# Outline of a Uniform coding system for Canada industrial data

Canada has a dual coding system. The LFS reports on four-digit NAICS codes, but the GDP and other datasets use six-digit ‘IOICC’ codes. These are close to NAICS codes but

1. IOICC codes are not reported at every level. Thus GDP for 4131,4132,4133 are not reported, but for 4130 (the sum of these), it is.
2. They contain ‘residuals’ - amalgams of several NAICS codes. For example, the iOICC code 31A000 combines NAICS 3130 and 3140.
3. They contain ‘superfluous aggregates’ like 333A00 which gives the sum of 3332 and 3333, although these are also reported individually. This can give rise to double counting.

We will create single 6-digit ‘Pseudo-NAICS’ code NAICS (‘PNAICS’) code. All primary and all foreign keys will use the ANAICS code.

The ANAICS code will be the primary key of the industry dimension table.

It will have seven characters. The first will be a letter code indicating the data source to which it applies (L for LFS, P for P&H, G for GDP, etc). The following six digits will be the NAICS code, padded out with zeroes to the six digit level as appropriate

The PNAICS entry in the industry table will be the lowest atomic level found in the data data. Thus G111000 (Crop production, a 3-digit code) is the lowest level at which GDP for Crop production is reported and so G110000 will not be included in the PNAICS table, nor will G111100 or G111200. On the other hand G211110 (oil and gas production except oil sands) is a 5-digit code, lower than the 4-digit level. Therefore it appear in the industry table but G221100 will not.

The result is a ‘ragged hierarchy’ – not all items exist at the same level. It will however be a genuine hierarchy with no double counting.

The industry dimension will include, for each PNAICS key, a column for every level from 2 to 5. If the data is not reported at any lower level, the entry will be blank. Thus in the row whose primary key is ‘A111000’ the two-digit entry will be 11, the three-digit level will be 111 and the four and five-digit levels will be blank.

This thus creates a ‘ragged hierarchy’ in that whilst the sum of all leaves connected to the same higher node will add up to the value of the node, not all leaves are at the same level of the hierarchy.

NAICS-coded data (LFS) will be pre-processed: the code will be padded to 6 characters but trailing zeroes will be replaced by ‘N’ to prevent double-counting