

Introduction to Research Data Management

Data documentation

<http://hdl.handle.net/1969.1/164246>

Introduction

Focus on documenting your data to help you

- share information about your project with your advisor,
- write your thesis or dissertation,
- and ensure your data are useful later.

Why document data

Documenting your data enables:

- You to get a mess of details out of your head.
- You to reproduce and improve workflows in the future.
- Your collaborators to find the right data and use them properly.

Discussion

As you are working on a research project, what sort of information do you document?

How do you document that information?

What to document

High-level documentation explains your research goals and the progress of your project.

Low-level documentation explains the details of the data, how it has been collected, stored, and changed over time.

Project documentation

- Rationale and context for data collection.
- Research questions, goals, and hypotheses.
- Data sources, collection methodology, protocols.
- Data validation and quality assurance actions.
- Transformation of raw or derived data for integration or analysis.
- Data confidentiality, access, and use conditions.

Dataset documentation

- Variable names and descriptions.
- Codes and classification schemes.
- Algorithms used to transform data.
- Structure and organization of files.
- Relationship among data files or tables in a database schema.
- Version information.
- File formats and software used.

How to document

A few documentation tools:

- Laboratory and field notebooks
- README files
- Codebooks

Notebooks

Common in laboratory settings and fieldwork.

Document context, project, and dataset level information.

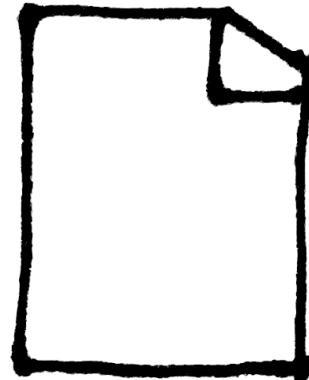
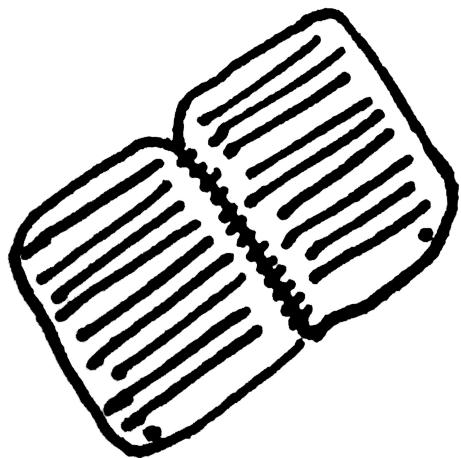
Give you one place to go throughout the research process.

Encourage a thoughtful examination.

Enable continuity in case of unexpected events or time passing.

Establish a legal and scientific provenance (historical record) of your work.

Analog and digital notes



~~3/10~~ 1-May 1999 Exp 30 cont.

Results of ligation/Transformation
- Colonies counted

Plasmid	Number	Colony	Count
①	- negative control	-	2 colonies
②	vector ①	-	0
③	" ② "	28	-
④	positive control	2300	-
⑤	- no Amp ^r (vector for construct analysis)	-	Confidence Colonies

- OK - have some "colonies" on ①
=> suggests need to check colonies from ③ for recombinant clones.
- Should have done a ligation control!!
- Will need to do a PCR
- to check colonies from ③

Robert H. Lloyd

~~3/11~~ 1-May 1999 Exp 31

PCR of colonies from Exp 30 - ③ Recombinant Clones

- To see when clones are recombinant, will do PCR with primers to MCS outside insert - will get a 130 bp fragment if no insert - and a 1.1 kb fragment if 900 bp insert is present.

- Some colonies to wire loop - Transfer to 2nd wire loop
in 50 ml Erlenmeyer
(only R.M. 18 colonies).

PCR Reaction setup	Exp. No. - SI	Date - 1-May-99	No. Reactions - 18
Builder	SI - Ligation	0.2	36
dNTP's	(100 microliters each)	-	-
MgCl ₂	75 microliters	1	18
Primer 1	1.7 microliters	1	18
Primer 2	1.6 microliters	1	18
TAQ	0.25 microliters	0.2	3.6
DNA	5 microliters	4.8	90
Water		10	183.6
TOTAL VOLUME			= 210 microliters

Reactions set up to 9 ml of mixture mix (labeled)
and 1 ml of each colony.

