How to create tidy and reproducible spreadsheets

Some (very basal) good practice guidance for data handling

Roman M. Link

University of Würzburg

July 31, 2019

How datasheets often look like

kopiert werden soll.

		C	D	E	F	G	H	I	J	K	L	M	N	
1														
2	Height (m)\Sampling tree ring number from pith	H3	H2	H1	T	S	mean-H3	sd-H3	mean-H2	sd-H2	mean-H1	sd-H1	mean-T	sd-
3	0	7-9	11-13	15-18	24-25	27-30	126.8105	66.4122	127.8238	72.1244	158.4160	69.5139	134.8277	
4	1.3	6-7	10-11	14-16	22-24	25-28	156.3900	66.3379	185.7622	63.8564	192.1357	81.7497	162.2033	
5	3.8	5-6	8-9	15-16	22-23	25-27	153.3162	62.5290	192.4373	52.2249	189.2416	68.6717	186.5378	
6	6.3	3-4	7-8	12-13	17-18	19-21	152.1864	44.3087	155.4903	39.5131	165.6237	57.8282	164.6438	
7	8.8	1-4	5–7	8–9	9–10	10–14	143.8754	36.6903	164.9568	57.3062	172.7266	40.6129	164.8847	
8	11.3		1–2	3–4	5–6	7–9			130.9386	33.5562	155.0439	38.4145	163.8724	
9	0	10-11	18-20	25-26	35-37	39-43	177.4023	50.5611	175.1471	57.7171	211.6613	65.5032	199.4911	
10	1.3	8-9	14-16	22-24	30-34	36-40	124.1157	44.0238	166.6570	73.4100	107.0477	47.5352	163.6820	
11	3.8	7-8	12-14	19-21	30-32	34-37	202.0969	34.8620	193.8636	59.9496	177.2197	60.7483	226.5688	
12	6.3	6-8	11-12	15-17	23-25	27-30	209.3379	52.7908	215.6250	41.6247	219.8794	56.8417	207.2086	
13	8.8	6-8	11-12	14-15	19-21	24-27	186.9061	45.2097	207.8444	23.1942	225.1272	40.6940	224.8664	
14	11.3	6-7	10-11	13-16	18-19	21-24	165.9997	27.8235	197.1963	35.4516	216.0156	33.0659	190.6444	
15	13.8	4-5	7-9	11-12	13-14	16-17								
16	0	9-10	16-18	24-25	28-30	31-33	164.3226	39.6517	157.4609	50.5740	185.9277	71.1383	178.3198	
17	1.3	7-8	11-12	17-18	26-28	29-32	179.4890	45.3813	194.7687	48.9111	188.9610	49.6464	170.5074	
18	3.8	7-8	12-14	19-20	26-27	29-30	168.3417	42.8019	168.2942	61.7713	211.1269	67.8582	204.5259	
19	6.3	4-6	9-10	14-15	20-22	24-27	170.4199	44.6880	190.3797	42.9845	171.9804	57.7621	173.5818	
20	8.8	1–5	6–9	10–16	17	18–22	127.4685	28.3830	165.1716	49.0346	189.2385	43.7274	194.3022	
21	11.3		1–4	5–9	10	11–16			135.5505	34.5744	152.8270	48.5895	152.5273	
22														
23	Height (m)\Sampling tree ring number from pith	H3	H2	H1	T	S	mean-H3	sd-H3	mean-H2	sd-H2	mean-H1	sd-H1	mean-T	sd-
24	0	7-9	11-13	15-18	24-25	27-30	15.7414	4.0328	14.3419	3.7495	20.6260	6.2806	22.0372	
25	1.3	6-7	10-11	14-16	22-24	25-28	15.6771	5.0043	21.0863	6.2775	21.9254	6.7489	19.8336	
26	3.8	5-6	8-9	15-16	22-23	25-27	14.4573	2.5162	22.5611	4.6616	20.8423	3.4980	21.7049	
27	6.3	3-4	7-8	12-13	17-18	19-21	18.2293	5.3181	17.6986	3.9953	25.0464	8.9790	23.0737	
28	8.8	1-4	5–7	8–9	9–10	10–14	24.5550	10.7211	18.3327	6.8659	20.0307	6.2324	20.9739	
29	11.3		1–2	3–4	5–6	7–9			14.3876	5.0128	20.2123	6.2546	19.4379	
30	0	10-11	18-20	25-26	35-37	39-43	18.0858	6.4663	17.0532	6.3837	23.9316	8.8978	20.4616	
31	1.3	8-9	14-16	22-24	30-34	36-40	14.7886	4.7394	19.5521	6.0184	20.2647	7.0007	21.8543	
32	3.8	7-8	12-14	19-21	30-32	34-37	22.2514	8.4005	21.9845	5.6868	22.1407	5.3998	22.6555	
33	6.3	6-8	11-12	15-17	23-25	27-30	19.8574	7.8061	21.8686	6.4527	29.7834	7.2834	25.0980	
34	8.8	6-8	11-12	14-15	19-21	24-27	19.2573	5.3629	21.2129	7.3059	22.8607	6.0112	25.1364	
35	11.3	6-7	10-11	13-16	18-19	21-24	20.7304	6.3607	21.7833	4.9576	32.4353	6.6055	24.7970	
36	13.8	4-5	7-9	11-12	13-14	16-17								
37	0	9-10	16-18	24-25	28-30	31-33	16.1888	3.9260	17.0750	4.5888	19.0203	7.3010	32.2862	
38	1.3	7-8	11-12	17-18	26-28	29-32	15.4653	3.6609	19.1954	6.1743	18.8216	3.9954	25.9690	
39	3.8	7-8	12-14	19-20	26-27	29-30	25.8628	10.4058	19.5091	4.0772	23.1992	7.3450	26.5840	

