences etc.	http://www.usgs.gov/core_science_systems/csas/biocomplexity_thesaurus/Thesauri.html

[1] More information about file formats can be found on the RDMSG website (http://data.research.cornell.edu) and the DataOne website: "Document and store data using stable file formats"

Recommended Content:

Recommended minimum content for data re-use is in **bold.**

1. *Introductory information*

a. For each filename, a short description of what data it contains
b. Format of the file if not obvious from the file name
c. If the data set includes multiple files that relate to each other, the relationship between the files or a description of the file structure that holds them (possible terminology might include "dataset" or "study" or "data package")
d. Name/institution/address/email information for
    i. Principle investigator (or person responsible for collecting the data)
    ii. Associate or co-investigators
    iii. Contact person for questions
e. Date of data collection (can be a single date, or a range)
f. Information about geographic location of data collection
g. Date that the file was created
h. Date(s) that the file(s) was updated and the nature of the update(s), if applicable
i. Keywords used to describe the data topic
j. Language information

2. *Methodological information*

a. Method description, links or references to publications or other documentation containing experimental design or protocols used in data collection
b. Any instrument-specific information needed to understand or interpret the data
c. Standards and calibration information, if appropriate
d. Describe any quality-assurance procedures performed on the data
e. Definitions of codes or symbols used to note or characterize low quality/questionable/outliers that people should be aware of
f. People involved with sample collection, processing, analysis and/or submission

3. *Data-specific information*

a. Full names and definitions (spell out abbreviated words) of column headings for tabular data
b. Units of measurement
c. Definitions for codes or symbols used to record missing data
d. Specialized formats or abbreviations used

4. *Sharing/Access information*

a. Licenses or restrictions placed on the data[2]
b. Links to publications that cite or use the data
c. Links to other publicly accessible locations of the data
d. Recommended citation for the data
e. Information about funding sources that supported the collection of the data

[2] See also: "Introduction to Intellectual Property Rights in Data Management" by Peter Hirtle, Senior Policy Advisor, Cornell University Library

Example 1: Single Data File readme [3]

Data Filename: 2002_NCAGL_Inventory.txt
readme File Name: README_2002_NCAGL_Inventory.txt
Data file contents:

Site	DateTime	Plot	Sp	PA	S	Bm	P	N
VO	20020530T1000	1	S1	1	3	16.25	3.19	0.01
VO	20020530T1000	1	S2	1	3	16.25	3.19	0.01
VO	20020530T1000	1	S3	1	3	16.25	3.19	0.01
VO	20020530T1430	2	S1	1	2	16.25	3.19	0.02
VO	20020530T1430	2	S2	1	2	16.25	3.19	0.02
VO	20020530T1430	2	S3	0	2	16.25	3.19	0.02
CH	20020601T0900	1	S1	1	3	88.82	11.91	0.02
CH	20020601T0900	1	S2	1	3	88.82	11.91	0.02
CH	20020601T0900	1	S3	1	3	88.82	11.91	0.02
CH	20020601T1315	2	S1	1	2	65.62	-999	0.03
CH	20020601T1315	2	S2	0	2	65.62	-999	0.03
CH	20020601T1315	2	S3	1	2	65.62	-999	0.03

readme File Contents:

Introductory Information:

Data Filename: 2002_NCAGL_Inventory.txt

Data File Description: These data were collected as part of the Northern California Grasslands (NCAGL) diversity research program. Data collected include species richness, presence absence of plant species, peak standing biomass and nitrogen and phosphorus soil content. The relationship between diversity and productivity can take many different shapes. Soil nutrients can affect species composition, diversity and productivity. This research program will attempt to investigate soil nutrients to as a possible factor in determining the shape of the diversity productivity curve.

Data File Format: File is a tab separated text file, originally created in Microsoft Excel. File should also be readable in any basic text editor such as Notepad, Open Office, TextEdit etc.

Related Data Files: None

Principle Investigator / Data Owner: Jane Q. Researcher; Department of Biology; Northern California University; University Town, CA 95666; (123) 456-7890; researcher@uncal.edu

Associated Investigators: None

[3] Adapted from: http://knb.ecoinformatics.org/eml_metadata_guide.html. Now available at: https://web.archive.org/web/20120424124714/http://knb.ecoinformatics.org/eml_metadata_guide.html.

Contact Person: Sally R. Labmanager; Department of Biology; Northern California University; University Town, CA 95666; (123) 456-7891; labmanager@uncal.edu

Data Collection Dates: 20020530 - 20020601

Geographic Coverage: Data were collected in the coastal mountains of Northern California, in the Valley Oak Reserve. The Valley Oak Reserve is adjacent to and managed by Northern California University (NCU). NCU is located in University Town in Sonoma County, approximately 150 km northeast of San Francisco. Bounding coordinates are West: -120°15'00", East: -120°30'00", North: -39°15'00", South: 38°45'00". Exact locations of sites and plots are available upon request.

