DIS08 – Data Modeling

05 – Open Data, data file formats, spreadsheets, data cleaning

Philipp Schaer, Technische Hochschule Köln, Cologne, Germany

Version: SS2020

Technology
Arts Sciences
TH Köln

Disclaimer

This lesson is based on the Library Carpentry

- https://librarycarpentry.org/lc-spreadsheets/

Chapter 14 of Automating the Boring Stuff

And some slides by Kai Dührkop

- https://bio.informatik.uni-jena.de/wp/wp-content/uploads/2015/03/web.pdf

Agenda

Last week

- Regular expressions
- Mining or searching in files

This week

- Open Data and where to get it
- Data formats: CSV, JSON, XML
- Working with CSV data
- Data cleaning in Excel in a nutshell

Next week

- Hands-on Python

Definition of „Openness“

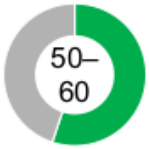
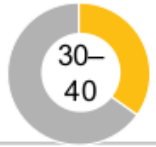
- “Open means **anyone** can **freely access, use, modify, and share** for **any purpose** (subject, at most, to requirements that preserve provenance and openness).”

- “Open data and content can be **freely used, modified, and shared** by **anyone** for **any purpose**”

  - Open Data
  - Open Access
  - Open Science
  - Open Source
  - Open Government ...

http://opendefinition.org/

Potential Impact of Open Data

	Potential impact: 2011 research	Value captured %	Major barriers
Location-based data	▪ $100 billion+ revenues for service providers ▪ Up to $700 billion value to end users		▪ Penetration of GPS-enabled smartphones globally
US retail[1]	▪ 60%+ increase in net margin ▪ 0.5–1.0% annual productivity growth		▪ Lack of analytical talent ▪ Siloed data within companies
Manufacturing[2]	▪ Up to 50% lower product development cost ▪ Up to 25% lower operating cost ▪ Up to 30% gross margin increase		▪ Siloed data in legacy IT systems ▪ Leadership skeptical of impact
EU public sector[3]	▪ ~€250 billion value per year ▪ ~0.5% annual productivity growth		▪ Lack of analytical talent ▪ Siloed data within different agencies
US health care	▪ $300 billion value per year ▪ ~0.7% annual productivity growth		▪ Need to demonstrate clinical utility to gain acceptance ▪ Interoperability and data sharing

1 Similar observations hold true for the EU retail sector.
2 Manufacturing levers divided by functional application.
3 Similar observations hold true for other high-income country governments.

SOURCE: Expert interviews; McKinsey Global Institute analysis

Open Government

Barack Obama, 21 January 2009 (his first day in the office): "Memorandum for the Heads of Executive Departments and Agencies"

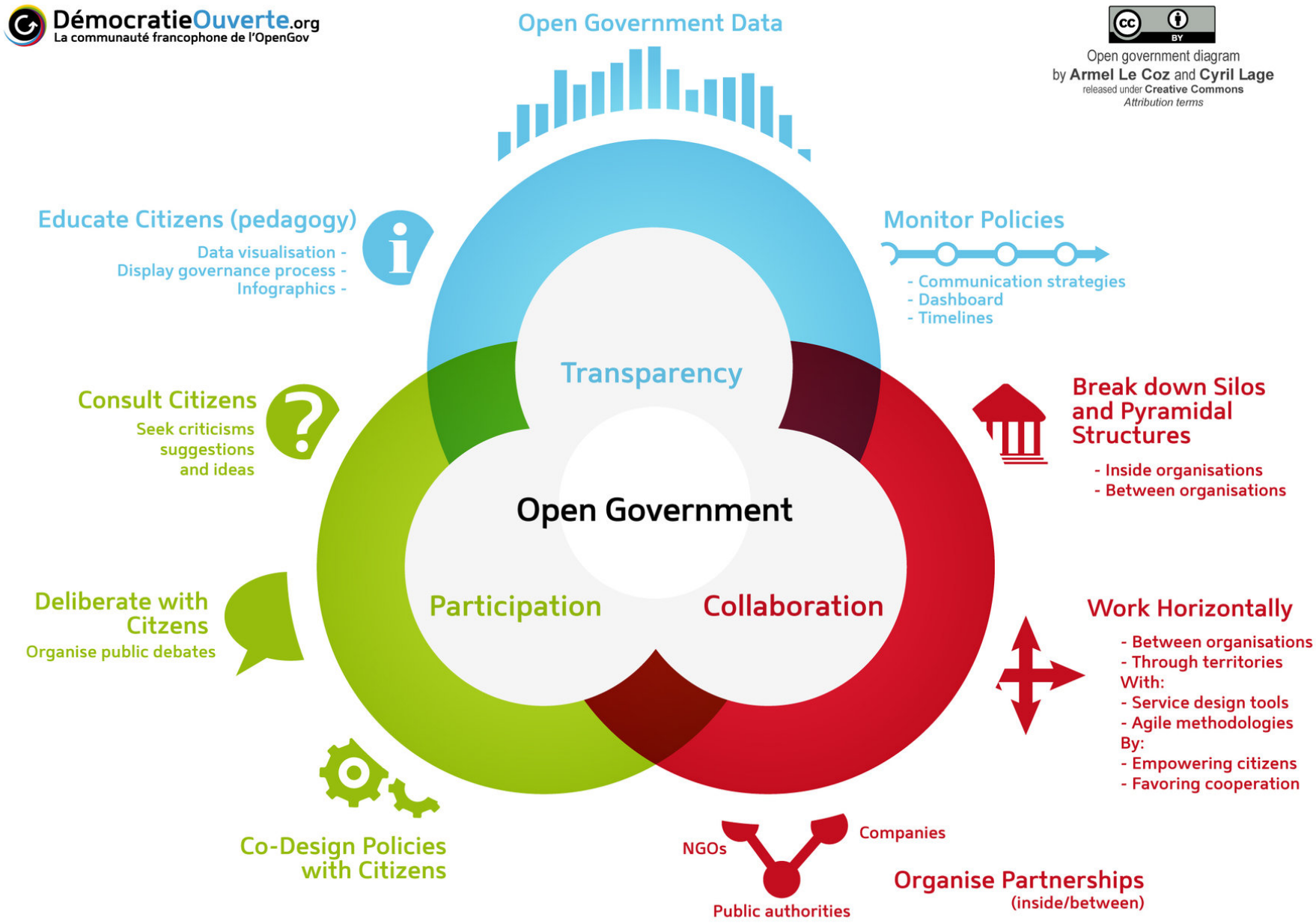
- **Transparency and Open Government**
  - Government should be transparent.
  - Government should be participatory.
  - Government should be collaborative.
- **Freedom of Information Act (FOIA)**
  - "In the face of doubt, openness prevails.“
  - "commitment to accountability and transparency"