PCR Conditions

92°C	- 5 minutes
92°C	- 10 Sec.
58°C	- 20 Sec.
72°C	- 60 Sec.
72°C	- 10 minutes
15°C	- Hold

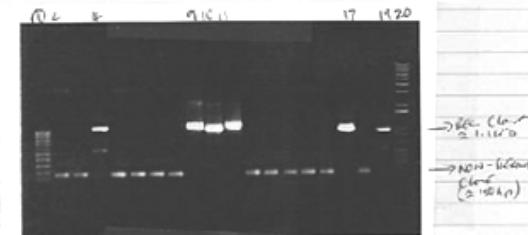
Reactions started @ 10:53 - Estimate end @ 12:47

Robert H. Lloyd

1-May 1999 Exp 31 Cont'd

0.8% Agarose gel in TAE

- Lane 1 - 100 bp ladder
- 2-14 - Sample from PCR reactions (+ 2nd loop controls)
- > 1 kb plus ladder
- 10 kb for 60 minutes



- Party Good! - 6 recombinants
- 16. looks strange - 3 in 3rd from bottom lane 2-14
- Select 4, 8, 13 & 17 for sequencing.
- Add 100 microliters + amp to the tube in the sequencing tubes.

Robert H. Lloyd
Reported for May

Note-taking software



OneNote





sciNOTE

RDM Workshop / Demo project - qPCR / My first experiment

OVERVIEW SAMPLES REPORTS

+ New experiment

Navigation:

- Demo project - qPCR
- My first experiment
 - Experiment design
 - Sampling biological...
 - RNA isolation
 - RNA quality & quantity ...
 - Reverse transcription
 - qPCR
 - Data quality control
 - Data analysis - ddCq

Edit experiment Actions Zoom: 100% 150% 200%

Experiment design Sampling biological material RNA isolation RNA quality & quantity - BIOAN...

Due date 12.10.2017 Due date 26.10.2017 Due date 09.11.2017 Due date 23.11.2017

Data analysis - ddCq Data quality control qPCR Reverse transcription

Due date 18.01.2018 Due date 04.01.2018 Due date 21.12.2017 Due date 07.12.2017





Experiment 1 — Evernote

+ New Note in Experiment 1 | New Chat | Search notes

Shortcuts

Experiment 1

Notes

Notebooks

Tags

Trash

Upgrade

All Notes E... 📁 🗂️ 🎥 🖍️

Experiment 1 plant study

Created: Sep 30, 2017 Updated: Sep 30, 2017

2017-08-29

Today Spikey the plant At 16:00 the plant is alive.



2017-08-30

Today Spikey the plant At 16:00 the plant is dead.



2017-08-30

Today Spikey the plant At 16:00 the plant is dead.





The matplotlib object-oriented API

The main idea with object-oriented programming is to have objects that one can apply functions and actions on, and no object or program states should be global (such as the MATLAB-like API). The real advantage of this approach becomes apparent when more than one figure is created, or when a figure contains more than one subplot.

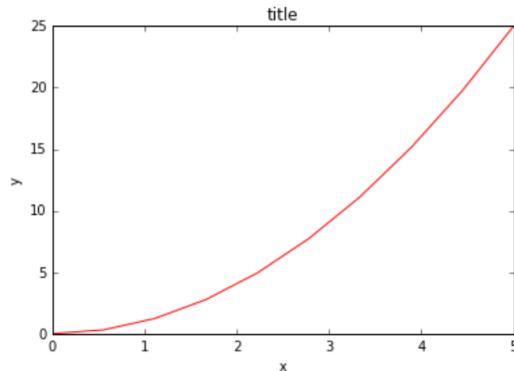
To use the object-oriented API we start out very much like in the previous example, but instead of creating a new global figure instance we store a reference to the newly created figure instance in the `fig` variable, and from it we create a new axis instance `axes` using the `add_axes` method in the `Figure` class instance `fig`:

```
In [9]: fig = plt.figure()

axes = fig.add_axes([0.1, 0.1, 0.8, 0.8]) # left, bottom, width, height (range 0 to 1)

axes.plot(x, y, 'r')

axes.set_xlabel('x')
axes.set_ylabel('y')
axes.set_title('title');
```



