# DESCRIPCIÓN DE LA BASE DE DATOS DE PROYECTOS DE CORDIS

[DESCRIPCIÓN DE LA BASE DE DATOS DE PROYECTOS DE CORDIS 1](#_Toc526342223)

[1. Objetivo del documento 1](#_Toc526342224)

[2. Description of auxiliary tables 2](#_Toc526342225)

[2.1 Table countries 2](#_Toc526342226)

## 1. Version control

Current versión: 1.0

Date: Oct 4, 2018

Author: Jerónimo Arenas García

## 2. Document objective

El presente documento describe la base de datos MySQL que contiene el corpus de Proyectos de CORDIS. La documentación oficial y los datos se pueden obtener del portal de datos abiertos

https://data.europa.eu/

De entre todos los datasets disponibles se han incorporado a la base de datos los siguientes datasets:

* CORDIS reference data

<https://data.europa.eu/euodp/en/data/dataset/cordisref-data>

Contiene tablas con posibles valores de campos de uso común en el resto de las bases de datos:

* + listados de países con su acrónimo y nombre en diferentes idiomas
  + funding schemes/types of action for projects
  + Activity types of organizations
  + SIC codes (subjects of research)
  + programmes (FP6, FP7 and H2020)
  + topics (H2020 only)
* CORDIS - EU research projects under Horizon 2020 (2014-2020)

<https://data.europa.eu/euodp/en/data/dataset/cordisH2020projects>

It contains project data, organization data, as well as information of participating individuals (Main researchers) in certain projects

* CORDIS - EU research projects under FP7 (2007-2013)

<https://data.europa.eu/euodp/en/data/dataset/cordisfp7projects>

It contains project data and organization data

* CORDIS - EU research projects under FP6 (2002-2006)

<https://data.europa.eu/euodp/en/data/dataset/cordisfp6projects>

It contains project data and organization data

* CORDIS - EU research projects under FP5 (1998-2002)

<https://data.europa.eu/euodp/en/data/dataset/cordisfp5projects>

It contains project data and organization data

* CORDIS - EU research projects under FP4 (1994-1998)

<https://data.europa.eu/euodp/en/data/dataset/cordisfp4projects>

It contains project data and organization data

* CORDIS - EU funded projects under FP3 (1990–1994)

<https://data.europa.eu/euodp/en/data/dataset/cordis-fp3>

It contains project data and organization data

* CORDIS - EU funded projects under FP2 (1987–1991)

<https://data.europa.eu/euodp/en/data/dataset/cordis-fp2>

It contains project data and organization data

* CORDIS - EU funded projects under FP1 (1984–1987)

<https://data.europa.eu/euodp/en/data/dataset/fp1-cordis>

It contains project data and organization data

Other datasets that could be of interest are

* CORDIS - ESA FP6,FP7, projects

<https://data.europa.eu/euodp/en/data/dataset/cordis-esa-fp6-fp7-projects>

* CORDIS - FP7 projects with participants from Africa

<https://data.europa.eu/euodp/en/data/dataset/cordis-fp7-projects-with-participants-from-africa>

However, we have checked that these projects are already included in the corresponding datasets for FP6 and FP7.

Data importing has been implemented in Python using the project <https://github.com/jeroarenas/PTL_data>. The code imports the Excel files, since these are the ones that are available for all previous datases. A certain structure is also assumed for organization of the datafiles.

El acceso a la base de datos en los servidores del departamento puede hacerse usando los siguientes datos:

dbUSER = 'PTLprojects'

dbPASS = 'Kts93\_u17a'

dbNAME = 'db\_Pr\_CORDIS'

dbSERVER = 'vanir.tsc.uc3m.es'

## 3. Database structure

The database contains the following tables:

* **projects**: 99.266 rows corresponding to the same number of projects. The distribution of projects according to the 8 datasets is given in the table below.



* **countries**: EU code assigned to each country. There is a total of 251 countries, but the number of rows in the table is larger given that country names are provided in several languages
* **orgactivities**: codes for the different activity codes for organizations (5 rows)
* **fundingschemes:** comprises all 187 different funding schemes for the different calls in the different programmes.

## 4. Description of ‘projects’ table

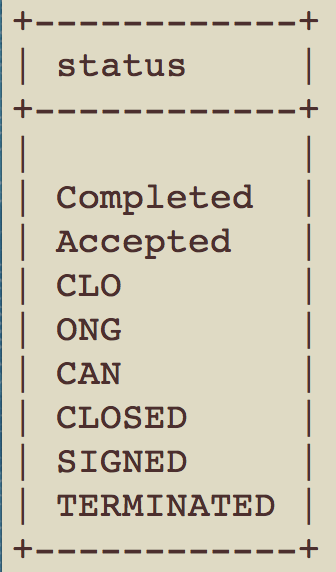
The following table denotes the presence of fields in the available Excel files. Highlighted fields in green are present as a header in the corresponding Excel file. However, it should be noted that sometimes these green fields have a lot of missing values. In some cases, it may even be the case that a green field is missing in the entire framework programme.