https://de.slideshare.net/nice/vorlesung-open-data-informationen-zur-vorlesung-und-einfhrung-ins-thema-open-data-und-open-government

Open Data Principles

- **Complete** *All public data is made available.*
- **Primary** *Data is as collected at the source, with the highest possible level of granularity, not in aggregate/modified forms.*
- **Timely** *Data is made available as quickly as necessary to preserve the value of the data.*
- **Accessible** *Data is available to the widest range of users for the widest range of purposes.*
- **Machine processable** *Data is reasonably structured to allow automated processing.*
- **Non-discriminatory** *Data is available to anyone, with no requirement of registration.*
- **Non-proprietary** *Data is available in a format over which no entity has exclusive control.*
- **License-free** *Data is not subject to any copyright, patent, …*

https://public.resource.org/8_principles.html


DémocratieOuverte.org
La communauté francophone de l'OpenGov
Open Government Data
Open government diagram
by Armel Le Coz and Cyril Lage
released under Creative Commons
Attribution terms
Educate Citizens (pedagogy)
Data visualisation -
Display governance process -
Infographics -
Monitor Policies
- Communication strategies
- Dashboard
- Timelines
Transparency
Consult Citizens
Seek criticisms
suggestions
and ideas
Break down Silos
and Pyramidal
Structures
- Inside organisations
- Between organisations
Open Government
Deliberate with
Citzens
Organise public debates
Participation
Collaboration
Work Horizontally
- Between organisations
- Through territories
With:
- Service design tools
- Agile methodologies
By:
- Empowering citizens
- Favoring cooperation
Co-Design Policies
with Citizens
NGOs
Companies
Public authorities
Organise Partnerships
(inside/between)

Open Government in Germany

Der Bundestag hat am **Donnerstag, 18. Mai 2017**, das sogenannte **E-Government-Gesetz** geändert. Dem Gesetzentwurf der Bundesregierung (18/11614) in der vom Innenausschuss geänderten Fassung (18/12406) stimmten die Koalitionsfraktionen zu, die Opposition enthielt sich. Mit dem Gesetz werde die Grundlage für die aktive Bereitstellung von elektronischen Daten der Behörden der unmittelbaren Bundesverwaltung geschaffen, schreibt die Regierung. Um dem Anspruch auf eine Vorreiterrolle Deutschlands gerecht zu werden, orientiert sich die Regelung nach eigenen Angaben an international anerkannten Open-Data-Prinzipien, wie sie beispielsweise in der Internationalen Open-Data-Charta (IODC) oder in der Open-Data-Charta der sogenannten G8-Staaten beschrieben werden.

https://www.bundestag.de/dokumente/textarchiv/2017/kw20-de-e-government/505120

What to do with open public data...

OPEN.NRW
AKTUELLE PROJEKTE

RHEINSPANGE553
WWW.STRASSEN.NRW.DE

Frühzeitige Beteiligung zu neuer Rheinquerung

Der Landesbetrieb Straßenbau NRW plant derzeit eine Rheinquerung zwischen Köln und Bonn. Der konkrete Verlauf ist noch offen und wird unter Beteiligung der Menschen und Interessengruppen der...

Verwaltung einfacher machen

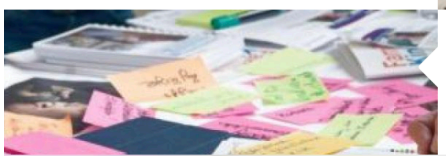
Wo ist künftig noch eine Unterschrift nötig und wann muss man persönlich erscheinen? Zwischen August und November 2017 konnten Interessierte sich beteiligen und helfen, bürokratische Hürden...