<h3>Regression Multiple Gaussian Targets</h3>

Now *assume* that we have K Gaussian distributed target variables, $\{t_1, \dots, t_K\}$, each with a mean that is independently conditional on \mathbf{x} , i.e. the mean of t_k is defined by some function $\mu_k(\mathbf{x})$. Also assume that all K variables share the same variance, $\sigma^2 = 1/\beta$. Assuming the network output layer has K nodes where $y_k(\mathbf{x}, \mathbf{w}) \approx \mu_k(\mathbf{x})$ and letting $\mathbf{y}(\mathbf{x}, \mathbf{w}) = [y_1(\mathbf{x}, \mathbf{w}), \dots, y_K(\mathbf{x}, \mathbf{w})]$, and that we again have N training target values \mathbf{t} (\mathbf{t} is a $K \times N$ matrix of the training values), the conditional distribution of the target training values is given by

$$p(\mathbf{t} | \mathbf{x}, \mathbf{w}) = ND(\mathbf{t} | \mathbf{y}(\mathbf{x}, \mathbf{w}), \beta^{-1} \mathbf{I})$$

The parameter estimates, $\mathbf{w}^{(ML)}$ are again found by minimizing the sum-of-squares error function, $E(\mathbf{w})$, and the estimate for β is found from

$$\frac{1}{\beta_{ML}} = \frac{1}{NK} \sum_{n=1}^N \|y(\mathbf{x}_n, \mathbf{w}^{(ML)}) - t_n\|^2$$



Regression Multiple Gaussian Targets

Now assume that we have K Gaussian distributed target variables, $\{t_1, \dots, t_K\}$, each with a mean that is independently conditional on \mathbf{x} , i.e. the mean of t_k is defined by some function $\mu_k(\mathbf{x})$. Also assume that all K variables share the same variance, $\sigma^2 = 1/\beta$. Assuming the network output layer has K nodes where $y_k(\mathbf{x}, \mathbf{w}) \approx \mu_k(\mathbf{x})$ and letting $\mathbf{y}(\mathbf{x}, \mathbf{w}) = [y_1(\mathbf{x}, \mathbf{w}), \dots, y_K(\mathbf{x}, \mathbf{w})]$, and that we again have N training target values \mathbf{t} (\mathbf{t} is a $K \times N$ matrix of the training values), the conditional distribution of the target training values is given by

$$p(\mathbf{t} | \mathbf{x}, \mathbf{w}) = ND(\mathbf{t} | \mathbf{y}(\mathbf{x}, \mathbf{w}), \beta^{-1} \mathbf{I})$$

The parameter estimates, $\mathbf{w}^{(ML)}$ are again found by minimizing the sum-of-squares error function, $E(\mathbf{w})$, and the estimate for β is found from

$$\frac{1}{\beta_{ML}} = \frac{1}{NK} \sum_{n=1}^N \|y(\mathbf{x}_n, \mathbf{w}^{(ML)}) - t_n\|^2$$

Good practices and tips

- Date each entry.
- List full names and contact information for collaborators.
- Take notes on meetings and discussions.
- Write justifications for research methods and data source(s).
- Include protocols.
- Capture conditions in the field, lab, interview.

Good practices and tips

- Note mistakes and corrections that need to be made.
- Annotate calculations with units.
- Record relevant digital file names and locations.
- Describe location of physical materials.
- Use images and print-outs to simplify documentation.

README files

Coopted from software development.

Describe the files and folders in a project.

Document mainly at the dataset level.

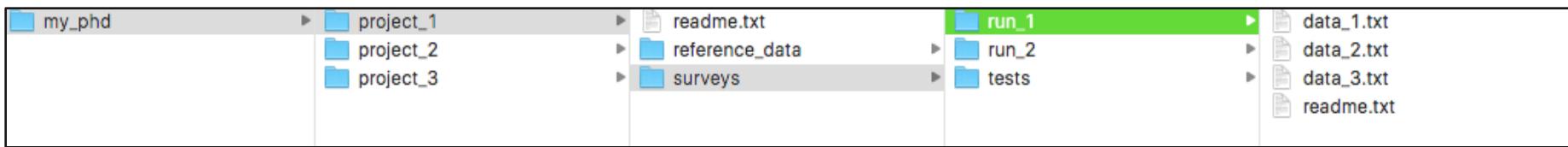
Primarily aimed at an external audience and your future self.

Example: README template

Cornell University.

<https://cornell.app.box.com/v/ReadmeTemplate>

README tips



Codebooks and data dictionaries

Document at the level of variables and data within the dataset.