Data File Submission Date: 20050601

Data File Updated: None

Keywords: richness, productivity, grasslands, biomass, northern California, soil nutrients

Methodological Information:

Methods: Twenty five 1 m^2 plots were randomly placed throughout the Valley Oak Reserve. Due to destructive biomass harvest, plots are relocated each year. Two plastic sample bags (Ziplock) are labeled with the randomly assigned plot number and contents (plant or soil). If more than one bag is needed, all bags are labeled with the contents, plot number and bag number (i.e. the second of four bags of plant clippings from plot 6 will be labeled "Plant, Plot 6, Bag 2/4").

At each plot, one person, starting at the south-east corner of the plot, identifies each plant according to the Jepson manual. Species names are recorded in the field notebook.

Species names are used to calculate species richness per plot.

All plant material within each 1 m^2 plot is clipped at soil level and placed in the sample bag. Sample bags containing plant matter are brought to the laboratory. If wet, plant matter is dried using paper towels. Plant material is dried in a drying oven at 80 degrees C for 24 (+/-2) hours. Plant matter is weighed within 2 hours of drying.

Approximately 0.5 g of soil, free from plant debris, is collected from the middle of the plot. Soil is placed in appropriate sample bag. Soil samples are placed into aluminum sample trays and placed into a drying oven and dried at 80 degrees C for 24 (+/- 2) hours. Soil is ground using a ball mill until powdery and weighed. Soil sample is analyzed for Phosphorus and Nitrogen using a SoilPro v. 10 machine. Lower limits of detection for Phosphorus is 0.005 and for Nitrogen is 0.01. All procedures for this machine are followed.

Quality Assurance: All sampling was done by Jane Researcher, Sally Labmanager and John Fieldassistant. Data was plotted and reviewed for data entry errors by a second lab member before submission.

Column Headers: Site: Site at which data were collected.

Site	Code
Coastal Hills Reserve	CH
Valley Oaks Reserve	VO

DateTime: Date data were collected YYYYMMDDThhmm format
Plot: Randomly assigned number of plot
Sp: Species inventoried

Species Name	Code
Avena fatua	S1
Bromus hordeaceus	S2
Calochortus lutens	S3

PA: Observed presence or absence of each of three species inventoried. For each species, a value of 1 indicates presence and a value of 0 indicates absence.
S: Species richness, calculated as total number of species per plot
Bm: Peak standing biomass, measured in grams
P: Phosphorous in soil, recorded in ppm (parts per million)
N: Nitrogen in soil, recorded as a percentage; BLD indicates "Below Level of Detection" of instrument, as detailed in the methods (not detected).

Missing Data: Missing data (not collected or sample lost in processing) are indicated with a "-999". BLD indicates "Below Level of Detection" of instrument (see Methods).

Sharing and Access Information:

Licensing: This data is freely available for re-use. Please acknowledge the Knowledge Network for Biocomplexity, NSF Grant #12345 and Dr. Jane Researcher in any publications that use this data.

Related Publications: Reseacher, Jane Q and Labmanager, Sally R. (2003) Soil Nutrients and the Relationship between Diversity and Productivity. *Science* 5959:1234-1235.

Data Source: This dataset is publicly available through the Knowledge Network for Biocomplexity (http://www.knb.ecoinformatics.org).

Recommended Citation: Reseacher, Jane Q and Labmanager, Sally R. (2005) Data from: Northern California Grasslands diversity research project. Knowledge Network for Biocomplexity. http://dx.doi.org/11.1111/knb.1111t1. [4]

Funding Information: Collection of the data was funded by NSF grant #12345.

---

[4] Please note that this example includes a DOI in the citation. Persistent identifiers such as DOIs, Handles, ARKs etc. are an ideal way for people to obtain shared data, but in this example, this is for illustrative purposes only, as KNB does not currently offer DOI's for the datasets.