Arbeit im digitalen Wandel – Ihre Ideen für NRW sind gefragt

Die Digitalisierung verändert unsere Arbeitswelt massiv. Doch wie lassen sich digitaler Wandel und die Grundsätze guter und fairer Arbeit in Einklang bringen? Diskutieren und kommentieren Sie im...

Vernetzte Nachbarschaften

Grünflächen, Geschäfte, eine gute Verkehrsanbindung: Wer sein Wohnumfeld gestalten möchte, braucht Unterstützung und Partner. Deshalb hat das Ministerium für Bauen, Wohnen, Stadtentwicklung...



Digitale Lebenswelten in NRW: Wir sind dabei – und Sie?

Naturschutzinformationen NRW: OSIRIS

Ohne vertrauensvolle und konstruktive Zu-

Where to get Open Data

- https://www.europeandataportal.eu/
- https://www.govdata.de/
- https://open.nrw/
- https://offenedaten-koeln.de
- https://www.kaggle.com/datasets
- and many more...

Data formats

Binary

- not human readable
- memory efficient and fast to parse
- but: platform-dependent
- difficult to convert the format (e.g. open a Word Document in Open Office…)

Text

- (mostly) human readable
- waste more memory and relatively slow to parse
- platform-independent (but: still encoding problems!)
- format can be easily transformed

Common (text) data formats

- **CSV**: simplest format possible: Just strings separated by commas and newlines.
- **XML**: most common text data format. XHTML is the language of the web.
- **JSON**: uses javascript syntax. Much better readable and more sparse than XML.

CSV - Comma seperated values #1

```
4/5/2015 13:34,Apples,73
4/5/2015 23:41,Cherries,85
4/6/2015 12:46,Pears,14
4/8/2015 08:59,Oranges,52
```

- Column separator:
  - default is comma (“,”), but also tabulator (“\t”)
  - very common row separator: usually newline (“\n”)
- Strings can be enclosed in quotation marks to escape special characters
- Quotation marks are escaped by another quotation marks

CSV - Comma seperated values #2

- The advantage of CSV files is **simplicity**.
- CSV files are **widely supported** by many types of programs, can be viewed in text editors, and are a straightforward way to represent data.
- Whenever your data has **no nested structure**: **USE CSV**!
- Can be easily read in every programming language.
- Can be opened in text editors and in **Excel**!
- Comma is default, but tabs are often better as you rarely have to escape strings in tab separated files.

XML - Extensible Markup Language

```
<lectures>
  <lecture title="xml">
    <day>thursday</day>
    <slides>many</slides
  </lecture>
</lectures>
```

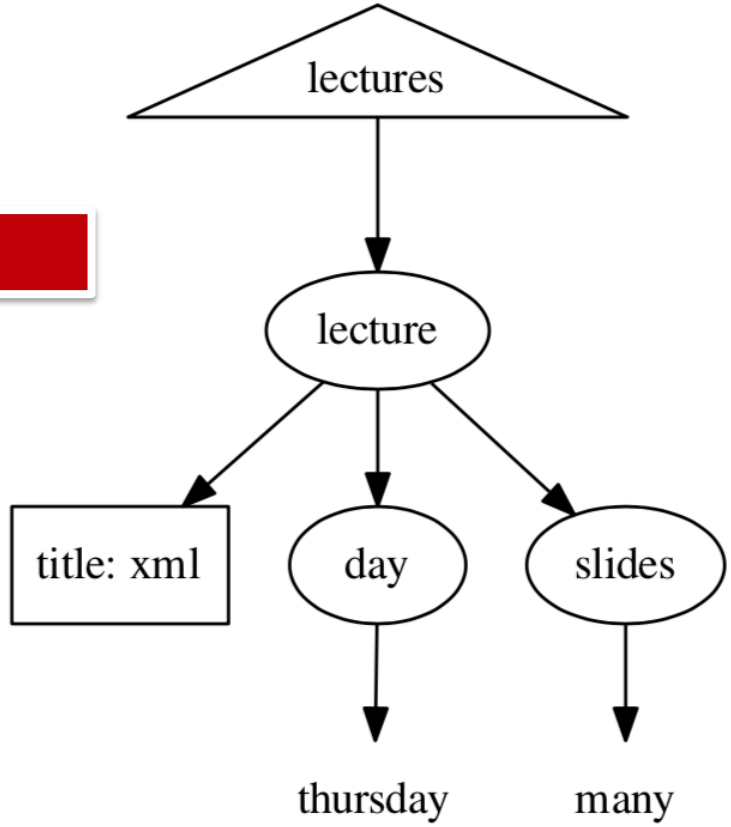
root

attribute

element

value

closing tag



- Very common data format
- Solves a lot of problems (namespaces, encoding, embedded data)
- Not very readable, very verbose

JSON - JavaScript Object Notation

```
[
  {
  "name": "Glucose",
  "formula": "C6H12O6",
  "similarTo": ["Hexose", "Fructose"]
   }
]
```

- Just Javascript data. Human readable but not easy to parse.
- Popular use: Client and web server often use JSON to communicate. Javascript can naturally work with json.
- Also almost compatible to Pythons syntax booleans, numbers, strings, arrays, objects (dictionaries).