Raw data | Note | tree infomation | ring width

Common problems

- Not **computer readable**
  - Multiple headers
  - Several tables per sheet
  - Special characters in column titles (units etc)

Common problems

- Not **computer readable**
  - Multiple headers
  - Several tables per sheet
  - Special characters in column titles (units etc)
- Not **human readable**
  - Not well documented
  - Unclear variable names
  - Messy structure spread over several sheets and/or files

The concept of "tidy data"

Happy families are all alike; every unhappy family is unhappy in its own way

Leo Tolstoy, quoted after Hadley Wickham (2014)

The concept of "tidy data"

- General structure of data: collection of *values* that belong to a *variable* and an *observation*

The concept of "tidy data"

- General structure of data: collection of *values* that belong to a *variable* and an *observation*
- Datasets are tidy when
  1. Each *variable* forms a column
  2. Each *observation* forms a row
  3. Each type of *observational unit* forms a table

The concept of "tidy data"

- General structure of data: collection of *values* that belong to a *variable* and an *observation*
- Datasets are tidy when
  1. Each *variable* forms a column
  2. Each *observation* forms a row
  3. Each type of *observational unit* forms a table
- Leading paradigm for data handling in the `tidyverse`

Minimum requirements for data input

- **Create datasets having in mind their analysis**
  - Computer-readable headers with single rows for variable names
  - Short but easy to remember column titles
  - One table per sheet
  - No unnecessary whitespace or fancy formatting (united cells etc.)

Minimum requirements for data input

- **Create datasets having in mind their analysis**
  - Computer-readable headers with single rows for variable names
  - Short but easy to remember column titles
  - One table per sheet
  - No unnecessary whitespace or fancy formatting (united cells etc.)
- **Keep the data tidy if possible**
  - One variable per column
  - One observation per row
  - One table per observational unit (e.g. tree-level vs. plot-level data)

Minimum requirements for data input

- **Create datasets having in mind their analysis**
  - Computer-readable headers with single rows for variable names
  - Short but easy to remember column titles
  - One table per sheet
  - No unnecessary whitespace or fancy formatting (united cells etc.)
- **Keep the data tidy if possible**
  - One variable per column
  - One observation per row
  - One table per observational unit (e.g. tree-level vs. plot-level data)
- **But keep in mind that...**
  - ...exceptions may make sense at the data entry stage
  - ...rearranging clean data is easy

