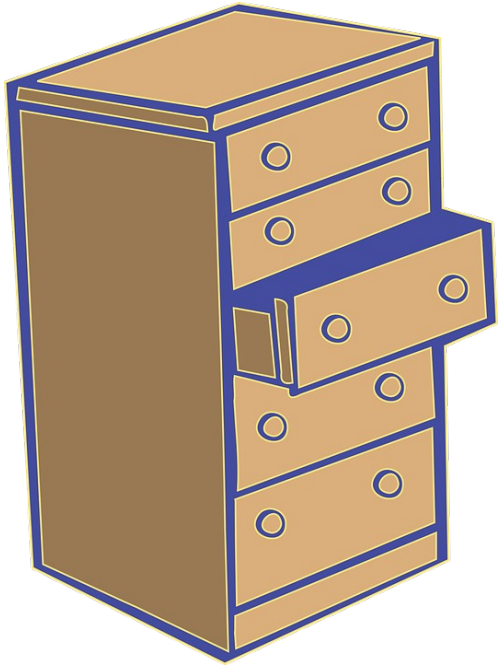
Data normalisation and formats

Sarah Faulwetter & Pieter Provoost

A Practical introduction to acquisition, validation, quality control and access to (biodiversity) data

Vlissingen, Netherlands, 14.06.2015

Purpose



Storage & sharing

Analyses

Format depends on software, type of analysis, etc.

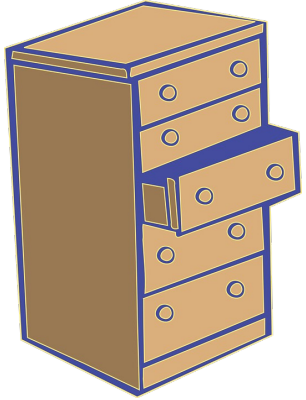
Format depends on software, type of analysis, etc.

	A	B	C	D	E	F	
1	scientificName	Cardigan Bay	Celtic Deep	Central Channel	Humber	Irish Sea	Li
2	Abra alba	23	159	9	25		
3	Abra nitida						
4	Abra prismatica						
5	Achelia echinata						
6	Acidostoma obesum						
7	Acidostoma sarsi						
8	Acteon tornatilis						
9	Aega crenulata			7			
10	Aequipecten opercularis			4			
11	Aglaophamus rubella					2	
12	Ammodytes marinus						
13	Ampelisca brevicornis						
14	Ampelisca diadema						
15	Ampelisca macrocephala		3				
16	Ampelisca spinipes			10			
17	Ampelisca tenuicorni	1	6				

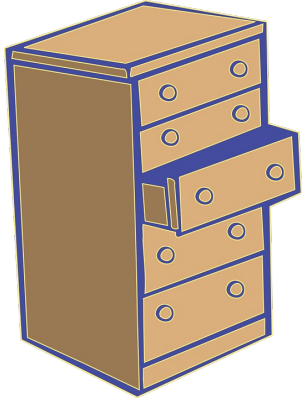
Format depends on software, type of analysis, etc.

	A	B	C	D	E	F
1	scientificName	Cardigan Bay	Celtic Deep	Central Channel	Humber	Irish Sea
2	Abra alba	23	159	9	25	
3	Abra nitida					
4	Abra prismatica					
11	Aglaophamus rubella					2
12	Ammodytes marinus					
13	Ampelisca brevicornis					
14	Ampelisca diadema					
15	Ampelisca macrocephala		3			
16	Ampelisca spinipes			10		
17	Ampelisca tenuicorni	1	6			

limited to three types of information



taxonNameAsInFile	samplingLevel	value
Hydrobia	subtidal	1
Macoma balthica	subtidal	2
Mytilus	subtidal	1
Bathyporeia pilosa	subtidal	9
Polychaeta	subtidal	34
Mytilus	subtidal	1
Polychaeta	subtidal	10
Mytilus	subtidal	1
Bathyporeia pilosa	subtidal	36
Polychaeta	subtidal	23
Polychaeta	subtidal	14
Hydrobia	subtidal	3
Bathyporeia pilosa	subtidal	19
Polychaeta	subtidal	9
Hydrobia	subtidal	1
Bathyporeia pilosa	subtidal	20
Neomysis integer	subtidal	1
Polychaeta	subtidal	35
Hydrobia	subtidal	1
Bathyporeia pilosa	subtidal	16
Polychaeta	subtidal	20
Nephtys	intertidal	1
Oligochaeta	intertidal	1