Question time!

- How many people have used spreadsheets in their work?
- What kind of operations do you do in spreadsheets?
- Which ones do you think spreadsheets are good for?

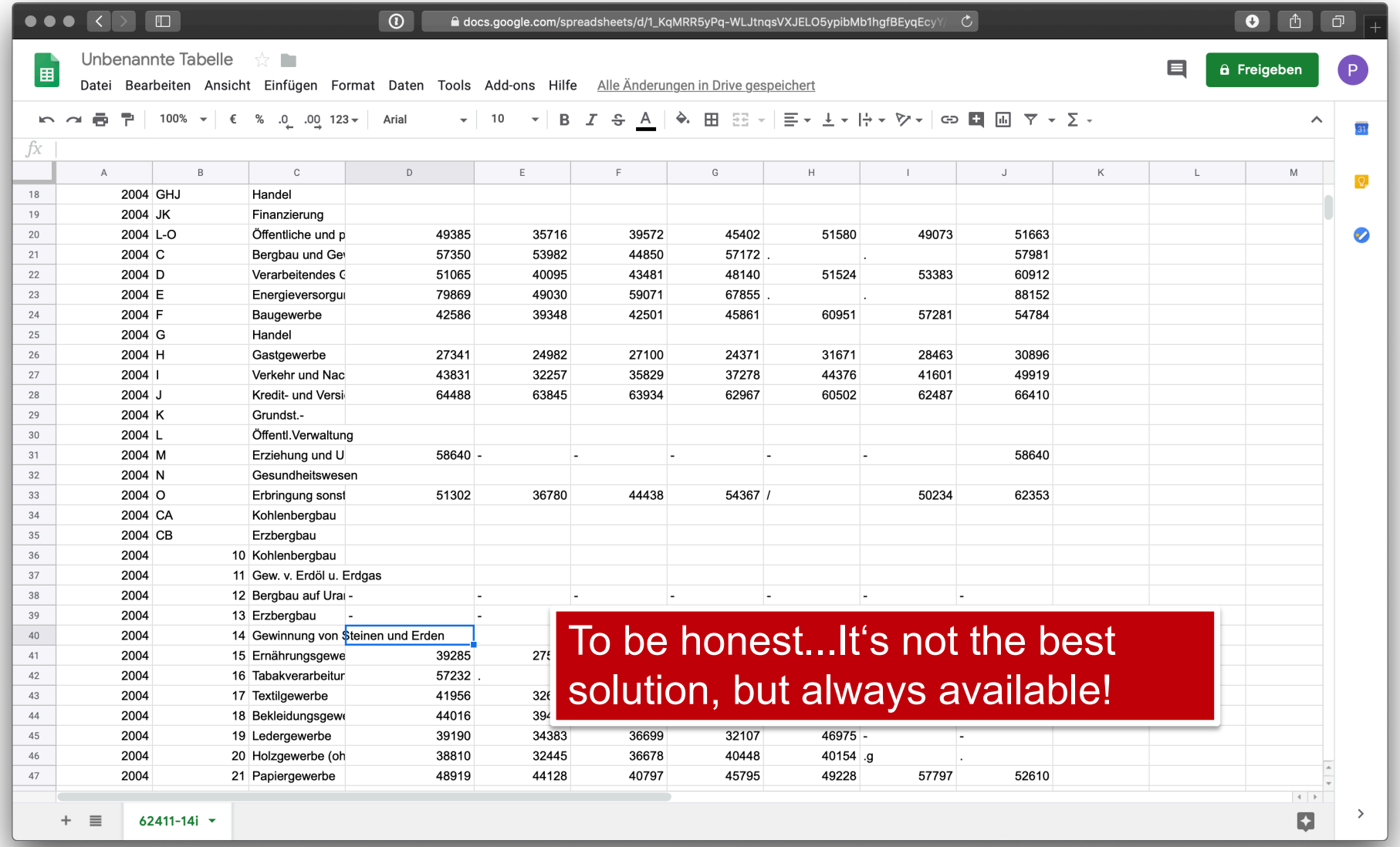
Why do we care about spreadsheets?

- Spreadsheets are **good for data entry**, but in reality we **tend to use spreadsheet programs for much more** than data entry. We use them to create data tables for publications, to generate summary statistics, and make figures.
- Generating **tables for reports** in a spreadsheet is not optimal. We advise you to do this sort of operation within your document editing software.
- **HOWEVER**, there are circumstances where you might want to use a spreadsheet program to produce “quick and dirty” calculations or figures, and some of these features can be used in **data cleaning**, prior to importation into a statistical analysis program. We will show you how to use some features of spreadsheet programs to check your **data quality** along the way and produce preliminary summary statistics.