Example

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	stem	species	pressure_leve	HFD_ratio1	HFD_ratio2	HFD_ratio3	HFD_ratio4	HFD_ratio5	HFD_ratio6	HFD_ratio7	HFD_ratio8	weighted_HFD_ratio	U1	U2	U3	U4	U5	U6	U7
2	FS5S1	FS	P150	1.012	1.089	0.976	0.634	0.412	0.223	0.151	0.122	0.935	30.522	32.848	29.562	19.216	12.528	6.817	4.659
3	FS5S1	FS	P200	1.220	1.320	1.123	0.683	0.429	0.220	0.142	0.114	1.100	38.373	41.442	35.698	21.940	13.858	7.141	4.628
4	FS5S1	FS	P250	1.655	1.800	1.444	0.824	0.495	0.250	0.164	0.130	1.455	55.327	60.292	48.309	27.543	16.546	8.402	5.553
5	FS5S1	FS	P300	2.129	2.355	1.768	0.933	0.544	0.269	0.174	0.139	1.834	74.471	82.353	61.835	32.784	19.106	9.436	6.129
6	FS5S1	FS	P350	2.703	2.958	2.081	1.036	0.592	0.290	0.186	0.147	2.257	98.226	107.323	75.648	37.687	21.473	10.558	6.817
7	FS5S1	FS	P400	3.174	3.485	2.344	1.127	0.635	0.312	0.199	0.153	2.613	118.904	130.084	87.546	41.970	23.714	11.639	7.436
8	FS5S1	FS	P450	3.714	3.998	2.619	1.220	0.691	0.336	0.214	0.159	2.995	142.545	153.450	100.644	46.952	26.625	12.955	8.290
9	FS5S1	FS	P500	4.142	4.422	2.889	1.320	0.746	0.368	0.231	0.170	3.316	162.093	172.950	113.099	51.790	29.290	14.450	9.093
10	FS5S2	FS	P150	1.104	1.200	1.035	0.691	0.474	0.259	0.167	0.138	1.017	32.186	34.952	30.616	20.694	14.303	7.804	5.032
11	FS5S2	FS	P200	1.611	1.786	1.401	0.856	0.550	0.279	0.171	0.131	1.435	50.724	55.916	44.849	27.888	18.236	9.415	5.855
12	FS5S2	FS	P250	2.142	2.393	1.750	1.005	0.630	0.325	0.201	0.157	1.861	70.607	78.791	58.659	34.056	21.452	11.075	6.885
13	FS5S2	FS	P300	2.723	3.122	2.136	1.157	0.705	0.356	0.224	0.172	2.338	94.888	108.225	75.384	41.457	25.409	12.911	8.120
14	FS5S2	FS	P350	3.274	3.800	2.537	1.309	0.780	0.389	0.238	0.180	2.796	121.006	140.378	94.114	48.661	29.041	14.497	8.937
15	FS5S2	FS	P400	3.724	4.438	2.962	1.476	0.855	0.418	0.254	0.191	3.216	145.020	172.874	115.514	57.619	33.333	16.324	9.922
16	FS5S2	FS	P450	4.094	4.998	3.471	1.656	0.926	0.443	0.268	0.194	3.610	166.647	204.698	142.001	67.701	37.892	18.104	10.920
17	FS5S2	FS	P500	4.379	5.505	4.006	1.839	0.996	0.464	0.268	0.185	3.965	186.336	234.608	170.616	78.336	42.496	19.777	11.412
18	FS5S3	FS	P150	0.741	0.701	0.648	0.495	0.367	0.258	0.182	0.178	0.673	27.064	25.730	23.826	18.261	13.621	9.616	6.836
19	FS5S3	FS	P200	1.149	1.040	0.915	0.653	0.460	0.305	0.205	0.191	0.996	45.789	41.803	37.072	26.818	19.126	12.776	8.711
20	FS5S3	FS	P250	1.435	1.259	1.078	0.744	0.511	0.324	0.215	0.196	1.210	59.544	52.639	45.455	31.649	21.935	14.042	9.383
21	FS5S3	FS	P300	1.936	1.646	1.372	0.902	0.593	0.357	0.229	0.206	1.588	83.734	71.696	60.203	39.991	26.514	16.143	10.510
22	FS5S3	FS	P350	2.348	1.958	1.632	1.028	0.657	0.385	0.243	0.211	1.900	105.253	88.027	73.402	46.652	30.013	17.672	11.238
23	FS5S3	FS	P400	2.769	2.292	1.907	1.160	0.717	0.406	0.255	0.217	2.225	127.686	106.098	88.383	53.944	33.440	18.985	11.957
24	FS5S3	FS	P450	3.090	2.552	2.139	1.257	0.767	0.425	0.262	0.222	2.477	146.377	121.148	101.671	59.756	36.470	20.203	12.484
25	FS5S3	FS	P500	3.380	2.817	2.384	1.360	0.812	0.441	0.273	0.227	2.721	164.599	136.906	115.946	66.224	39.541	21.493	13.300
26	FS5S3	FS	P550	3.613	3.059	2.627	1.449	0.852	0.457	0.280	0.231	2.934	180.127	152.344	130.950	72.409	42.648	22.880	14.060
27	FS6S2	FS	P150	0.928	1.148	1.241	0.903	0.634	0.352	0.229	0.163	1.034	29.515	36.496	39.600	28.813	20.233	11.278	7.399
28	FS6S2	FS	P200	1.309	1.688	1.822	1.191	0.771	0.402	0.247	0.167	1.472	42.423	53.986	58.363	39.426	26.174	13.931	8.785
29	FS6S2	FS	P250	1.818	2.476	2.679	1.586	0.970	0.488	0.298	0.200	2.090	57.871	77.896	84.347	50.849	31.431	15.968	9.805
30	FS6S2	FS	P300	2.292	3.301	3.688	1.946	1.126	0.541	0.321	0.213	2.726	72.470	102.878	114.158	62.092	36.457	17.760	10.646
31	FS6S2	FS	P350	2.667	4.017	4.665	2.266	1.250	0.580	0.335	0.219	3.285	83.068	123.829	142.429	70.893	39.628	18.567	10.833
32	FS6S2	FS	P400	2.985	4.622	5.654	2.514	1.349	0.610	0.348	0.225	3.785	91.192	139.840	168.245	77.059	41.851	19.058	10.937
33	FS6S2	FS	P450	3.237	5.157	6.570	2.787	1.442	0.644	0.360	0.232	4.235	97.182	153.111	192.372	83.240	43.595	19.556	11.034
34	FS6S2	FS	P500	3.442	5.544	7.372	2.978	1.529	0.666	0.372	0.237	4.595	100.529	161.636	211.715	87.153	44.834	19.696	11.028
35	FS6S2	FS	P550	3.641	5.882	8.218	3.174	1.623	0.691	0.379	0.240	4.952	103.702	167.474	230.055	90.324	46.279	19.863	10.960

