

Open & Tidy Data

Programmierkurs 2 Data Science WS24/25

Leonard Traeger
M. Sc. Information Systems
leonard.traeger@fh-dortmund.de



Disclaimer

This lecture part is designed to give you a rough understanding of rule of thumbs in data modelling. If your group works with spreadsheets, these tips will help.

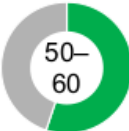
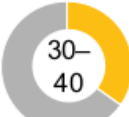
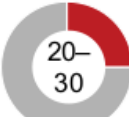

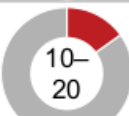
- This lesson is based on slides by Philipp Schaer, Technische Hochschule Köln, Cologne, Germany

Definition of „Openness“ <http://opendefinition.org/>

“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
- Open Access
- Open Science
- Open Source
- Open Government ...

Potential Impact of Open Data

	Potential impact: 2011 research	Value captured %	Major barriers
Location-based data	<ul style="list-style-type: none"> ▪ \$100 billion+ revenues for service providers ▪ Up to \$700 billion value to end users 		<ul style="list-style-type: none"> ▪ Penetration of GPS-enabled smartphones globally
US retail¹	<ul style="list-style-type: none"> ▪ 60%+ increase in net margin ▪ 0.5–1.0% annual productivity growth 		<ul style="list-style-type: none"> ▪ Lack of analytical talent ▪ Siloed data within companies
Manufacturing²	<ul style="list-style-type: none"> ▪ Up to 50% lower product development cost ▪ Up to 25% lower operating cost ▪ Up to 30% gross margin increase 		<ul style="list-style-type: none"> ▪ Siloed data in legacy IT systems ▪ Leadership skeptical of impact
EU public sector³	<ul style="list-style-type: none"> ▪ ~€250 billion value per year ▪ ~0.5% annual productivity growth 		<ul style="list-style-type: none"> ▪ Lack of analytical talent ▪ Siloed data within different agencies
US health care	<ul style="list-style-type: none"> ▪ \$300 billion value per year ▪ ~0.7% annual productivity growth 		<ul style="list-style-type: none"> ▪ 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

<https://obamawhitehouse.archives.gov/the-press-office/transparency-and-open-government>



"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)