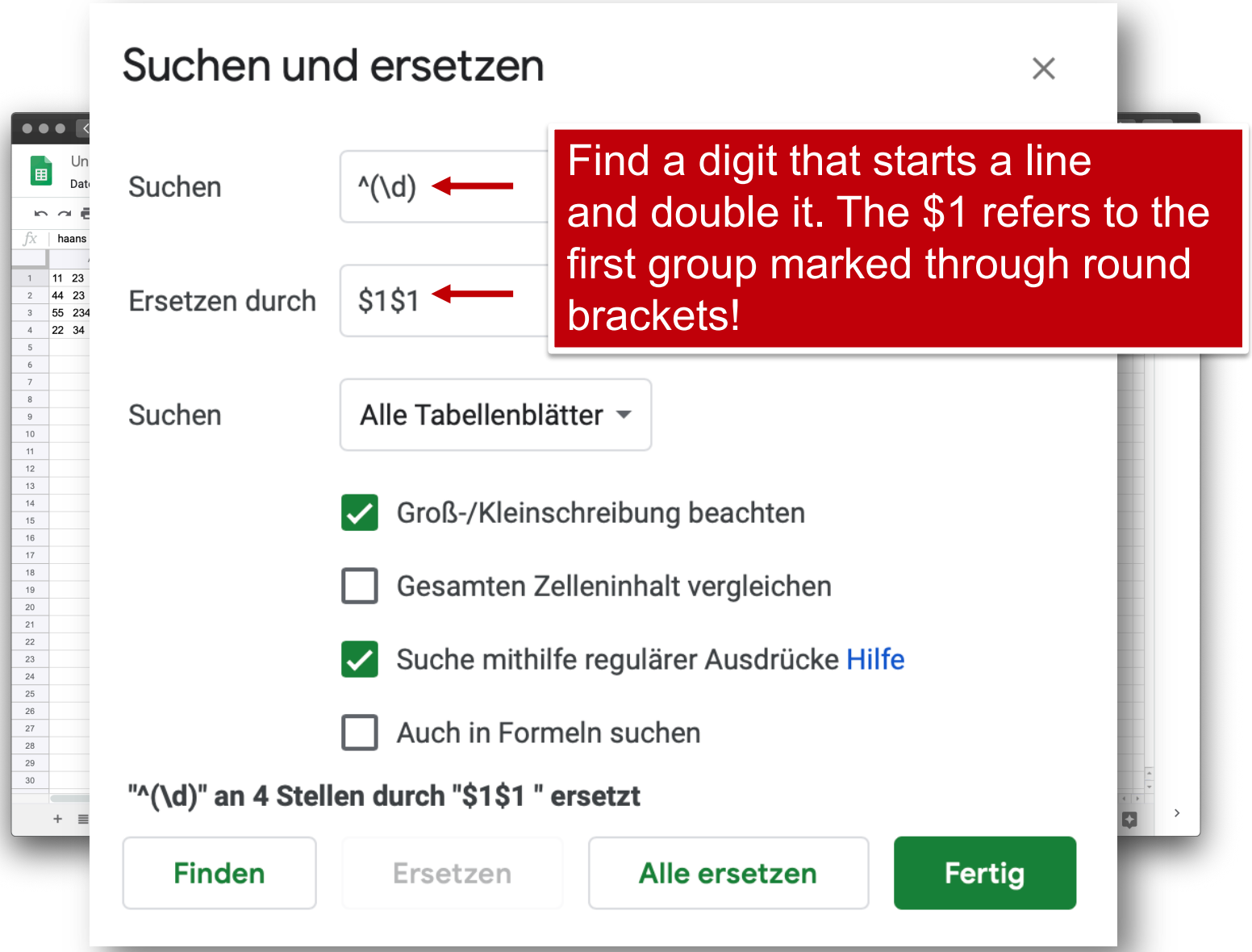
Spreadsheet Software

- Microsoft Excel
- Libre/Open Office
- Apple Numbers
- OnlyOffice (in Sciebo)
- **...**