Metadata Data

Why metadata?

- Your data are valuable: they are the main outcome of your scientific work

Why metadata?

- Your data are valuable: they are the main outcome of your scientific work
- If data are not well documented, they become worthless very fast

Why metadata?

- Your data are valuable: they are the main outcome of your scientific work
- If data are not well documented, they become worthless very fast
- To make them last, datasheets have to be created having in mind...

Why metadata?

- Your data are valuable: they are the main outcome of your scientific work
- If data are not well documented, they become worthless very fast
- To make them last, datasheets have to be created having in mind. . .
    - you

Why metadata?

- Your data are valuable: they are the main outcome of your scientific work
- If data are not well documented, they become worthless very fast
- To make them last, datasheets have to be created having in mind...
    - you
    - future you

Why metadata?

- Your data are valuable: they are the main outcome of your scientific work
- If data are not well documented, they become worthless very fast
- To make them last, datasheets have to be created having in mind...
  - you
  - future you
  - others

Why metadata?

- Your data are valuable: they are the main outcome of your scientific work
- If data are not well documented, they become worthless very fast
- To make them last, datasheets have to be created having in mind...
  - you
  - future you
  - others

$\Rightarrow$ Need for documentation

Minimum requirements for metadata

- **What** was measured?

Minimum requirements for metadata

- **What** was measured?
- **Who** did the work? (contact)

Minimum requirements for metadata

- **What** was measured?
- **Who** did the work? (contact)
- **Where** and **when** was it done?

Minimum requirements for metadata

- **What** was measured?
- **Who** did the work? (contact)
- **Where** and **when** was it done?
- **How** was it measured?
    - Equipment
    - Software
    - Equations used for calculation
    - Data source of publicly available datasets
    - Measurement units

Minimum requirements for metadata

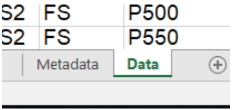
- **What** was measured?
- **Who** did the work? (contact)
- **Where** and **when** was it done?
- **How** was it measured?
  - Equipment
  - Software
  - Equations used for calculation
  - Data source of publicly available datasets
  - Measurement units
- **Additional information**, e.g.
  - Associated publications (must-cites)
  - Licenses for data use

How to document metadata

- Fancy method: associated meta data (e.g. .xml) files linked to each dataset in a large a database

How to document metadata

- Fancy method: associated meta data (e.g. .xml) files linked to each dataset in a large a database
- Day-to-day method: add a metadata sheet to each of your datasets in Calc, Excel etc...