taxonNameAsInFile	samplingLevel	value	sampleArea	samples.unit	replicate(	samplingMinDepth	sampledSedimentD
Hydrobia	subtidal	1	133	cm2	1	0.5	30
Macoma balthica	subtidal	2	133	cm2	1	0.5	30
Mytilus	subtidal	1	133	cm2	1	0.5	30
Bathyporeia pilosa	subtidal	9	133	cm2	1	0.5	30
Polychaeta	subtidal	34	133	cm2	1	0.5	30
Mytilus	subtidal	1	133	cm2	2	0.5	30
Polychaeta	subtidal	10	133	cm2	2	0.5	30
Mytilus	subtidal	1	133	cm2	3	0.5	30
Bathyporeia pilosa	subtidal	36	133	cm2	3	0.5	30
Polychaeta	subtidal	23	133	cm2	3	0.5	30
Polychaeta	subtidal	14	133	cm2	1	0.5	30
Hydrobia	subtidal	3	133	cm2	1	0.5	30
Bathyporeia pilosa	subtidal	19	133	cm2	1	0.5	30
Polychaeta	subtidal	9	133	cm2	1	0.5	30
Hydrobia	subtidal	1	133	cm2	2	0.5	30
Bathyporeia pilosa	subtidal	20	133	cm2	2	0.5	30
Neomysis integer	subtidal	1	133	cm2	2	0.5	30
Polychaeta	subtidal	35	133	cm2	2	0.5	30
Hydrobia	subtidal	1	133	cm2	2	0.5	30
Bathyporeia pilosa	subtidal	16	133	cm2	2	0.5	30
Polychaeta	subtidal	20	133	cm2	2	0.5	30
Nephtys	intertidal	1	133	cm2	2		30
Oligochaeta	intertidal	1	133	cm2	2		30

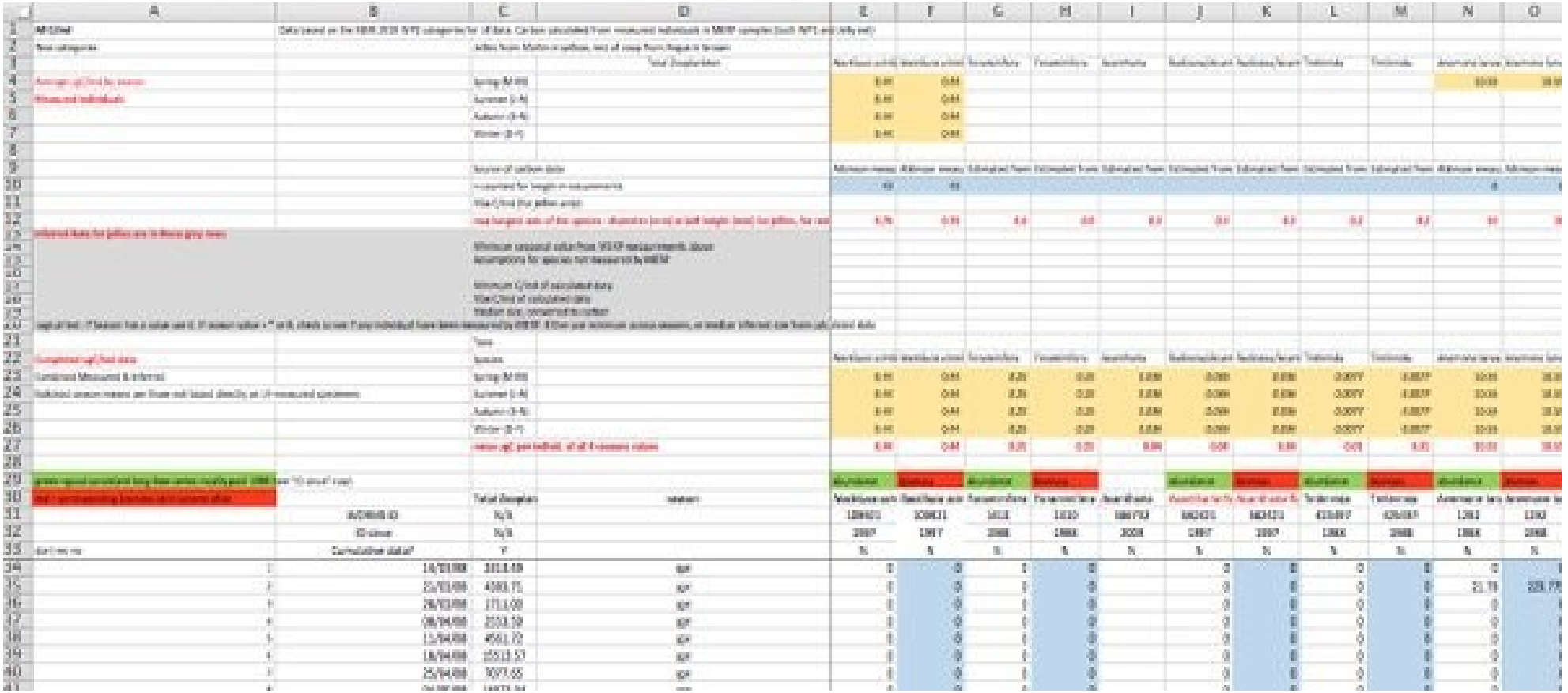
What's the most used tool for data sharing in biology?