Example 2: Multiple Data Files readme – Longitudinal Study[5]

DataSet Name: NCAGL_Inventory_v3_1.zip
Data Filenames: 2002_NCAGL_Inventory.txt
2003_NCAGL_Inventory.txt
2004_NCAGL_Inventory_v2.txt
readme Filename: README_NCAGL_Inventory_v3_1.txt[6]
Data file 2002_NCAGL_Inventory.txt contents:

Site	DateTime	Plot	Sp	PA	S	Bm	P	N
VO	20020530T1000	1	S1	1	3	16.25	3.19	0.01
VO	20020530T1000	1	S2	1	3	16.25	3.19	0.01
VO	20020530T1000	1	S3	1	3	16.25	3.19	0.01
VO	20020530T1430	2	S1	1	2	16.25	3.19	0.02
VO	20020530T1430	2	S2	1	2	16.25	3.19	0.02
VO	20020530T1430	2	S3	0	2	16.25	3.19	0.02
CH	20020601T0900	1	S1	1	3	88.82	11.91	0.02
CH	20020601T0900	1	S2	1	3	88.82	11.91	0.02
CH	20020601T0900	1	S3	1	3	88.82	11.91	0.02
CH	20020601T1315	2	S1	1	2	65.62	-999	0.03
CH	20020601T1315	2	S2	0	2	65.62	-999	0.03
CH	20020601T1315	2	S3	1	2	65.62	-999	0.03

Data file 2003_NCAGL_Inventory.txt contents:

Site	DateTime	Plot	Sp	PA	S	Bm	P	N
VO	20030528T1100	1	S1	1	3	16.25	3.25	0.01
VO	20030528T1100	1	S2	1	3	16.25	3.25	0.01
VO	20030528T1100	1	S3	1	3	16.25	3.25	0.01
VO	20030528T1400	2	S1	1	3	16.25	3.58	0.02
VO	20030528T1400	2	S2	1	3	16.25	3.58	0.02
VO	20030528T1400	2	S3	1	3	16.25	3.58	0.02
CH	20030529T0800	1	S1	1	3	80.34	10.23	0.02
CH	20030529T0800	1	S2	1	3	80.34	10.23	0.02
CH	20030529T0800	1	S3	1	3	80.34	10.23	0.02
CH	20030529T1115	2	S1	1	3	65.62	7.50	0.01
CH	20030529T1115	2	S2	1	3	65.62	7.50	0.01
CH	20030529T1115	2	S3	1	3	65.62	7.50	0.01

[5] Adapted from: http://knb.ecoinformatics.org/eml_metadata_guide.html. Now available at: https://web.archive.org/web/20120424124714/http://knb.ecoinformatics.org/eml_metadata_guide.html.
[6] Note the filename here – the readme file indicates that it too has changed with each addition of new data to the dataset.

Data file 2004_NCAGL_Inventory_v2.txt contents:

Site	DateTime	Plot	Sp	PA	S	Bm	P	N
VO	20040529T1030	1	S1	1	3	16.25	3.19	0.01
VO	20040529T1030	1	S2	1	3	16.25	3.19	0.01
VO	20040529T1030	1	S3	1	3	16.25	3.19	0.01
VO	20040529T1300	2	S1	1	2	16.25	3.19	0.02
VO	20040529T1300	2	S2	1	2	16.25	3.19	0.02
VO	20040529T1300	2	S3	0	2	16.25	3.19	0.02
CH	20040601T0900	1	S1	1	1	1.23	38.20	0.02
CH	20040601T0900	1	S2	0	1	1.23	38.20	0.02
CH	20040601T0900	1	S3	0	1	1.23	38.20	0.02
CH	20040601T1100	3	S1	0	0	4.68	40.93	BLD
CH	20040601T1100	3	S2	0	0	4.68	40.93	BLD
CH	20040601T1100	3	S3	0	0	4.68	40.93	BLD

readme File Contents:

Introductory Information:

Dataset Name: NCAGL_Inventory_v3_1.zip

Filenames: 2002_NCAGL_Inventory.txt
2003_NCAGL_Inventory.txt
2004_NCAGL_Inventory_v2.txt

DataSet Description: These data are annually collected as part of the Northern California Grasslands (NCAGL) diversity research program. Data collected include species richness, presence absence of plant species, peak standing biomass and nitrogen and phosphorus soil content. The relationship between diversity and productivity can take many different shapes. Soil nutrients can affect species composition, diversity and productivity. This research program attempts to investigate soil nutrients to as a possible factor in determining the shape of the diversity productivity curve, and monitors changes in this factor over time.

Data File Format: Files are tab separated text files, originally created in Microsoft Excel. Files should also be readable in any basic text editor such as Notepad, Open Office, TextEdit etc.

Related Data Files: SCAGL_Inventory_v2.zip

Principle Investigator / Data Owner: Jane Q. Researcher; Department of Biology; Northern California University; University Town, CA 95666; (123) 456-7890; researcher@uncal.edu

Associated Investigators: None

Contact Person: Sally R. Labmanager; Department of Biology; Northern California University; University Town, CA 95666; (123) 456-7891; labmanager@uncal.edu

Data Collection Dates: 20020530 - 20040601

Geographic Coverage: Data were collected in the coastal mountains of Northern California, in the Valley Oak Reserve. The Valley Oak Reserve is adjacent to and managed by Northern Caifornia University (NCU). NCU is located in University Town in Sonoma County, approximately 150 km northeast of San Francisco. Bounding coordinates are West: -120°15'00", East: -120°30'00", North: -39°15'00", South: 38°45'00"

Data File Submission Date: 20070601

Dataset Updates: All previous versions available upon request.

