# Multi-Source Open Data Quality

Dr. Wassim Derguech – Derilinx.com

**Due Date:** 14/02/2018
**Lab dates:** Wednesday 31/01/2018, Wednesday 7/02/2018,

**Assignment**:
The objective of the project is to create a complete and clean dataset of playing pitches around Dublin region. The resulting dataset should be a csv file containing as much informative columns as possible. This is an individual assignment.

**Input datasets**:
Dataset are available on https://data.gov.ie

- [DCC] Playing pitches in Dublin City Council: https://data.gov.ie/dataset/dublin-city-council-parks-playing-pitches
- [DLR] Playing pitches in DLR: https://data.gov.ie/dataset/dlr_pitches
- [F] Playing Pitches in Fingal: https://data.gov.ie/dataset/playing-pitches

**Data format:**
Each of these datasets exhibits resources using various format: csv, xml, kml, html, etc. You should use at least 2 different formats (you will need to manage 3 files with at least 2 different formats).

**Marking scheme:**

- Report: 50%
- Code/Analytics: 50%

**Deliverables required:**

- Source code of your solution should be either in Python or Java with detailed comments to facilitate its readability (Using Jupiter notebook is highly recommended).
- A report outlining your decisions for the different challenges of this project
- The resulting dataset
- Optionally you can share your code on github.

**Methodology:**
You will have to create the resulting dataset by merging three different resources by resolving the following the three-step pipeline introduced in the lecture:

- **Step1: Observation.** Perform an observatory analysis of the datasets and define a data management plan for creating the resulting dataset.
- **Step 2:** D**ata modelling**: Create a unified model that should include the resulting fields of each record in your dataset. The minimal fields required are: location, and geographical coordinates (use x and y).
- **Step 3: Data Quality Enhancement:** You will be faced with two main challenges:
    - **Data cleaning challenge**: Merge the input resource and make sure that there are no duplicates, Locations can be created by merging various fields, etc.

- **Incomplete data challenge:** The first dataset [DCC] does not have any geographical coordinates. A solution to complete this information is required.

**Data observation and modelling: (20 points)**
The selected datasets do not share the same data model. Attribute names are different, some attributes or values are missing, etc.

You have to define a data model that describes best the resulting dataset without losing any information. That is, if one of dataset has a field that describes the type of playing pitch, it should be part of the resulting dataset.

A key requirement at this step is to make sure that at least three fields are considered: location, x and y (x and y are the geographical coordinates of the playing pitch).

In your report, indicate what issues where encountered: missing field, different field names, missing values, etc. and indicate what strategy can you use to resolve the issue.

**Data Quality Enhancement:**

- **Data cleaning challenge: (40 points)**

While merging the datasets, you will notice data issues that you need to resolve. For example, in the second dataset [DLR], you will find that the field "location" is empty, it should take the values of the previous line.

Analyse the datasets, identify data issues and propose a solution for each of them.

- **Incomplete data challenge: (40 points)**

The first dataset [DCC] does not contain any geographical coordinates. To complete the dataset, you will need to complete the dataset by the solution that you think is most suitable (google geocoding service, geocoder python package index, etc.).

You are free to use the method that you like with convincing reasons included in your report.

A suggested method is the following: Use another dataset that helps identify the geographical coordinates as follows:

- [OSiNPG] Dataset: https://data.gov.ie/dataset/townlands-osi-national-placenames-gazetteer2fe62
- Find out the best fields from the first dataset [DCC] that help you find a matching entry in [OSiNPG]. Fields can be the name of the park or the club.

Please note that there is not one correct answer, but there are many **convincing** answers. It is critical to justify the approaches used in the report.

**Submission Instructions**

- Please put your code into a single .zip archive with name "YourName_CaseStudyAssignment1_code.zip", submit via Blackboard
- Include a screenshot of the output of your application for any relevant output.

- Include all source code files (that is, files with name ending .java etc) required to compile and run your code.
- Use comments to explain your source code. Insufficient comments can lead to mark deductions.
- Please put your report into a single .pdf file with name "YourName_CaseStudyAssignment1_report.pdf", submit via Blackboard
- Please note that all submissions (both code and report) will be checked for plagiarism.