Excel.

What's the most hated tool for data sharing?

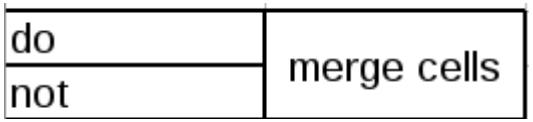
Excel.

—— Bruce Wilson, DataONE



source: @tomjwebb

1



1

do	merge cells
not	

2

do not	use
colour	coding

1

do not	merge cells

2

do not	use
colour	coding

3

use “talking” column headers

st.	ab_t	d	s
B1	12	0.33	4
B2	5	0.135	6

station	abundanceTotal	density	nrOfSpecies
B1	12	0.33	4
B2	5	0.135	6

1

do not	merge cells

2

do not	use
colour	coding

3

use “talking” column headers

st.	ab_t	d	s
B1	12	0.33	4
B2	5	0.135	6

station	abundanceTotal	density	nrOfSpecies
B1	12	0.33	4
B2	5	0.135	6

4

include units:

B	C
Temp_C	Oxygen_%
12	86.9

1

do not	merge cells

2

do not	use
colour	coding

3 use “talking” column headers

st.	ab_t	d	s
B1	12	0.33	4
B2	5	0.135	6

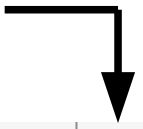
station	abundanceTotal	density	nrOfSpecies
B1	12	0.33	4
B2	5	0.135	6

4 include units:

B	C
Temp_C	**Oxygen_%**
12	86.9

5 Only one piece of information per cell

B
Location
Crete, Greece
Sicily, Italy



F	G
Location	**Country**
Crete	Greece
Sicily	Italy

1

do not	merge cells

2

do not	use
colour	coding

3

use “talking” column headers

st.	ab_t	d	s
B1	12	0.33	4
B2	5	0.135	6

station	abundanceTotal	density	nrOfSpecies
B1	12	0.33	4
B2	5	0.135	6

4

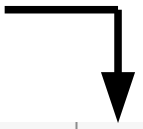
include units:

B	C
Temp_C	**Oxygen_%**
12	86.9

5

Only one piece of information per cell

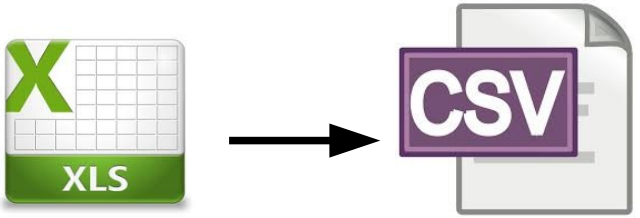
B
Location
Crete, Greece
Sicily, Italy



F	G
Location	**Country**
Crete	Greece
Sicily	Italy

6

final version not in proprietary format



atomization

Name	Contact details
Robert Ingram	robert@ingram.com, New York, Tel. 555-861-2025
Jane Wright	jwright@gmail.com, New York, Tel. 555-861-2033
Maria Fernandez	mfernandez@gmail.com, Chicago, Tel. 393-3345-2235

Name	Contact details
Robert Ingram	robert@ingram.com, New York, Tel. 555-861-2025
Jane Wright	jwright@gmail.com, New York, Tel. 555-861-2033
Maria Fernandez	mfernandez@gmail.com, Chicago, Tel. 393-3345-2235