- **Google Spreadsheets**
  - http://spreadsheets.google.com


docs.google.com/spreadsheets/d/1_KqMRR5yPq-WLJtnqsVXJELO5ypibMb1hgfBEyqEcyY/
Unbenannte Tabelle
Datei Bearbeiten Ansicht Einfügen Format Daten Tools Add-ons Hilfe
Alle Änderungen in Drive gespeichert
Freigeben
P
100% € % .0 .00 123 Arial 10
fx
A B C D E F G H I J K L M
18 2004 GHJ Handel
19 2004 JK Finanzierung
20 2004 L-O Öffentliche und p 49385 35716 39572 45402 51580 49073 51663
21 2004 C Bergbau und Ge 57350 53982 44850 57172 . . 57981
22 2004 D Verarbeitendes G 51065 40095 43481 48140 51524 53383 60912
23 2004 E Energieversorgu 79869 49030 59071 67855 . . 88152
24 2004 F Baugewerbe 42586 39348 42501 45861 60951 57281 54784
25 2004 G Handel
26 2004 H Gastgewerbe 27341 24982 27100 24371 31671 28463 30896
27 2004 I Verkehr und Nac 43831 32257 35829 37278 44376 41601 49919
28 2004 J Kredit- und Versi 64488 63845 63934 62967 60502 62487 66410
29 2004 K Grundst.-
30 2004 L Öffentl.Verwaltung
31 2004 M Erziehung und U 58640 - - - - - 58640
32 2004 N Gesundheitswesen
33 2004 O Erbringung sonst 51302 36780 44438 54367 / 50234 62353
34 2004 CA Kohlenbergbau
35 2004 CB Erzbergbau
36 2004 10 Kohlenbergbau
37 2004 11 Gew. v. Erdöl u. Erdgas
38 2004 12 Bergbau auf Urai - - - - - - -
39 2004 13 Erzbergbau - -
40 2004 14 Gewinnung von Steinen und Erden
41 2004 15 Ernährungsgewe 39285 275
42 2004 16 Tabakverarbeitur 57232 .
43 2004 17 Textilgewerbe 41956 326
44 2004 18 Bekleidungsgewe 44016 394
45 2004 19 Ledergewerbe 39190 34383 36699 32107 46975 - -
46 2004 20 Holzgewerbe (oh 38810 32445 36678 40448 40154 .g .
47 2004 21 Papiergewerbe 48919 44128 40797 45795 49228 57797 52610
62411-14i
To be honest...It‘s not the best solution, but always available!

Intermission! Using regex to change data



Structuring data in spreadsheets

The cardinal rules of using spreadsheet programs for data:

- Put all your **variables in columns** - the thing you're measuring, like 'weight' or 'temperature'.
- Put each **observation in its own row**.
- **Don't combine multiple pieces of information in one cell**. Sometimes it just seems like one thing, but think if that's the only way you'll want to be able to use or sort that data.
- **Leave the raw data raw** - don't mess with it!
- Export the cleaned data to a **text based format** like CSV. This ensures that anyone can use the data, and is the format required by most data repositories.

An example - Messy data?

- For instance, we have data from a **survey of small mammals** in a desert ecosystem. **Different people** have gone to the field and **entered data in to a spreadsheet**. They keep track of things like species, plot, weight, sex and date collected.

Date collected	Plot	Species-Sex	Weight
1/9/78	1	DM-M	40
1/9/78	1	DM-F	36
1/9/78	1	DS-F	135
1/20/78	1	DM-F	39
1/20/78	2	DM-M	43
1/20/78	2	DS-F	144
3/13/78	2	DM-F	51
3/13/78	2	DM-F	44
3/13/78	2	DS-F	146

An example - Messy data!

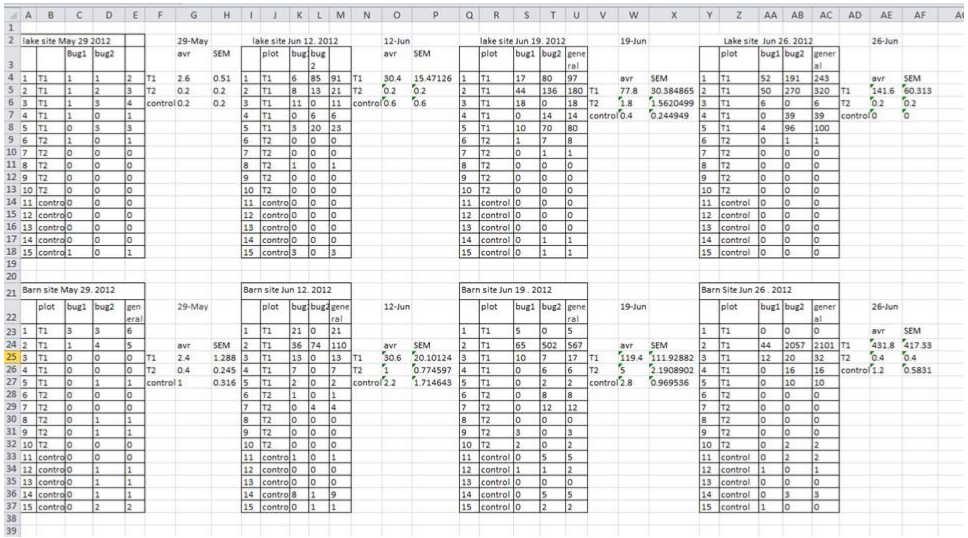
- The problem is that **species** and **sex** are in the same field.
- If you want to look at all of one species or look at different weight distributions by sex, it would be hard to do so.
- Put sex and species in **different columns**!

Date collected	Plot	Species	Sex	Weight
1/9/78	1	DM	M	40
1/9/78	1	DM	F	36
1/9/78	1	DS	F	135
1/20/78	1	DM	F	39
1/20/78	2	DM	M	43
1/20/78	2	DS	F	144
3/13/78	2	DM	F	51
3/13/78	2	DM	F	44
3/13/78	2	DS	F	146

Common Spreadsheet Errors #1

Multiple tables

- A common strategy is creating multiple data tables within one spreadsheet. **This confuses the computer!**
- When you create multiple tables within one spreadsheet, you're drawing false associations between things for the computer, which sees each row as an observation. You're also potentially using the same field name in multiple places!