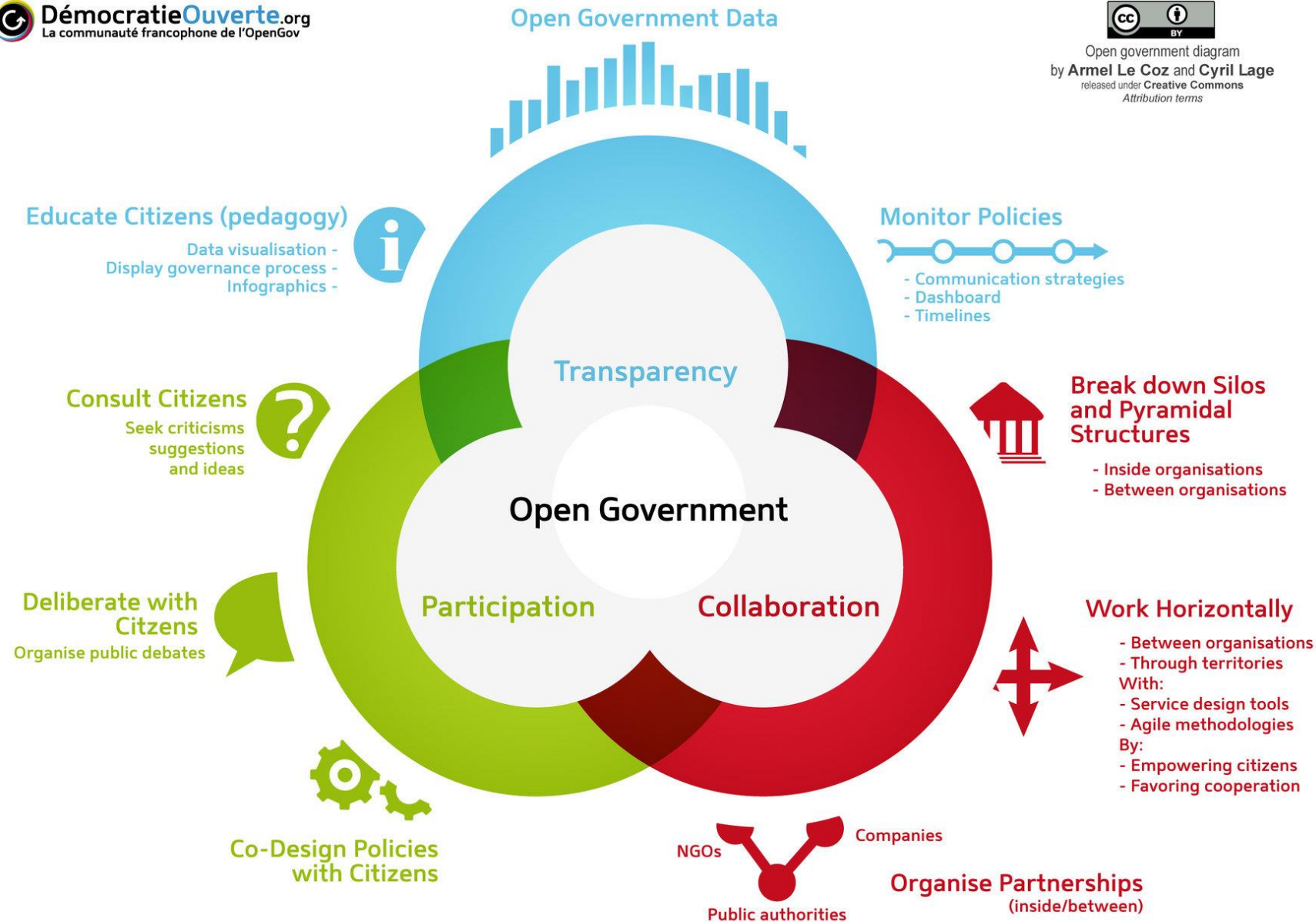
- "In the face of doubt, openness prevails."
- "Commitment to accountability and transparency"



Barack Obama, 21 January 2009 (first day in the office)

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, ...



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>

Trends and Scoring of Open Data in EU

<https://data.europa.eu/en/publications/open-data-maturity/2022>

National open data portals become more

- User-friendly
- Data quality improves
- Policies more sustainable
- Higher Impact

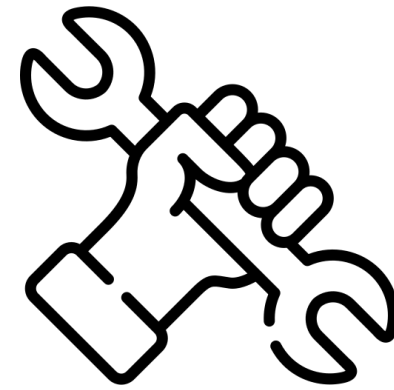
High-value datasets:

- COVID 19 vaccination (France)
- Waterlevels (Ireland)

Dimension	Key metrics	Scoring	Weight
Open Data Policy		648	25%
	Policy framework	275	
	Governance of open data	190	
	Open data implementation	175	
Open Data Impact		640	25%
	Strategic awareness	170	
	Measuring reuse	110	
	Created impact	320	
	Governmental impact	80	
	Social impact	80	
	Environmental impact	80	
	Economic impact	80	
Open Data Portal		650	25%
	Portal features	240	
	Portal usage	150	
	Data provision	110	
	Portal sustainability	150	
Open Data Quality		650	25%
	Currency and completeness	150	
	Monitoring and measures	150	
	DCAT-AP compliance	180	
	Deployment quality and linked data	170	
Total		2540	100%

Open Data Hubs

- Eine **Sammlung** von Datensätzen aus einer Vielzahl von Domänen, zu der über 100 Personen beigetragen haben:
<https://github.com/awesomedata/awesome-public-datasets>
- **UCI ML Repository**: <http://archive.ics.uci.edu/ml/>
- **Kaggle Datasets**: <https://www.kaggle.com/datasets>
- **Open Data Europa**: <https://data.europa.eu/en>
- **Datenportal für Deutschland**: <https://www.govdata.de/>
- **Landesdatenbank NRW**: <https://www.landesdatenbank.nrw.de/ldbnrw/online>
- **Open Data Dortmund**: <https://opendata.dortmund.de/Informationsportal/>



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:

- Create data tables for publications.
- Generate summary statistics.
- Make figures.

