GIS references in data.un.org

I've been playing with most of the data.un.org dataset: all 5577 csv files and 21195188 rows of it. And it's a fascinating dataset when you see it all together. I've already written about how to access the data.un.org datasets from an external application: now it's time to look at the headings and indices (first row and first column) in them all.

Details: I looked at 5577 csv files in the data.un.org dataset, and automatically excluded the footnotes at the end of each dataset. Data.un.org data that was excluded from the investigation were as follows: the UN interface limits downloads to 50000 rows of data, so 159 files in the set are incomplete; and 25 files were excluded because they're in a format (multi-sheet Excel files) that needs further work to separate comments from data. In all, there are 21195188 rows of data in the remaining dataset, so much of the following work had to be automated. Every cell in the first row of each of these files was added to file "headings.txt"•; every cell in the first column of each of these files (excluding the first row) was added to file "indices.txt"•.

Data.un.org files excluded (because they were in Excel format) were:

- Human Development Indices: A statistical update 2011
- Indicators on Women and Men
- OECD Data
- World Tourism Data
- World Fertility Data
- World Marriage Data
- World Contraceptive Use
- Key Indicators of the Labour Market, 7th Edition
- WTI Data

My goal here was to survey the types of headers used, so I could create lists of errors, inconsistencies etc. for known index types (e.g. country names), and start the work of cross-matching indices and headers to each other.

So. The results are: the indices contain these types:

- Most files: Country, region and economic group names, with a variety of spelling errors.
- International Monetary Fund files: a code called OID. The second column (country name) from these files has also been added to the list of indices.

Headings are more varied:

- Dates/times (Jan, Feb, year, number of years etc)
- Age-related (e.g. age, age groups)
- Geographical (e.g. country, OID, national station id, station name, residence area, WMO station number, city, country of birth, country of citizenship, country or territory, country or area, reference area, area etc)
- City type
- Birth/death details (cause of death, birth weight etc)
- Number of people (e.g. number of refugees, children etc)
- Personal (gender, sex, religion, marital status)
- Nationality/language (citizenship, ethnic group, native or foreign born, language)
- Household (household size, type of household etc)
- Financial value (USD, wealth quintile, currency, SNA)
- Employment/education (occupation, industry, education, literacy)
- Trade (Commodity/ISIC rev 3, flow=import or export, etc)
- General measure words (observation value, quantity, quantity name, weight, measure etc)
- Classification words (subgroup, item, type etc)
- Metadata (Description/source/series, survey coverage, activity status, record type, reliability, variant, footnotes “” these headings are used to refer to footnotes at the end of the file)

More detailed data should be attached to this post.