Sort by last name

All employees from New York

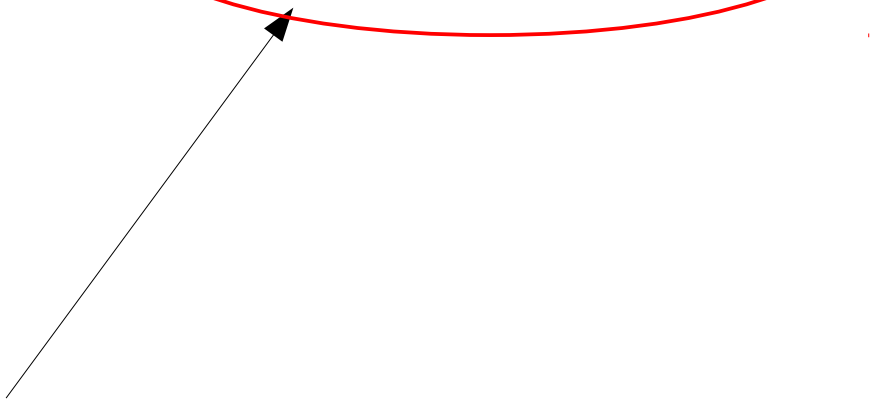
→ Problem

Name	Contact details
Robert Ingram	robert@ingram.com, New York, Tel. 555-861-2025
Jane Wright	jwright@gmail.com, New York, Tel. 555-861-2033
Maria Fernandez	mfernandez@gmail.com, Chicago, Tel. 393-3345-2235

First name	Last name	e-mail	City	Phone Number
Robert	Ingram	robert@ingram.com	New York	555-861-2025
Jane	Wright	jwright@gmail.com	New York	555-861-2033
Maria	Fernandez	mfernandez@gmail.com	Chicago	393-3345-2235

First name	Last name	e-mail	City	Phone Number
Robert	Ingram	robert@ingram.com	New York	555-861-2025
Jane	Wright	jwright@gmail.com	New York	555-861-2033
Maria	Fernandez	mfernandez@gmail.com, mfernandez@company.com	Chicago	393-3345-2235

First name	Last name	e-mail	City	Phone Number
Robert	Ingram	robert@ingram.com	New York	555-861-2025
Jane	Wright	jwright@gmail.com	New York	555-861-2033
Maria	Fernandez	mfernandez@gmail.com, mfernandez@company.com	Chicago	393-3345-2235



Bad format!

First name	Last name	e-mail	City	Phone Number
Robert	Ingram	robert@ingram.com	New York	555-861-2025
Jane	Wright	jwright@gmail.com	New York	555-861-2033
Maria	Fernandez	mfernandez@gmail.com	Chicago	393-3345-2235
Maria	Fernandez	mfernandez@company.com	Chicago	393-3345-2235

First name	Last name	e-mail	City	Phone Number
Robert	Ingram	robert@ingram.com	New York	555-861-2025
Jane	Wright	jwright@gmail.com	New York	555-861-2033
Maria	Fernandez	mfernandez@gmail.com	Chicago	393-3345-2235
Maria	Fernandez	mfernandez@company.com	Chicago	393-3345-2235

Bad format!
(Duplicate information)

First name	Last name	e-mail	City	Phone Number
Robert	Ingram	robert@ingram.com	New York	555-861-2025
Jane	Wright	jwright@gmail.com	New York	555-861-2033
Maria	Fernandez	mfernandez@gmail.com	Chicago	393-3345-2235
Maria	Fernandez	mfernandez@company.com	Chicago	393-3345-2235

Bad format!
(Duplicate information)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Species/Station**	**I- 50**	**I-60**	**I-70**	**I-72**	**I-75**	**I-77**	**I-79**	**I-81**	**I-83**	**I-87**	**I-89**	**I-91**
2													
3	**PORIFERA**												
4	Axinella domicornis	0	0	1	0	0	0	0	0	0	0	0	0
5	Suberites carnosus	0	0	0	0	0	0	0	0	0	0	0	0
6	'----------												
7	**HYDROZOA**												
8	Laomedea angulata	0	0	0	0	0	0	0	0	0	0	0	0
9	Obelia geniculata	0	0	0	0	0	0	0	0	0	0	0	0
10	Podocoryna carnea	0	0	0	0	0	0	0	0	0	0	0	0
11	'----------												
12	**ANTHOZOA**												
13	Amphianthus dohrnis	1	0	0	0	0	0	0	1	0	1	0	0
14	Edwardsia calimorpha	0	0	0	0	0	0	0	0	0	0	0	0
15	Sarcodictyon coralloides	0	0	0	0	0	0	0	0	0	0	0	0
16	'----------												
17	**TUBELLARIA**												
18	Stylochus pilidium	0	0	0	0	0	0	0	0	0	0	0	0
19	'----------												
20	**NEMERTINEA**												
21	Sp. 1	0	0	2	0	0	0	0	0	0	0	2	0
22	Sp. 2	0	2	3	0	0	0	0	0	0	0	0	0
23	Sp. 3	0	0	0	0	0	0	0	0	0	0	0	0
24	'----------												
25	**POLYCHAETA**												
26	Arabella iricolor	0	0	0	1	0	0	0	0	0	0	0	0
27	Drilonereis filum	0	0	0	0	0	0	0	0	0	0	0	0
28	Eteone siphonodonta	0	0	1	0	0	0	2	0	0	0	0	0