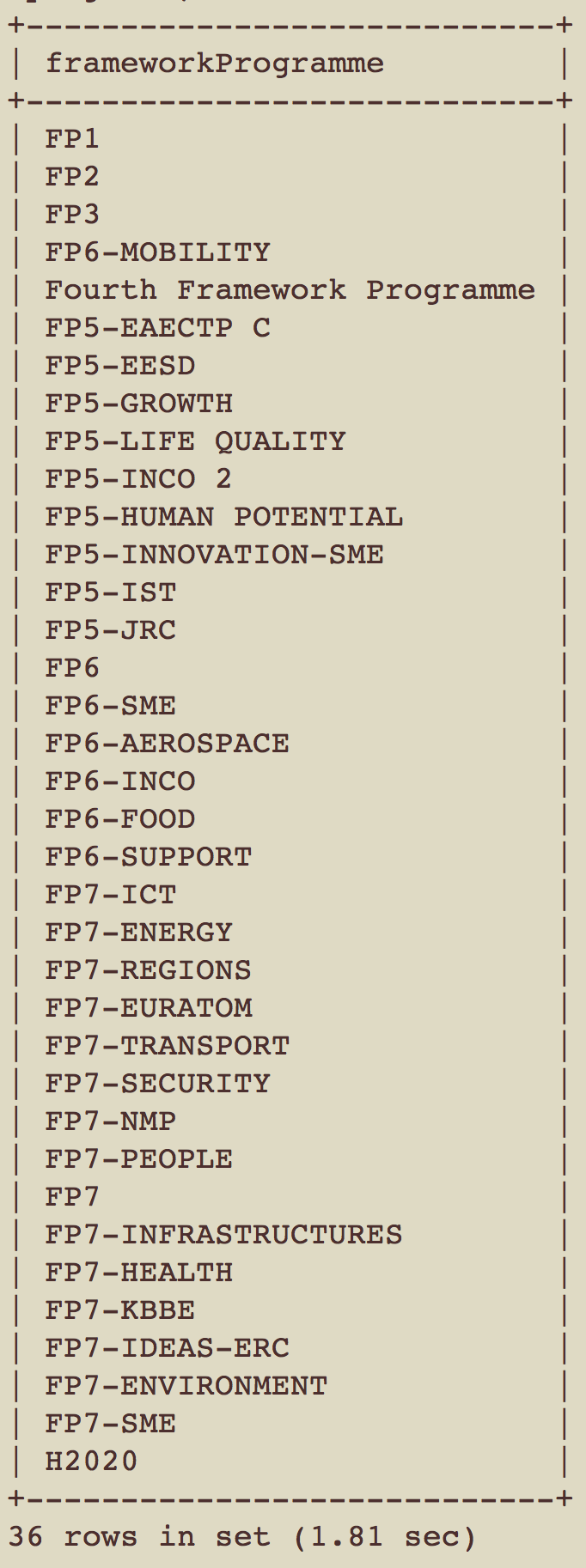
Fields highlighted in blue contain textual information that can be used for training embeddings, topic modelling, etc. However, fields ‘keywords’ and ‘GeneralInfo’ are present just in FP5 and are therefore not included. It can be seen that the field names are quite coherent across files, except for FP5, where some fields are omitted altogether, and other fields have a different name (indicated in the previous figure).

Next, we outline the different columns available in this table:

* frame\_programme: An identifier for the frame\_programme to which each project is associated (H2020/FP7/FP6/FP5/FP4/FP3/FP2/FP1)
* rcn: The unique identifier for each project, and primary key for the table
* reference: A second identifier for the project, not necessarily unique for the dataset.
* acronym: Missing for 24.199 projects
* status. Not available for 39.956 projects. When available, it takes one of the following values:



* programme / topics / call : Identifiers of the programme, topic and call associated to a project. This is only widely available for FP7 and H2020, but even in this case, I doubt these fields can be of direct applications given the large number of values they take. Maybe they can still be exploited by pruning the values in some way (Joseba input here)
* frameworkProgramme: This field provides some detailed information for FP5, FP6 and FP7 that could be useful. The values taken by this field are in the following list:



* title: Title of the project (no missing values)
* startDate: datetime object with the starting date of the project (missing for 2.209 projects)
* endDate: datetime object with the ending date of the project (missing for 3.042 projects)
* projectURL: website of the project. Only available for 9.277 projects.
* Objective: Abstract of the project (missing for 15.234 projects)
* totalCost: Missing for 27.609 projects
* ecMaxContribution: Missing for 22.933 projects
* fundingScheme: Missing for 5.019 projects. Currently this field takes 163 different values, but most likely the real number of missing values is around 100. This fields needs consolidation!
* coordinator: Name of the organization or person coordinating the project
* coordinatorCountry: Acronym (2 letters) of the country coordinating the project. This field is missing in 858 projects
* participants
* participantCountries: List of acronyms of the participating countries. This field is missing for 44.518 projects, so it is much less reliable than the coordination data. It would be interesting to find out whether projects with missing information in this field are individual projects in some cases (e.g. ERC starting or advance grants).
* subjects:
* LEMAS\_UC3M\_ENG: English lemmas obtained using the lemmatization tool provided by UPM. These lemmas are extracted keeping their order of appearance in the original text, in case they are used to train a Word Embedding model. We also keep certain structure to mark sentence limits:

The format of this field is as follows:

LEMMATIZED TITLE + ‘\*\*\*\*\*’ + LEMMATIZED OBJECTIVE

In each of the fragments, the separation between sentences is denoted by the new line character ‘\n’.

N-grams can also be easily identified because they include character ‘\_’ to separate the elements in the n-gram.

## 7. Description of auxiliary tables

### 7.1. Table ‘countries’

It contains the acronyms of the country names used in the projects and organization tables. It contains more than one entry per country, since full name (column ‘name’) is available in multiple languages (column ‘language’). A total of 251 countries are included.

### 7.2. Table ‘orgactivities’

It contains 5 different types of organizations according to their activity. This value is then used in the organizations table (in column ‘activityType’).

### 7.3. Table ‘fundingschemes’

It contains the different types of Funding Schemes to which a project can belong to. The codes in this table are therefore used to fill in the field ‘fundingScheme’ of the projects table. This table contains codes for 187 different fundingSchemes (e.g. ERC-AG -> ERC Advanced Grant)