Indices and headings in data.un.org

'Data.un.org is the UN Statistics Division's Internet-accessible repository for data. It's potentially incredibly useful to the world, but is lagging behind sites like data.worldbank.org and data.gov because it doesn't have a machine API, it isn't normalized (the datasets aren't in a form that makes them easy to use with each other) and there are spelling errors in some of the headers and indices, notably in the names used for geographical locations (e.g. "USA"• and "United States"• are both used).

I've written already about how to access the data.un.org datasets from an external application and about the types of headers and indices within it. Now it's time to make the geographical references more usable.

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.

I collected the indices and headers from all the datasets into lists: the headers list was searched for geographical references, and the indices list was used to produce a list of corrections from the data.un.org geographical indices into both ISO standard 3166 (countrynames) and UNSTATS' list "Country and Region Codes for Statistical Use"• of region, country and economic group names on data.un.org.

Geographical references in the headers are:

- country of birth, country of citizenship, country or area, country or territory, country or territory of asylum or residence, country or territory of origin, reference area.
- OID.
- WMO station number, station name, national station id number.
- City.
- Area, residence area, city type.

The GIS naming standards found in the dataset were:

Country names: Most of the indices and country references are a close match to the UNSTATS standard or ISO3166 (as used by the World Bank etc), although country names in particular are very inconsistent in this dataset. The indices discussion below gives instructions on

- how to correct all of the country and region names to either of these standards. This applies

to headers country of birth, country of citizenship, country or area, country or territory, country or territory of asylum or residence, country or territory of origin and reference area. OID: International Monetary Fund's internal GIS standard.
- WMO station number, Station Name, National Station Id Number: World Meteorological Organisation references to meteorological stations.
- City: UNSD Demographic Statistics (code: POP) includes city names. No standard has yet been identified for these names.
- Area: UNSD Demographic Statistics (code: POP) and World Health Organisation (code: WHO) use the headings Area and Residence Area to classify the geographic extent of coverage (e.g. "Total"•, "Urban"•, "Rural"•). UNSD Demographic Statistics also uses the heading "City type"• to subclassify cities too (e.g. "City proper"•, "Urban agglomeration"•).

Most of the data.un.org datasets contain information that is listed by country (e.g. Yemen), region (e.g. West Africa) or economic group (e.g. Developing Regions). Looking at the placenames in the indices, we see that they are a mix of country, region and economic group names, with different spellings and formats for similar names (e.g. "Yemen"•, "YEMEN"•, "Yemen,Rep."•, "Yemen, Republic of"• etc).

Two standards are similar to the placenames used in these files: ISO3166 and the "composition of regions"• list published by data.un.org.

- ISO3166 is a widely-used standard, but contains code for countries and their subregions only (e.g. has no official lists of larger regions or economic areas) and is published as tables online and available (although without the list of withdrawn codes) in the Python library pycountry.
- The Unstats list (which ISO3166 is partially based on) contains countries, regions and economic areas, but is available only as an html table at http://unstats.un.org/unsd/methods/m49/m49regin.htm. The countries list is available as a ScraperWiki dataset which needs some editing to make it usable. The regions list has been scraped by hand for now. There are two main lists in it: the regions, subregions and countries by physical location, and the economic status (e.g. "Developing regions"•, "Least developed countries"•) of each country and region. These are mostly consistent, with a couple of oddities. For instance, Netherland Antilles doesn't appear on the list of countries, but does appear on the list of small island developing states. Luckily it has a country code, so it's been included in the list of countries.

The indices were checked against both these standards. Suggested improvements to the indices and standards include:

- Make the regions list available as a csv file online, to include withdrawn country codes, assignment dates and withdrawal dates (these are needed to match names for earlier

years).
- Make the economic status list available as a csv file online.
- Lobby ISO to create a region (Africa, West Africa, North America etc.) code standard, if it doesn't already exist.
- Lobby ISO to correct inconsistencies in the ISO countries list (e.g. republic not Republic in Bolivia's name).
- Make a definitive statement about which GIS naming standard (ISO, UNstats etc) UN online data should attempt to adhere to.
- Change all the data.un.org datafiles to meet this standard.

Against the ISO3166 standard, the data.un.org csv index errors were:

- Withdrawn countries with no ISO3166 code: "East Timor\", \"Czechoslovakia, Czechoslovak Socialist Republic"•, \"USSR, Union of Soviet Socialist Republics\", "Yemen, Yemen Arab Republic\", "Yemen, Democratic, People's Democratic Republic of\", "Yugoslavia, Socialist Federal Republic of"•, "Germany, Federal Republic of"•, "German Democratic Republic"•, "US Miscellaneous Pacific Islands\", "Wake Island\", "Serbia and Montenegro\".
- Abbreviation, e.g. "Rep."• for "Republic"•, "St."• for "Saint"•, "Is."• For "Island"•, "Isds"• for "Islands"•, "&"• for "and"•.
- Added markers, e.g. "+"• added to the end of region names, to differentiate them from countrynames.
- Capitalisation, e.g. "YEMEN"• for "Yemen"•, "republic"• for "Republic"•, "The"• for "the"• and the"• for "The"•.
- Brackets: UNICEF in particular uses brackets "()"• instead of commas in placenames
- Standards confusion: the ISO3166 labels "name"• and "official_name"• were both used in the same datasets ("name"• is available for all countries; "official_name"• is not).

Tables of other misspellings against both standards are given below. Some of these errors are the use of familiar names (e.g. Brunei, Ivory Coast, China) or issues with character translation (e.g. Cote d'Ivoire). Some names could not be resolved: remaining queries include the code for French Polynesia, whether "Christmas Is.(Aust)"• is Christmas Island, whether St. Helena refers to just the island of Saint Helena, or "Saint Helena, Ascension and Tristan da Cunha"• and whether Palestine and Palestinian Territories refer to "Palestinian Territory, Occupied"•