



DIS08 – Data Modeling

06 – Open data and the tidy data principle

Philipp Schaer, Technische Hochschule Köln, Cologne, Germany

Version: WS 2021

Disclaimer

This lesson is based on the Library Carpentry

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

Chapter 14 of Automating the Boring Stuff

Agenda

Last week

- Data formats: CSV, JSON, XML

This week

- Open Data and where to get it
- Working with CSV data
- Data cleaning in Excel in a nutshell

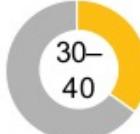
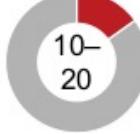
Next week

- Submit your first assignment!
- Take a breath! ☺

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 ...

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 	 50–60	<ul style="list-style-type: none"> Penetration of GPS-enabled smartphones globally
US retail ¹	<ul style="list-style-type: none"> 60%+ increase in net margin 0.5–1.0% annual productivity growth 	 30–40	<ul style="list-style-type: none"> Lack of analytical talent Siloed data within companies
Manufacturing ²	<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 	 20–30	<ul style="list-style-type: none"> Siloed data in legacy IT systems Leadership skeptical of impact
EU public sector ³	<ul style="list-style-type: none"> ~€250 billion value per year ~0.5% annual productivity growth 	 10–20	<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 	 10–20	<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.

Open Government

The screenshot shows the official website of the White House. At the top left, it says "the WHITE HOUSE PRESIDENT BARACK OBAMA". In the center is the official seal of the White House. To the right are links for "Get Email Updates" and "Contact Us". Below the main header is a navigation bar with links for "BLOG", "PHOTOS & VIDEO", "BRIEFING ROOM", "ISSUES", "the ADMINISTRATION", "the WHITE HOUSE", and "our 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"



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 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



Transparency

Consult Citizens

- Seek criticisms
- suggestions
- and ideas



Deliberate with Citizens

- Organise public debates



Co-Design Policies with Citizens



Open Government

Participation

Collaboration

Monitor Policies

- Communication strategies
- Dashboard
- Timelines



Break down Silos and Pyramidal Structures

- Inside organisations
- Between organisations

Work Horizontally

- Between organisations
- Through territories
- With:

 - Service design tools
 - Agile methodologies

- By:

 - Empowering citizens
 - Favoring cooperation



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.

What to do with open public data...

open.nrw/open-government/opennrw-projekte

The grid displays 12 project cards:

- Gemeinde Kranenburg: Weg zur Open Government-Gemeinde
- Kolpingstadt Kerpen: Digitalisierte Archivalien
- Gemeinde Stemwede: Open Government Angebote
- Stadt Paderborn: Transparenz und Partizipation
- Stadt Moers: Offene Daten in Schulen
- Südwestfalen IT Zahlen, Daten und Fakten
- kdvz Rhein-Erft-Rur: Transparenzportal für Gremiendaten
- Stadt Münster: Das Projekt „Lebenslage Familie“
- Stadt Köln: Civic Tech als neues Kollaborationsmodell
- Kreis Lippe: Wege zum Open Government

Aktuelle Daten und Zahlen zu Corona in NRW

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...



kaggle

Common (text) data formats

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

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>

Unbenannte Tabelle Freigeben P

Alle Änderungen in Drive gespeichert

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 C	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	326								
43	2004	17	Textilgewerbe	41956	326								
44	2004	18	Bekleidungsgew	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			

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

Intermission! Using regex to change data

The screenshot shows the 'Suchen und ersetzen' (Search and Replace) dialog in Google Sheets. On the left, there's a preview of a spreadsheet with data from row 1 to 4. The search term is '^(\d)' and the replace term is '\$1\$1'. A red callout box highlights these two entries with arrows pointing to them. Below the search term, there are dropdown menus for 'Suchen' (Search) set to 'Alle Tabellenblätter' (All spreadsheets) and 'Ersetzen durch' (Replace with). Under 'Ersetzen durch', there are several checkboxes: 'Groß-/Kleinschreibung beachten' (Check for case sensitivity) is checked, while 'Gesamten Zelleninhalt vergleichen' (Compare entire cell content), 'Suche mithilfe regulärer Ausdrücke Hilfe' (Help for regular expressions), and 'Auch in Formeln suchen' (Search in formulas) are unchecked. At the bottom, it says '"^(\d)" an 4 Stellen durch "\$1\$1 " ersetzt' (Replaced '^(\d)' at 4 places with "\$1\$1"). There are four buttons at the bottom: 'Finden' (Find), 'Ersetzen' (Replace), 'Alle ersetzen' (Replace All), and a large green 'Fertig' (Done) button.