Original File	Date Deposited	Change Made
2002_NCAGL_Inventory.txt	20050601	(original file)
NCAGL_Inventory_v2.zip	20060601	Updated to include 2003 data
NCAGL_Inventory_v3.zip	20070601	Updated to include 2004 data
NCAGL_Inventory_v3_1.zip	20070603	Updated to correct typo in date/time stamp of 2004 data file

Keywords: richness, productivity, grasslands, biomass, northern California, soil nutrients

Methodological Information:

Methods: Twenty five 1 m^2 plots were randomly placed throughout the Valley Oak Reserve. Due to destructive biomass harvest, plots are relocated each year. Two plastic sample bags (Ziplock) are labeled with the randomly assigned plot number and contents (plant or soil). If more than one bag is needed, all bags are labeled with the contents, plot number and bag number (i.e. the second of four bags of plant clippings from plot 6 will be labeled "Plant, Plot 6, Bag 2/4").

At each plot, one person, starting at the south-east corner of the plot, identifies each plant according to the Jepson manual. Species names are recorded in the field notebook.

Species names are used to calculate species richness per plot.

All plant material within each 1 m^2 plot is clipped at soil level and placed in the sample bag. Sample bags containing plant matter are brought to the laboratory. If wet, plant matter is dried using paper towels. Plant material is dried in a drying oven at 80 degrees C for 24 (+/-2) hours. Plant matter is weighed within 2 hours of drying.

Approximately 0.5 g of soil, free from plant debris, is collected from the middle of the plot. Soil is placed in appropriate sample bag. Soil samples are placed into aluminum sample trays and placed into a drying oven and dried at 80 degrees C for 24 (+/- 2) hours. Soil is ground using a ball mill until powdery and weighed. Soil sample is analyzed for Phosphorus and Nitrogen using a SoilPro v. 10 machine. Lower limits of detection for Phosphorus is 0.005 and for Nitrogen is 0.01. All procedures for this machine are followed.

Note that in 2004, CH Plot #2 was inaccessible; plot #3, 500m NNW of Plot #2 was inventoried as a proxy.

Quality Assurance: All sampling was done by Jane Researcher, Sally Labmanager and John Fieldassistant. Data were plotted and reviewed for data entry errors by a second lab member before submission.

Column Headers: Site: Site at which data were collected.

Site	Code
Coastal Hills Reserve	CH
Valley Oaks Reserve	VO

DateTime: Date data were collected YYYYMMDDThhmm format
Plot: Randomly assigned number of plot
Sp: Species inventoried

Species Name	Code
Avena fatua	S1
Bromus hordeaceus	S2
Calochortus lutens	S3

PA: Observed presence or absence of each of three species inventoried. For each species, a value of 1 indicates presence and a value of 0 indicates absence.
S: Species richness, calculated as total number of species per plot
Bm: Peak standing biomass, measured in grams
P: Phosphorous in soil, recorded in ppm (parts per million)
N: Nitrogen in soil, recorded as a percentage; BLD indicates "Below Level of Detection" of instrument as detailed in the methods (not detected).

Missing Data: Missing data (not collected or sample lost in processing) are indicated with a "-999". "BLD" indicates measured levels are Below the Level of Detection (see Methods).

Sharing and Access Information:

Licensing: This data is freely available for re-use. Please acknowledge the Knowledge Network for Biocomplexity, NSF Grant #12345, #12346 and #12347 and Dr. Jane Researcher in any publications that use this data.

Related Publications: Reseacher, Jane Q and Labmanager, Sally R. (2003) Soil Nutrients and the Relationship between Diversity and Productivity. *Science* 5959:1234-1235.

Data Source: This dataset is publicly available through the Knowledge Network for Biocompletxity. www.knb.ecoinformatics.org

Recommended Citation: Reseacher, Jane Q and Labmanager, Sally R. (2007) Data from: Northern California Grasslands diversity research project . Knowledge Network for Biocomoplexity. http://dx.doi.org/33.3333/knb.3333t3. [7]

Funding Information: Collection of the data was funded by NSF grant #12345, #12346 and #12347.

[7] Please note that this example includes a DOI in the citation. Persistent identifiers such as DOIs, Handles, ARKs etc. are an ideal way for people to obtain shared data, but in this example, this is for illustrative purposes only, as KNB does not currently offer DOI's for the datasets.