

DataScience for Development and Social Change, 2015

Cleaning Data part ii

Getting data into usable
formats

Data Cleaning

- ❖ The right format (e.g. CSV)
- ❖ The right shape (e.g. no 'extra' rows on top, one row per piece of data)
- ❖ Consistent labels
- ❖ Consistent “no data” values
- ❖ No junk symbols, whitespace etc.

Applying Standards

DR Congo in [data.UN.org](https://data.un.org): “Congo, Democratic Republic of the”, “Congo Democratic”, “Democratic Republic of the Congo”, “Congo (Democratic Republic of the)”, “Congo, Dem. Rep.”, “Congo Dem. Rep.”, “Congo, Democratic Republic of”, “Dem. Rep. of Congo”, “Dem. Rep. of the Congo”

DR Congo in common standards: “Democratic Republic of the Congo” (UN Stats), “Congo, The Democratic Republic of the” (ISO3166), “Congo, Democratic Republic of the” (FIPS10, Stanag), “180” (UN Stats), “COD” (ISO3166, Stanag), “CG” (FIPS10)

Even the standards don't agree. Your datasets won't either

Because

census	Arusha	Arusha	Daraja 2
shapefile	Arusha	Arusha Urban	Daraja Mbili
shapefile	Arusha	Ngorongoro	Endulen
census	Arusha	Ngorongoro	Enduleni
shapefile	Arusha	Ngorongoro	Engusero Sambu
census	Arusha	Ngorongoro	Enguserosambu
shapefile	Arusha	Longido	Gelai lumbwa
census	Arusha	Longido	Gelai Lumbwa
census	Arusha	Longido	Ketumbeine
shapefile	Arusha	Longido	Kitumbeine
census	Arusha	Arusha	Levolos
shapefile	Arusha	Arusha Urban	Levolosi

Happens almost every time you combine two datasets