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
1																																
2	lake site May 29 2012						29-May		lake site Jun 12. 2012						12-Jun			lake site Jun 19. 2012					19-Jun			Lake site Jun 26. 2012					26-Jun	
3			Bug1	bug2			avr	SEM		plot	bug1	bug 2			avr	SEM		plot	bug1	bug2	gene ral					plot	bug1	bug2	gener al			
4	1	T1	1	1	2	T1	2.6	0.51	1	T1	6	85	91	T1	30.4	15.47126	1	T1	17	80	97		avr	SEM	1	T1	52	191	243		avr	SEM
5	2	T1	1	2	3	T2	0.2	0.2	2	T1	8	13	21	T2	0.2	0.2	2	T1	44	136	180	T1	77.8	30.384865	2	T1	50	270	320	T1	141.6	60.313
6	3	T1	1	3	4	control	0.2	0.2	3	T1	11	0	11	control	0.6	0.6	3	T1	18	0	18	T2	1.8	1.5620499	3	T1	6	0	6	T2	0.2	0.2
7	4	T1	1	0	1				4	T1	0	6	6				4	T1	0	14	14	control	0.4	0.244949	4	T1	0	39	39	control	0	0
8	5	T1	0	3	3				5	T1	3	20	23				5	T1	10	70	80				5	T1	4	96	100			
9	6	T2	1	0	1				6	T2	0	0	0				6	T2	1	7	8				6	T2	0	1	1			
10	7	T2	0	0	0				7	T2	0	0	0				7	T2	0	1	1				7	T2	0	0	0			
11	8	T2	0	0	0				8	T2	1	0	1				8	T2	0	0	0				8	T2	0	0	0			
12	9	T2	0	0	0				9	T2	0	0	0				9	T2	0	0	0				9	T2	0	0	0			
13	10	T2	0	0	0				10	T2	0	0	0				10	T2	0	0	0				10	T2	0	0	0			
14	11	contro	0	0	0				11	contro	0	0	0				11	control	0	0	0				11	control	0	0	0			
15	12	contro	0	0	0				12	contro	0	0	0				12	control	0	0	0				12	control	0	0	0			
16	13	contro	0	0	0				13	contro	0	0	0				13	control	0	0	0				13	control	0	0	0			
17	14	contro	0	0	0				14	contro	0	0	0				14	control	0	1	1				14	control	0	0	0			
18	15	contro	1	0	1				15	contro	3	0	3				15	control	0	1	1				15	control	0	0	0			
19																																
20																																
21	Barn site May 29. 2012								Barn site Jun 12. 2012								Barn site Jun 19 . 2012								Barn Site Jun 26 . 2012							
22		plot	bug1	bug2	gen eral		29-May			plot	bug1	bug2	gene ral		12-Jun			plot	bug1	bug2	gene ral		19-Jun			plot	bug1	bug2	gener al		26-Jun	
23	1	T1	3	3	6				1	T1	21	0	21				1	T1	5	0	5				1	T1	0	0	0		avr	SEM
24	2	T1	1	4	5		avr	SEM	2	T1	36	74	110		avr	SEM	2	T1	65	502	567		avr	SEM	2	T1	44	2057	2101	T1	431.8	417.33
25	3	T1	0	0	0	T1	2.4	1.288	3	T1	13	0	13	T1	30.6	20.10124	3	T1	10	7	17	T1	119.4	111.92882	3	T1	12	20	32	T2	0.4	0.4
26	4	T1	0	0	0	T2	0.4	0.245	4	T1	7	0	7	T2	1	0.774597	4	T1	0	6	6	T2	5	2.1908902	4	T1	0	16	16	control	1.2	0.5831
27	5	T1	0	1	1	control	1	0.316	5	T1	2	0	2	control	2.2	1.714643	5	T1	0	2	2	control	2.8	0.969536	5	T1	0	10	10			
28	6	T2	0	0	0				6	T2	1	0	1				6	T2	0	8	8				6	T2	0	0	0			
29	7	T2	0	0	0				7	T2	0	4	4				7	T2	0	12	12				7	T2	0	0	0			
30	8	T2	0	1	1				8	T2	0	0	0				8	T2	0	0	0				8	T2	0	0	0			
31	9	T2	0	1	1				9	T2	0	0	0				9	T2	3	0	3				9	T2	0	0	0			
32	10	T2	0	0	0				10	T2	0	0	0				10	T2	2	0	2				10	T2	0	2	2			
33	11	contro	0	0	0				11	contro	1	0	1				11	control	0	5	5				11	control	0	2	2			
34	12	contro	0	1	1				12	contro	0	0	0				12	control	1	1	2				12	control	1	0	1			
35	13	contro	0	1	1				13	contro	0	0	0				13	control	0	0	0				13	control	0	0	0			
36	14	contro	0	1	1				14	contro	8	1	9				14	control	0	5	5				14	control	0	3	3			
37	15	contro	0	2	2				15	contro	0	1	1				15	control	0	2	2				15	control	1	0	0			
38																																
39																																

