

Webinar 3: Data cleaning

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


► Table 1. Working-hour losses, world and by region and subregion, first and second quarters of 2020 (full-time equivalent jobs and percentage)

Reference area	2020 Q1			2020 Q2		
	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)	Percentage hours lost (%)
World	185	155	5.4	480	400	14.0
Africa	11	9	2.4	55	45	12.1
<i>Northern Africa</i>	2	2	2.5	11	9	15.5
<i>Sub-Saharan Africa</i>	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
<i>Latin America and the Caribbean</i>	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
<i>Northern America</i>	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
<i>Eastern Asia</i>	115	95	11.6	100	85	10.4
<i>South-Eastern Asia and the Pacific</i>	7	6	2.1	44	37	12.6

Example: PDFs, scanned Documents

Solution: [Tabula](#), OCR software

BURIED DATA



WIKIPEDIA
The Free Encyclopedia

Main page

Contents

Current events

Random article

About Wikipedia

Contact us

Donate

Contribute

Help

Learn to edit

Community portal

Recent changes

Upload file

Tools

What links here

Related changes

Special pages

Permanent link

Page information

Cite this page

Wikidata item


Print/export

Download as PDF






Printable version

ArticleTalk

ReadEditView historySearch Wikipedia



WIKI loves monuments deutschland



Photograph a monument, help Wikipedia and win!

House of Representatives (Morocco)

From Wikipedia, the free encyclopedia

Coordinates: 34°01′03″N 6°50′12″W﻿ / ﻿34.01750°N 6.83667°W﻿ / 34.01750; -6.83667

The **House of Representatives** (Arabic: مجلس النواب [maǧlis ʔn.nu.wab]; Berber languages: ⵏⴰⵎⴰⵔⵉⵏ ⵏ ⵉⵎⵓⵔⴰⵏ, romanized: *Asqim n Imura*) is one of the **two chambers**—the other of which is the **House of Councillors**—of the Moroccan Parliament. The House of Representatives has 395 members elected for five-year terms, 305 of whom are elected in multi-seat **constituencies**, and 90 of whom are elected in two national lists dedicated to promote gender equality and national youth.


Composition after 2016 election [edit]

Party	Constituency			Nationwide					Total seats	+/-		
	Votes	%	Seats	Votes	%	Seats						
						Women	Youth					
Justice and Development Party	1,571,659	27.14	98	1,618,963	27.88	18	9	125	+18			
Authenticity and Modernity Party	1,205,444	20.82	81	1,216,552	20.95	14	7	102	+55			
Istiqlal Party	621,280	10.73	35	620,041	10.68	7	4	46	−14			
National Rally of Independents	558,875	9.65	28	544,118	9.37	6	3	37	−15			
Popular Movement	409,085	7.06	20	397,085	6.84	5	2	27	−5			
Socialist Union of Popular Forces	367,622	6.35	14	359,600	6.19	4	2	20	−19			
Party of Progress and Socialism	279,226	4.82	7	273,800	4.72	3	2	12	−6			
Constitutional Union	268,813	4.64	15	263,720	4.54	3	1	19	−4			

House of Representatives

مجلس النواب

ⵏⴰⵎⴰⵔⵉⵏ ⵏ ⵉⵎⵓⵔⴰⵏ



Type

TypeLower house

Term limits

5 years

Leadership

President of the House of Representatives

Habib El Malki, USFP since 16 January 2017^[1]

Structure

Seats

395

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	staat orig	B
Boumaïne-Dades	Marokko	
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)

	A	B	C
1	Schulen - Sch	ler - Klassen	Schulen in
2	Schulform	Anzahl Schulen	
3			
4			
5	Grundschulen		51
6	st		49
7	Gemeinschaftsgrundschul		29
8	kath. Grundschulen		18
9	ev. Grundschulen		2
10	nichtst		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
a	a	a	a	01100001
ä	Ã	í	√	11000011 10100100
é	Ã©	í©	√©	11000011 10101001
گ	Ú	ع	Ø	11011010 10101111

Example: “é” in UTF-8 becomes “Ã©” or “í©” in different encodings

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

FAULTY CSV IMPORT

year	country_origin_id	country_destination_id	sitc_pr
2013	GHA	BFA	0.0028%
2013	GHA	KEN	0.0015%
2013	GHA	TGO	0.031%
2013	GHA	ZAF	0.24%
2013	GHA	ARE	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
- [Table Capture](#): extract tables from web pages (see handout 2)
- [Regular Expressions](#): search and replace patterns in text

More resources

- [Top ten ways to clean your data](#) — 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.