Example

	A	B	C	D
1	**Description**			
2	*Calibration dataset for measurements performed with the HFD method (Nadezhdina et al., 1998, 2000).*			
3	*Measurements were performed by Sebastian Fuchs in summer 2014 using a HFD50 Sensor (ICT International Pty Ltd., Armidale, Australia),*			
4	*using stem segments with 1,4 m length and around 8.6-16 cm diameter. Water flow was induced by applying subatmospheric pressure*			
5	*to the upper end of the stem, and flow rates were tracked with a balance and consecutively converted to flux densities using sample length*			
6	*and sapwood areas based from cross-sections of paint-perfused stems.*			
7	*Data were aggregated over 1-hour measurement intervals at different pressure levels, excluding each the observations from the first 15 min*			
8	*(which were affected by a time-lag in pressure equilibration). Additionally, the first three pressure levels (0 mbar, 50 mbar and 100 mbar) were*			
9	*excluded since they were below the gravitative water potential over the stem length, which led to rather inconsistent measurements.*			
10	*Given are predictions of sap flux density based on the original equation of Nadezhdina et. al. (1998, 2012) from the ICP sap flow tool*			
11	*and new predictions of sap flux density based on our (species specific vs. species-independent) calibration equations (statistical analysis:*			
12	*R. Link, 2016).*			
13				
14	*Additional information about data collection and processing can be found in our paper:*			
15	*Fuchs, S., Leuschner, C., Link, R., Coners, H., & Schuldt, B., 2017: Calibration and comparison of thermal dissipation,*			
16	*heat ratio and heat field deformation sap flow probes for diffuse-porous trees. Agricultural and Forest Meteorology.*			
17				
18	**Variable**	**Unit**	**Description**	**Example**
19	**stem**	*NA*	*Unique identifier for each analyzed stem*	FS5S1
20	**species**	*NA*	*Species (FS = Fagus sylvatica, TC = Tilia cordata, AP = Acer pseudoplatanus)*	FS
21	**pressure_level**	*NA*	*Pressure level (number codes for the intended average suction in mbar)*	P150
22	**HFD_ratio1**	*unitless*	*HFD ratio readings from the HFD sensor at 0.5 cm depth.*	1.012
23	**HFD_ratio2**	*unitless*	*HFD ratio readings from the HFD sensor at 1.0 cm depth.*	1.089
24	**HFD_ratio3**	*unitless*	*HFD ratio readings from the HFD sensor at 1.5 cm depth.*	0.976
25	**HFD_ratio4**	*unitless*	*HFD ratio readings from the HFD sensor at 2.0 cm depth.*	0.634
26	**HFD_ratio5**	*unitless*	*HFD ratio readings from the HFD sensor at 2.5 cm depth.*	0.412
27	**HFD_ratio6**	*unitless*	*HFD ratio readings from the HFD sensor at 3.0 cm depth.*	0.223
28	**HFD_ratio7**	*unitless*	*HFD ratio readings from the HFD sensor at 3.5 cm depth.*	0.151
29	**HFD_ratio8**	*unitless*	*HFD ratio readings from the HFD sensor at 4.0 cm depth.*	0.122
30	**weighted_HFD_ratio**	*unitless*	*Weighted average of the HFD ratios at different depths weighted by the area contribution of the corresponding annulus to the total stem area (see paper for details)*	0.935
31	**U1**	*10^{-6}*m^3*m^2*s^{-1}*	*Calculated sap flux density for Sensor 1 from sap flow tool (based on the method of Nadezhdina et al., 1998, 2012)*	30.522

Data import and preparation in R

- Import is very easy if your data are tidy

Data import and preparation in R

- Import is very easy if your data are tidy
- Tools for data import
  - Base R: `read.csv()`, `read.table()`
  - Packages: `readr`, `readxl`, `gdata` etc.

Data import and preparation in R

- Import is very easy if your data are tidy
- Tools for data import
  - Base R: `read.csv()`, `read.table()`
  - Packages: `readr`, `readxl`, `gdata` etc.
- Tools for data wrangling
  - Packages from the `tidyverse`: `dplyr`, `tidyr` etc.

Data import and preparation in R

- Import is very easy if your data are tidy
- Tools for data import
  - Base R: `read.csv()`, `read.table()`
  - Packages: `readr`, `readxl`, `gdata` etc.
- Tools for data wrangling
  - Packages from the `tidyverse`: `dplyr`, `tidyr` etc.
- Most common problems
  - Font encoding (Windows, I am looking at you!)
  - Decimal and field separators
  - Formatting errors
  - Data entry errors (e.g. text in a numeric column)

Room for questions