Problem: Generating tables for reports in a spreadsheet is not optimal.

Advice: Do this sort of operation within your document editing software.

Exception: Produce “quick and dirty” data cleaning, calculations or figures (prior to importation into a statistical analysis program)

Spreadsheet Software

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

To be honest...It's not the best solution,
but always available!

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 into 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	lake site May 29 2012																29-May		SEM		lake site Jun 12 2012						12-Jun		SEM		lake site Jun 19 2012						19-Jun		lake site Jun 26 2012						26-Jun	
2			bug1		bug2																																									
3	4	1	T1	1	1	1	2	T1	2.6	0.51	1	T1	6	18	91	T1	30.4	15.47126	1	T1	17	80	97			avr	SEM	8	T1	51	191	243		avr	SEM											
4	5	2	T1	1	2	3	T2	0.2	0.2	2	T1	8	15	21	T2	0.2	0.2	2	T1	44	156	180	T1	77.8	30.384865	2	T1	50	270	320	T1	541.6	50.313													
5	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	5.8	1.5620499	3	T1	6	0	6	T2	0.2	0.2													
6	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													
7	8	5	T1	0	3	3				5	T1	3	20	23				5	T1	10	70	80				5	T1	4	96	100																
8	9	6	T2	1	0	1				6	T2	0	0	0				6	T2	1	7	8				6	T2	0	1	1																
9	10	7	T2	0	0	0				7	T2	0	0	0				7	T2	0	1	1				7	T2	0	0	0																
10	11	8	T2	0	0	0				8	T2	1	0	1				8	T2	0	0	0				8	T2	0	0	0																
11	12	9	T2	0	0	0				9	T2	0	0	0				9	T2	0	0	0				9	T2	0	0	0																
12	13	10	T2	0	0	0				10	T2	0	0	0				10	T2	0	0	0				10	T2	0	0	0																
13	14	11	control	0	0	0				11	control	0	0	0				11	control	0	0	0				11	control	0	0	0																
14	15	12	control	0	0	0				12	control	0	0	0				12	control	0	0	0				12	control	0	0	0																
15	16	13	control	0	0	0				13	control	0	0	0				13	control	0	0	0				13	control	0	0	0																
16	17	14	control	0	0	0				14	control	0	0	0				14	control	0	1	1				14	control	0	0	0																
17	18	15	control	1	0	1				15	control	0	0	0				15	control	0	1	1				15	control	0	0	0																
18	Barn site May 29 2012																29-May		SEM		Barn site Jun 12 2012						12-Jun		SEM		Barn site Jun 19 2012						19-Jun		Barn site Jun 26 2012						26-Jun	
19			bug1		bug2																																									
20	21	1	T1	3	3	6				1	T1	21	0	21				1	T1	5	0	5				1	T1	0	0	0																
21	22	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	531.8	517.33										
22	23	3	T1	0	0	0	T1	2.4	1.288	3	T1	13	0	13	T1	50.6	50.10124	3	T1	10	7	17	T1	519.4	111.92882	3	T1	12	20	32	T2	0.4	0.4													
23	24	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													
24	25	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	0.999356	5	T1	0	10	10																
25	26	6	T2	0	0	0				6	T2	1	0	1				6	T2	0	8	8				6	T2	0	0	0																
26	27	7	T2	0	0	0				7	T2	0	4	4				7	T2	0	12	12				7	T2	0	0	0																
27	28	8	T2	0	1	1				8	T2	0	0	0				8	T2	0	0	0				8	T2	0	0	0																
28	29	9	T2	0	1	1				9	T2	0	0	0				9	T2	0	0	0				9	T2	0	0	0																
29	30	10	T2	0	0	0				10	T2	0	0	0				10	T2	0	2	2				10	T2	0	2	2																
30	31	11	control	0	0	0				11	control	1	0	1				11	control	0	5	5				11	control	0	2	2																
31	32	12	control	1	1	1				12	control	0	0	0				12	control	1	1	1				12	control	1	0	1																
32	33	13	control	1	1	1				13	control	0	0	0				13	control	0	0	0				13	control	0	0	0																
33	34	14	control	1	1	1				14	control	0	1	9				14	control	0	5	5				14	control	0	3	3																
34	35	15	control	2	2	2				15	control	0	1	1				15	control	0	2	2				15	control	1	0	0																
35																																														
36																																														
37																																														
38																																														
39																																														

