# Information Visualization

# CHECKPOINT II: Data cleaning and processing

G01-A

**1. Initial Dataset**

We have a table (All Winners) containing all the editions of the games from 1896 until 2008, and for each there is the city and year they were at, and a set of podium finishes, which include the sport, discipline and event, and the athlete's name and country code (NOC, same as IOC and 3-letter ISO) and the medal they won. We have another table (Total) with the each country’s name, NOC, total medals and how many of gold, silver and bronze.

We have a table (Codes) with each country name, IOC (same as NOC) code and their ISO code.

We have another table (Population) with each country’s name, country code (3-letter ISO), indicator name and code, and a field of their population for each year from 1960 until 2014.

We have a table (Coordinates) to know the coordinates of each country, which has its 2-letter ISO code, latitude and longitude, and its name.

**2. Selected/Derived Data**

* + - 1. All Winners – Edition year, Sport, Medal won and the NOC of the country of the medallist.  
         Total – NOC, Total of medals.  
         Codes – Country, IOC, ISO code.  
         Population – ISO Code, Years from 1960 until 2008.  
         Coordinates – ISO Code, Latitude, Longitude.  
         Derived measure (*(medals won)/(population)* coefficient) – We want to compare the medals per capita over the years, so we will count the number of medals each country won in each year and divide it by its population in that year.
      2. **3. Data abstraction**

All Winners – A tree containing all the podium finishes of the countries since 1896 until 2008. It's organized first by edition which is the year of the games (*continuous* *sequential*); then by sport which can be “swimming”, “athletics”, etc. (*nominal*); and finally, the medals won which can be represented as “Bronze”, “Silver” or “Gold” (*ordinal*) and the **NOC** (*nominal*) of the medallist.  
Total – A simple table with the **NOC**, which is a 3-letter code representing a country (*nominal*); and an integer which is the total count of medals from all the editions of the games (*ratio*).  
Codes – A simple table with sets of three strings (*all* *nominal*): the country name, the **IOC** (country code, equal to NOC) and a **2-letter ISO** country code.  
Population – A table with a **3-letter ISO** country code matching the **IOC** (*nominal*), and a set of columns each pertaining to every fourth year between 1960 and 2008, containing the population of the country (*ratio*) in that year.  
Coordinates – A table with a **2-letter ISO** country code for each country (*nominal*), and a latitude and longitude for that country (*continuous*).

**4. Dataset processing**

* + - 1. To create the Codes table, we had to make sure the IOC codes matched the NOC on All Winners and Totals, and that the 2 and 3 letter ISO codes matched the same on other tables. Some values didn’t exist, because they were for older countries or united teams, so we checked the most representative countries related to those and made the association.  
         To get the *(medals won)/(population)* coefficient, we counted the amount of medals for each country (Pentaho’s *Group By* on All Winners) and divided it by the population of that country in that year (using Pentaho’s *Merge Join* between the two tables, then *Calculation*).
      2. **5. Mapping (Data sample / Questions)**

Some examples that show that with your data sample you will be able to provide the answers to the questions you formulated.