

When you find data, it's often not ready for analysis yet. It needs to be tidied up first.

What is tidy data?

- machine readable
- · only relevant information
- one column = one feature

Common problems

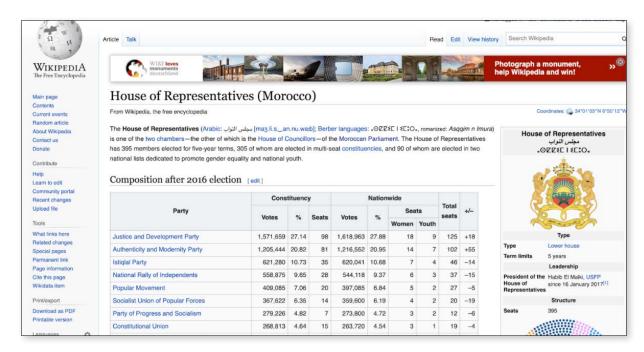
NOT MACHINE-READABLE

	2020 Q1		2020 Q2			
Reference area	Equivalent number of full-time jobs (40 hours/week) (millions)	Equivalent number of full-time jobs (48 hours/week) (millions)	Percentage hours lost (%)	Equivalent number of full-time jobs (40 hours/week) (millions)	Equivalent number of full-time jobs (48 hours/week) (millions)	Percentag hours lost (%)
World	185	155	5.4	480	400	14.0
Africa	11	9	2.4	55	45	12.1
Northern Africa	2	2	2.5	11	9	15.5
Sub-Saharan Africa	9	7	2.4	43	35	11.4
Central Africa	1	1	2.3	7	6	11.9
Eastern Africa	4	3	2.4	18	15	10.9
Southern Africa	0	0	1.6	3	2	12.2
Western Africa	3	3	2.5	15	13	11.6
Americas	13	11	3.0	80	70	18.3
Latin America and the Caribbean	10	9	3.6	55	47	20.0
Central America	1	1	1,1	16	13	19.2
South America	9	7	4.8	38	32	20.6
Northern America	3	2	1.8	25	21	15.3
Arab States	2	2	3.1	10	8	13.2
Asia and the Pacific	150	125	7.1	280	235	13.5
Eastern Asia	115	95	11.6	100	85	10.4
South-Fastern Asia and the Pacific	7	6	21	44	27	12.6

Example: PDFs, scanned Documents **Solution:** <u>Tabula</u>, OCR software



BURIED DATA



Example: Tables on websites that don't offer a download button

Solution: Scraping (s. Handout 2)

• A web scraping toolkit for journalists — Journocode

DIFFERENT SPELLINGS & TYPOS

gebort_original soumaine-Dages	staat_orie B
Casa Anfa	Marokko
Casablabca	Marokko
Casablanca	Marokko
Casablanca	Deutschland
Casablanca Anfa	Marokko
Casablanca/Hay Hassani	Marokko
Casbah	Marokko
Chefchaouen	Marokko

Example: Typing errors in city names

Solution: Change manually. For advanced users: Pattern recognition with

regular expressions or via Open Refine



INCORRECT ENCODING (A.K.A. FILE ORIGIN)

	Α	В	C		
1	Schulen - Sch	chulen - Schã¼ler - Klassen Schulen in			
2	Schulform	Anzahl Schulen			
3					
4					
5	Grundschulen		51		
6	städt.		49		
7	Gemeinschaft	29			
8	kath. Grundso	18			
9	ev. Grundsch	2			
10	nichtstädt.		2		
11					

Encoding determines how computers translate ones and zeros into characters

- different encodings are optimized for different languages
 - UTF-8: international standard, versatile
 - Win-1252 (a.k.a. Latin-1):
 Windows standard for Western
 European languages like German
 or French
- Win-1256: Windows standard for Arabic

If a file is opened with the wrong encoding, it might show the wrong symbols:

UTF-8	Win-1252	Win-1256	Mac Roman	Binary
<u></u>	а	а	а	01100001
ä	ä	ĺ¤	√§	11000011 10100100
é	é	ĺ©	é	11000011 10101001
گ	Ú-	_ع	/Ø	11011010 10101111

Example: "é" in UTF-8 becomes "é" or "[†]©" in different encodings

Solution: Look for strange symbols in the data, try out different encodings while

importing data

FAULTY CSV IMPORT

year,country_oridin_id,country_destination_id,sitc_pr
2013,GHA,BFA,0114,5696,0.0028%
2013,GHA,KEN,0114,3046,0.0015%
2013,GHA,TGO,0114,61769,0.031%
2013,GHA,ZAF,0114,479588,0.24%
2013 GHA ARE 0114 26950 0 013%

CSV: tabular data format with comma separated values:

- Rows are separated by line breaks
- Columns are separated by commas (or sometimes semicolons or tabs)

Example: ";" instead of "," as column separator or "," instead of "." as decimal separator

Solution: look at import settings while opening in Excel or LibreOffice

Tutorials:

- Import CSV files in Excel Copytrans.net
- Import CSV files in LibreOffice LibreOffice.org
- Import CSV files in Google Sheets Google.com



Workflow tips

DOCUMENTATION!

- Write down: What did you do? Why?
- Helps you and others retrace your steps

SPREADSHEET ORGANIZATION

- **Tipp:** Save your raw data, make a copy to work in!
- Move metadata to a new sheet: sources, date created, author, licenses etc.

TOOLS

Your best friends: spreadsheet applications like Excel, LibreOffice and Google Sheets!

More useful tools:

- Open Refine: Programm specifically for data cleaning, especially useful for messy text data
- Tabula: extract tables from PDFs
- <u>Table Capture:</u> extract tables from web pages (see handout 2)
- Regular Expressions: search and replace patterns in text

More resources

- <u>Top ten ways to clean your data</u> Microsoft.com
- Quartz/bad-data-guide: An exhaustive reference to problems seen in real-world data along with suggestions on how to resolve them. — Quartz
- How to prepare your data for analysis and charting in Excel & Google Sheets —
 Datawrapper.com

Help each other!

Many problems are easier to solve together. If you don't understand something, chances are others have the same question – or even an answer. Also, learning new skills in good company is simply a lot more fun.