Common Spreadsheet Errors #2

Using formatting to convey information

Plot: 2			
Date collect	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 collect	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**.
- 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.
- **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.

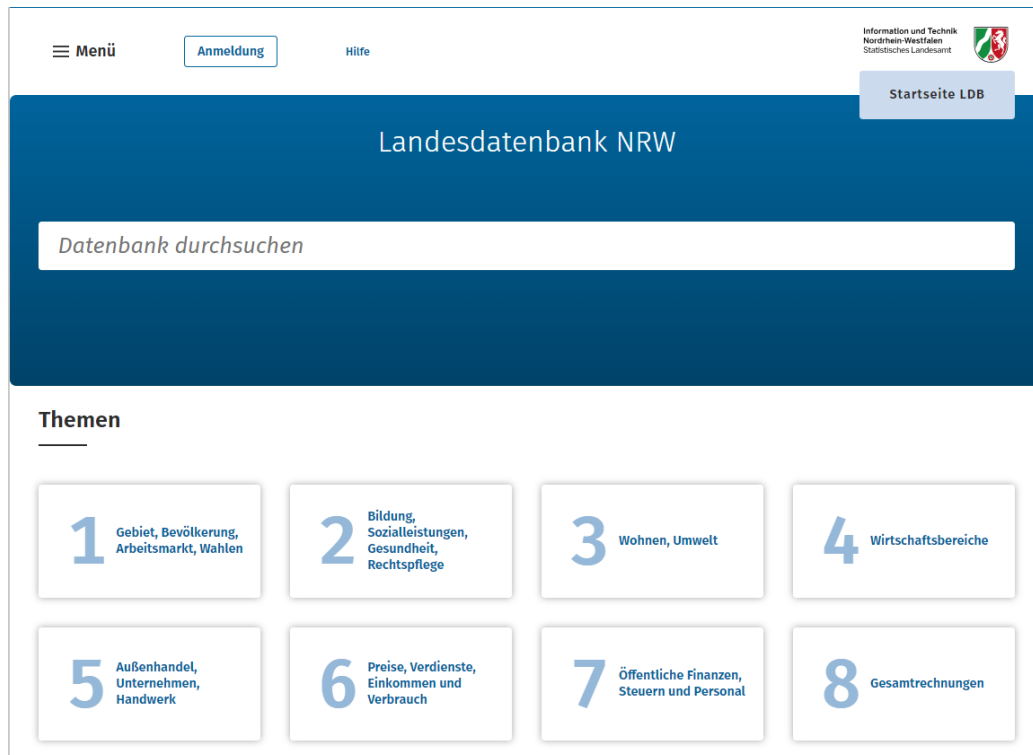
Common Spreadsheet Errors #6

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

White, E.P. et al. (2013) *Nine simple ways to make it easier to (re)use your data*, *Ideas in Ecology and Evolution*. Available at: <https://ojs.library.queensu.ca/index.php/IEE/article/view/4608> (Accessed: 19 September 2023).

Let's try it out!

<https://www.landesdatenbank.nrw.de/ldbnrw/online>



<https://opendata.dortmund.de/Informationsportal/>