Stand-alone or part of other files.

Primarily aimed at an external audience and your future self.

Example: Tabular data with a data dictionary

Data from the Duke Lemur Center about the weights of lemurs.

- Data file: <http://bit.ly/2wtleuh>
- README file: <http://bit.ly/2fL4UTH>

Example: Tabular data with a data dictionary

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Taxon	Hybrid	DLC_ID	Sex	Name	DOB	DOB_Estimated	Weight_g	Weight_Date	MonthOfWeight	AgeAtWt_d	AgeAtWt_wk	AgeAtWt_mo	AgeAtWt_mo_NoDec	AgeAtWt_y
2	CMED	N	1601	M	LINUS	16-Jul-83		77	9-Aug-83	8	24	3.43	0.79	0	0.07
3	CMED	N	1601	M	LINUS	16-Jul-83		242	21-Oct-83	10	97	13.86	3.19	3	0.27
4	CMED	N	1601	M	LINUS	16-Jul-83		258	10-Nov-83	11	117	16.71	3.85	3	0.32
5	CMED	N	1601	M	LINUS	16-Jul-83		337	8-Dec-85	12	876	125.14	28.8	28	2.4
6	CMED	N	1602	M	LUCAS	16-Jul-83		79	9-Aug-83	8	24	3.43	0.79	0	0.07
7	CMED	N	1602	M	LUCAS	16-Jul-83		176	25-Jul-93	7	3662	523.14	120.39	120	10.03
8	CMED	N	1602	M	LUCAS	16-Jul-83		177.7	7-Dec-93	12	3797	542.43	124.83	124	10.4
9	CMED	N	1602	M	LUCAS	16-Jul-83		181	31-Aug-93	8	3699	528.43	121.61	121	10.13
10	CMED	N	1602	M	LUCAS	16-Jul-83		193	26-May-93	5	3602	514.57	118.42	118	9.87
11	CMED	N	1602	M	LUCAS	16-Jul-83		194	13-Aug-91	8	2950	421.43	96.99	96	8.08
12	CMED	N	1602	M	LUCAS	16-Jul-83		194	20-Feb-97	2	4968	709.71	163.33	163	13.61
13	CMED	N	1602	M	LUCAS	16-Jul-83		196	5-Jun-95	6	4342	620.29	142.75	142	11.9
14	CMED	N	1602	M	LUCAS	16-Jul-83		199	25-Oct-91	10	3023	431.86	99.39	99	8.28
15	CMED	N	1602	M	LUCAS	16-Jul-83		200	11-Sep-96	9	4806	686.57	158.01	158	13.17
16	CMED	N	1602	M	LUCAS	16-Jul-83		200	16-Jan-97	1	4933	704.71	162.18	162	13.52
17	CMED	N	1602	M	LUCAS	16-Jul-83		200	4-Apr-97	4	5011	715.86	164.75	164	13.73
18	CMED	N	1602	M	LUCAS	16-Jul-83		206	10-Jul-96	7	4743	677.57	155.93	155	12.99
19	CMED	N	1602	M	LUCAS	16-Jul-83		207	7-Aug-96	8	4771	681.57	156.85	156	13.07
20	CMED	N	1602	M	LUCAS	16-Jul-83		208	7-Nov-96	11	4863	694.71	159.88	159	13.32
21	CMED	N	1602	M	LUCAS	16-Jul-83		213	8-Jul-92	7	3280	468.57	107.84	107	8.99
22	CMED	N	1602	M	LUCAS	16-Jul-83		219	8-Jun-95	6	4345	620.71	142.85	142	11.20
23	CMED	N	1602	M	LUCAS	16-Jul-83		221.7	2-May-84	5	291	41.57	9.57	9	0.8
24	CMED	N	1602	M	LUCAS	16-Jul-83		225	1-Apr-93	4	3547	506.71	116.61	116	9.72

DLC Weight File Variable Descriptions and Usage Notes

Table 1. List of taxa included in the data files, including taxonomic code used in all data files (Taxon), Latin name and common name of each taxon^{4, 5}.