Suchen und ersetzen

Suchen

Ersetzen durch

Suchen

Alle Tabellenblätter

Groß-/Kleinschreibung beachten

Gesamten Zelleninhalt vergleichen

Suche mithilfe regulärer Ausdrücke [Hilfe](#)

Auch in Formeln suchen

"`^\d`" an 4 Stellen durch "`$1$1`" ersetzt

Finden

Ersetzen

Alle ersetzen

Fertig

Find a digit that starts a line and double it. The \$1 refers to the first group marked through round brackets!

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!

Common Spreadsheet Errors #2

Using formatting to convey information

Plot: 2			
Date collected	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 collected	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

Enter: OpenRefine

Cleaning your data with spreadsheets is fine, but...

- hard to **reproduce** previous results
- **no log** file that tells us what was actually changed
- **keeping track** of all your changes is impossible
- adding **new data** let us start from the beginning. Again.
- ...

OpenRefine (previously Google Refine) is a powerful tool for working with **messy data!**

- **cleaning** it;
- **transforming** it from one format into another;
- and **extending** it with web services and external data.

OpenRefine 2014 01 31 JA america tsv Permalink

Facet / Filter Undo / Redo 0 / 0

Refresh Reset All Remove All

Facet by choice counts

27391 records

Show as: rows records Show: 5 10 25 50 100 500 1000 records « first < previous 1 of 2740 pages next » last »

Extensions: Wikidata ▾

Creator	Issue	Volume	Journal	ISSN	ID	Citation	Title	Place Label	Language	Publisher	Date
Herspring, D. R.	4	21	RUSSIAN HISTORY	0094-288X	(Uk)RN005361217	RUSSIAN HISTORY 21(4), 491-492. (1994)	"The Chinese Civil War, 1944-1946 Russia and America: From Rivalry to Reconciliation. Edited by George Ginsburgs, Alvin Z. Rubinstein, and Oles M. Smolansky"	xxu	eng	UNIVERSITY OF UTAH	1994
Helfman, B.	1	73	CALIFORNIA HISTORY	0162-2897	(Uk)RN009518447	CALIFORNIA HISTORY 73(1), 54-71. (1994)	"Sun Rising in an Eastern Sky": Japanese Americans in Washington Township, 1920-1942	xxu	eng	CALIFORNIA HISTORICAL SOCIETY	1994
Rothman, H. K.	1	9	MAGAZINE OF HISTORY	0882-228X	(Uk)RN010653563	MAGAZINE OF HISTORY 9(1), 28-33. (1994)	Environmental History in the American West	xxu	eng	ORGANIZATION OF AMERICAN HISTORIANS	1994
Bodnar, J.	1	15	JOURNAL OF AMERICAN ETHNIC HISTORY	0278-5927	(Uk)RN000189662	JOURNAL OF AMERICAN ETHNIC HISTORY 15(1), 3-27. (1995)	Remembering the Immigrant Experience in American Culture	xxu	eng	TRANSACTION PERIODICALS CONSORTIUM	1995
Erdman, H.	1	15	JOURNAL OF AMERICAN ETHNIC HISTORY	0278-5927	(Uk)RN000189674	JOURNAL OF AMERICAN ETHNIC HISTORY 15(1), 28-45. (1995)	M.B. Curtis and the Making of the American Stage Jew	xxu	eng	TRANSACTION PERIODICALS CONSORTIUM	1995
Barton, H. A.	1	15	JOURNAL	0278-	(Uk)RN000189686	JOURNAL	Where Have	xxu	eng	TRANSACTION	1995

Journal change
518 choices Sort by: name count Cluster
ANCIENT MESOAMERICA 9 ANGLOSAXON ENGLAND 1 edit include ANNALES DE BRETAGNE ET DES PAYS DE L OUEST 4 ANNALES DE L EST 1 ANNALES HISTORIQUES DE LA REVOLUTION FRANCAISE 3 ANNALS OF IOWA 66 ANTIPODE 40 ANTIQUITY -OXFORD- 53 APPLIED GEOGRAPHIC STUDIES 1 ADAP REVIEW LONDON

Creator change
17812 choices Sort by: name count Cluster
Abinales, P. 1 Abinales, P. N. 1 Abinales, P.N. 1 Abisaab, R. 1 Ablard, J. D. 1 Ablard, J.D. 1

Let's try it out!

- Download a free online open data set
 - <https://www.landesdatenbank.nrw.de/ldbnrw/online?operation=previous&levelindex=1&step=1&titel=Ergebnis&levelid=1637261702170>
 - Choose the CSV format
- Import it into **Google Spreadsheets** or **OpenRefine**
- Audit the data format and find common mistakes
- Correct the data format
- Export as CSV