Common Spreadsheet Errors #2

Using formatting to convey information

Plot: 2			
Date collecte	Species	Sex	Weight
1/8/14	NA		
1/8/14	DM	M	44
1/8/14	DM	M	38
1/8/14	OL		
1/8/14	PE	M	22
1/8/14	DM	M	38
1/8/14	DM	M	48
1/8/14	DM	M	43
1/8/14	DM	F	35
1/8/14	DM	M	43
1/8/14	DM	F	37
1/8/14	PF	F	7
1/8/14	DM	M	45
1/8/14	OT		
1/8/14	DS	M	157
1/8/14	OX		
2/18/14	NA	M	218
2/18/14	PF	F	7
2/18/14	DM	M	52
	measurement device not calibrated		

Plot: 2				
Date collecte	Species	Sex	Weight	Calibrated
1/8/14	NA			
1/8/14	DM	M	44	Y
1/8/14	DM	M	38	Y
1/8/14	OL			
1/8/14	PE	M	22	Y
1/8/14	DM	M	38	Y
1/8/14	DM	M	48	Y
1/8/14	DM	M	43	Y
1/8/14	DM	F	35	Y
1/8/14	DM	M	43	Y
1/8/14	DM	F	37	Y
1/8/14	PF	F	7	Y
1/8/14	DM	M	45	Y
1/8/14	OT			
1/8/14	DS	M	157	N
1/8/14	OX			
2/18/14	NA	M	218	N
2/18/14	PF	F	7	Y
2/18/14	DM	M	52	Y

Common Spreadsheet Errors #3

Using formatting to make the data sheet look pretty

- **Example**: merging cells.
- **Solution**: If you're not careful, formatting a worksheet to be more aesthetically pleasing can compromise your computer's ability to see associations in the data. Merged cells are an absolute formatting NO-NO if you want to make your data readable by statistics software. Consider restructuring your data in such a way that you will not need to merge cells to organize your data.

Common Spreadsheet Errors #4

Field name problems

- Choose **descriptive field names**, but be careful not to include: spaces, numbers, or special characters of any kind.
- **Spaces** can be misinterpreted by parsers that use whitespace as delimiters and some programs don't like field names that are text strings that start with numbers.
- **Underscores** (_) are a good alternative to spaces and consider writing names in camel-case to improve readability.
- Remember that **abbreviations** that make sense at the moment may not be so obvious in 6 months but don't overdo it with names that are excessively long. Including the units in the field names avoids confusion and enables others to readily interpret your fields.

Common Spreadsheet Errors #4

Good Name	Good Alternative	Avoid
Max_temp_C	MaxTemp	Maximum Temp (°C)
Precipitation_mm	Precipitation	precmm
Mean_year_growth	MeanYearGrowth	Mean growth/year
sex_mf	sex	M/F
weight_kg	weight	w.
cell_type	CellType	Cell Type
Observation_01	first_observation	1st Obs

Common Spreadsheet Errors #5

Special characters in data

- **Example**: You treat Excel as a word processor, even copying data directly from Word or other applications.
- **Solution**: This is a **common strategy**. For example, when writing longer text in a cell, people often include line breaks, em-dashes, et al in their spreadsheet.
- Worse yet, when copying data in from applications such as Word, formatting and fancy non-standard characters are included.
- **General best practice is to avoid adding characters such as newlines, tabs, and vertical tabs**. In other words, treat a text cell as if it were a simple web form that can only contain text and spaces.

Common Spreadsheet Errors #6

Not filling in zeroes

- It might be that when you're measuring something, it's usually a zero, say the number of times an elephant is observed in the object or the survey. Why bother writing in the number zero in that column, when it's mostly zeros?
- There's a **difference between a zero and a blank cell** in a spreadsheet. To the computer, a zero is actually data. You measured or counted it. A blank cell means that it wasn't measured and the computer will interpret it as a null value.

Table 1. Commonly used null values, limitations, compatibility with common software and a recommendation regarding whether or not it is a good option. Null values are indicated as compatible with specific software if they work consistently and correctly with that software. For example, the null value "NULL" works correctly for certain applications in R, but does not work in others, so it is not presented in the table as R compatible.

Null values	Problems	Compatibility	Recommendation
0	Indistinguishable from a true zero		Never use
Blank	Hard to distinguish values that are missing from those overlooked on entry. Hard to distinguish blanks from spaces, which behave differently.	R, Python, SQL	Best option
-999, 999	Not recognized as null by many programs without user input. Can be inadvertently entered into calculations.		Avoid
NA, na	Can also be an abbreviation (e.g., North America), can cause problems with data type (turn a numerical column into a text column). NA is more commonly recognized than na.	R	Good option
N/A	An alternate form of NA, but often not compatible with software		Avoid
NULL	Can cause problems with data type	SQL	Good option
None	Uncommon. Can cause problems with data type	Python	Avoid
No data	Uncommon. Can cause problems with data type, contains a space		Avoid
Missing	Uncommon. Can cause problems with data type		Avoid
-,+,.	Uncommon. Can cause problems with data type		Avoid

Let‘s try it out!

- Download a free online open data set
  - https://open.nrw/dataset/ldbnrw-service-62411-14i
  - Choose the CSV format
- Import it into Google Spreadsheets
- Audit the data format and find common mistakes
- Correct the data format
- Export as CSV