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Species/Station**	**I- 50**	**I-60**	**I-70**	**I-72**	**I-75**	**I-77**	**I-79**	**I-81**	**I-83**	**I-87**	**I-89**	**I-91**
2													
3	**PORIFERA**												
4	Axinella domicornis	0	0	1	0	0	0	0	0	0	0	0	0
5	Suberites carnosus	0	0	0	0	0	0	0	0	0	0	0	0
6	'----------												
7	**HYDROZOA**												
8	Laomedea angulata	0	0	0	0	0	0	0	0	0	0	0	0
9	Obelia geniculata	0	0	0	0	0	0	0	0	0	0	0	0
10	Podocoryna carnea	0	0	0	0	0	0	0	0	0	0	0	0
11	'----------												
12	**ANTHOZOA**												
13	Amphianthus dohrnis	1	0	0	0	0	0	0	1	0	1	0	0
14	Edwardsia calimorpha	0	0	0	0	0	0	0	0	0	0	0	0
15	Sarcodictyon coralloides	0	0	0	0	0	0	0	0	0	0	0	0
16	'----------												
17	**TUBELLARIA**												
18	Stylochus pilidium	0	0	0	0	0	0	0	0	0	0	0	0
19	'----------												
20	**NEMERTINEA**												
21	Sp. 1	0	0	2	0	0	0	0	0	0	0	2	0
22	Sp. 2	0	2	3	0	0	0	0	0	0	0	0	0
23	Sp. 3	0	0	0	0	0	0	0	0	0	0	0	0
24	'----------												
25	**POLYCHAETA**												
26	Arabella iricolor	0	0	0	1	0	0	0	0	0	0	0	0
27	Drilonereis filum	0	0	0	0	0	0	0	0	0	0	0	0
28	Eteone siphonodonta	0	0	1	0	0	0	2	0	0	0	0	0

1: “Headers” in file 2: graphical elements, not data 3: Data only understandable in combination with the “header” 4: White spaces in or before taxon names

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Species/Station**	**I- 50**	**I-60**	**I-70**	**I-72**	**I-75**	**I-77**	**I-79**	**I-81**	**I-83**	**I-87**	**I-89**	**I-91**
2	'----------												
3	'----------												
4	'----------												
5	'----------												
6	'----------												
7	Amphianthus dohrnis	1	0	0	0	0	0	0	1	0	1	0	0
8	**ANTHOZOA**												
9	Arabella iricolor	0	0	0	1	0	0	0	0	0	0	0	0
10	Axinella domicornis	0	0	1	0	0	0	0	0	0	0	0	0
11	Drilonereis filum	0	0	0	0	0	0	0	0	0	0	0	0
12	Edwardsia calimorpha	0	0	0	0	0	0	0	0	0	0	0	0
13	Eteone siphonodonta	0	0	1	0	0	0	2	0	0	0	0	0
14	**HYDROZOA**												
15	Laomedea angulata	0	0	0	0	0	0	0	0	0	0	0	0
16	**NEMERTINEA**												
17	Obelia geniculata	0	0	0	0	0	0	0	0	0	0	0	0
18	Podocoryna carnea	0	0	0	0	0	0	0	0	0	0	0	0
19	**POLYCHAETA**												
20	**PORIFERA**												
21	Sarcodictyon coralloides	0	0	0	0	0	0	0	0	0	0	0	0
22	Sp. 1	0	0	2	0	0	0	0	0	0	0	2	0
23	Sp. 2	0	2	3	0	0	0	0	0	0	0	0	0
24	Sp. 3	0	0	0	0	0	0	0	0	0	0	0	0
25	Stylochus pilidium	0	0	0	0	0	0	0	0	0	0	0	0
26	Suberites carnosus	0	0	0	0	0	0	0	0	0	0	0	0
27	**TUBELLARIA**												
28													

Same file, sorted alphabetically