count	Taxon	Latin_Name	Common_Name
1	CMED	<u>Cheirogaleus medius</u>	Fat-tailed dwarf lemur
2	DMAD	<u>Daubentonia madagascariensis</u>	Aye-aye
3	EALB	<u>Eulemur albifrons</u>	White-fronted brown lemur
4	ECOL	<u>Eulemur collaris</u>	Collared brown lemur
5	ECOR	<u>Eulemur coronatus</u>	Crowned lemur
6	EFLA	<u>Eulemur flavifrons</u>	Blue-eyed black lemur
7	EFUL	<u>Eulemur fulvus</u>	Common brown lemur
8	EMAC	<u>Eulemur macaco</u>	Black lemur
9	EMON	<u>Eulemur mongoz</u>	Mongoose lemur
10	ERUB	<u>Eulemur rubriventer</u>	Red-bellied lemur
11	ERUF	<u>Eulemur rufus</u>	Red-fronted brown lemur
12	ESAN	<u>Eulemur sanfordi</u>	Sanford's brown lemur
13	EUL	<u>Eulemur</u>	<u>Eulemur hybrid</u>

11	<u>AgeAtWt_d</u>	Age in days: Age of the animal when the weight was taken, in days (Weight_Date-DOB)
12	<u>AgeAtWt_wk</u>	Age in weeks: Age of the animal when the weight was taken, in weeks ((Weight_Date-DOB)/7))
13	<u>AgeAtWt_mo</u>	Age in months: Age of the animal when the weight was taken, in months (((Weight_Date-DOB)/365)*12))
14	<u>AgeAtWt_mo_NoDec</u>	Age in months with no decimal: <u>AgeAtWt_mo</u> value rounded down to a whole number for use in computing average individual weights (FLOOR(AgeAtWt_mo))
15	<u>AgeAtWt_y</u>	Age in years: Age of the animal when the weight was taken, in years ((weight date-DOB)/365))
16	<u>Days_Since_PrevWt</u>	Days difference: Difference, in days, between the date of this weight and the date of the animal's previous weight
		Weight difference: Difference, in grams, between this weight and the previous weight

Example: Data Documentation Initiative codebook

# BEDRM: Bedrooms, number of				
Information		[Type= discrete] [Format=numeric] [Range= 0-8] [Missing=*/8]		
Statistics [NW/ W]		[Valid=872972 / 32343616.174] [Invalid=14040 / 508706.825]		
Definition		Refers to all rooms designed mainly for sleeping purposes even if they are now used for other purposes, such as guest rooms and television rooms.		
Universe		Reported for: Persons in private households		
Notes		Data quality note – In the 2011 National Household Survey (NHS), a large proportion of records with 0 bedroom dwellings and 1 room dwellings was affected by respondent error (such as reporting more bedrooms than rooms). These errors were resolved during data processing and the results are consistent with other surveys. However, it is possible that in some instances or in small geographic areas the processed result is not consistent with the respondent's true situation. For more information, please consult the Housing Reference Guide, National Household Survey, Catalogue no. 99-014-X2011007.		
Value	Label	Cases	Weighted	Percentage (Weighted)
0	No bedroom	4889	178016.9	0.6%
1	1 bedroom	65590	2470756.9	7.6%
2	2 bedrooms	169285	6421315.9	19.9%
3	3 bedrooms	333708	12422735.6	38.4%
4	4 bedrooms	215433	7816595.4	24.2%
5	5 bedrooms or more	84067	3034195.4	9.4%
8	Not available	14040	508706.8	

Good practices and tips

- Variable name.
- Variable meaning.
- Variable format and how the variable was recorded.
- Units of measurement for scale variables.
- Numeric codes for categorical variables, and what they represent.
- Known issues and relationships.

Conclusion

- Identified relevant project-level and data-level documentation.
- Reviewed documentation tools including lab notebooks, README files, and codebooks.

References and resources

- Briney, Kristin. "Data Ab Initio" [Blog](<http://dataabinitio.com/?p=378>), [Blog](<http://dataabinitio.com/?p=454>)
- Cornell University. "Guide to writing "readme" style metadata" [Webpage](<https://data.research.cornell.edu/content/readme>)
- DDI. "Create a codebook" [Website](<http://www.ddialliance.org/training/getting-started-new-content/create-a-codebook>)
- DMPTool "Data Management General Guidance" [Website](https://dmptool.org/dm_guidance)
- Jupyter Team. "The Jupyter Notebook" [Website](<http://jupyter-notebook.readthedocs.io/en/stable/notebook.html>)

Break

