# DESCRIPCIÓN DE LA BASE DE DATOS DE PROYECTOS DE CORDIS

[DESCRIPCIÓN DE LA BASE DE DATOS DE PROYECTOS DE CORDIS 1](#_Toc526629848)

[1. Version control 1](#_Toc526629849)

[2. Document objective 1](#_Toc526629850)

[3. Database structure 3](#_Toc526629851)

[4. Project Information in CORDIS dataset 4](#_Toc526629852)

[4.1. Description of ‘projects’ table 4](#_Toc526629853)

[4. Description of auxiliary tables for the project data 7](#_Toc526629854)

[4.1. Table ‘countries’ 7](#_Toc526629855)

[4.2. Table ‘fundingschemes’ 7](#_Toc526629856)

[4.3. Table ‘programmes’ 7](#_Toc526629857)

[4.4. Table ‘topics2020’ 8](#_Toc526629858)

[4.5. Table ‘sicCodes’ 8](#_Toc526629859)

[5. Organization Information in CORDIS dataset 8](#_Toc526629860)

[5.1. Available organization information 8](#_Toc526629861)

[5.2. Disambiguation of organizations 9](#_Toc526629862)

[5.3. Data summary 11](#_Toc526629863)

[5.4. Auxiliary Table ‘orgactivities’ 11](#_Toc526629864)

[6. Researcher Information in CORDIS dataset 12](#_Toc526629865)

## 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.



* **organizations**: 118.581 organizations that have participated in at least some project in the database
* **researchers**: 22.213 researchers that have participated in at least some project in the H2020 program (CORDIS does not currently provide researchers data for older projects).
* **orgzproject**: pairs of organizations and projects
* **rschprojects**: pairs of researchers and projects
* **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.
* **programmes**: data on different funding programmes from FP6, FP7 and H2020
* **topics2020**: data on topics for H2020.

A summary of the different tables and the fields in each of them is provided in the EER diagram below.



## 4. Project Information in CORDIS dataset

### 4.1. Description of ‘projects’ table

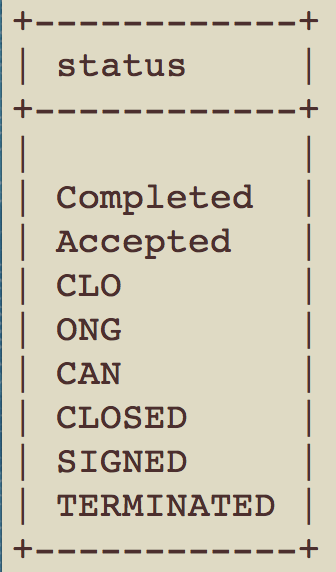
The following table denotes the presence of fields in the available Excel files. Highlighted fields in green are present as a header in the corresponding Excel file. However, it should be noted that sometimes these green fields have a lot of missing values. In some cases, it may even be the case that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

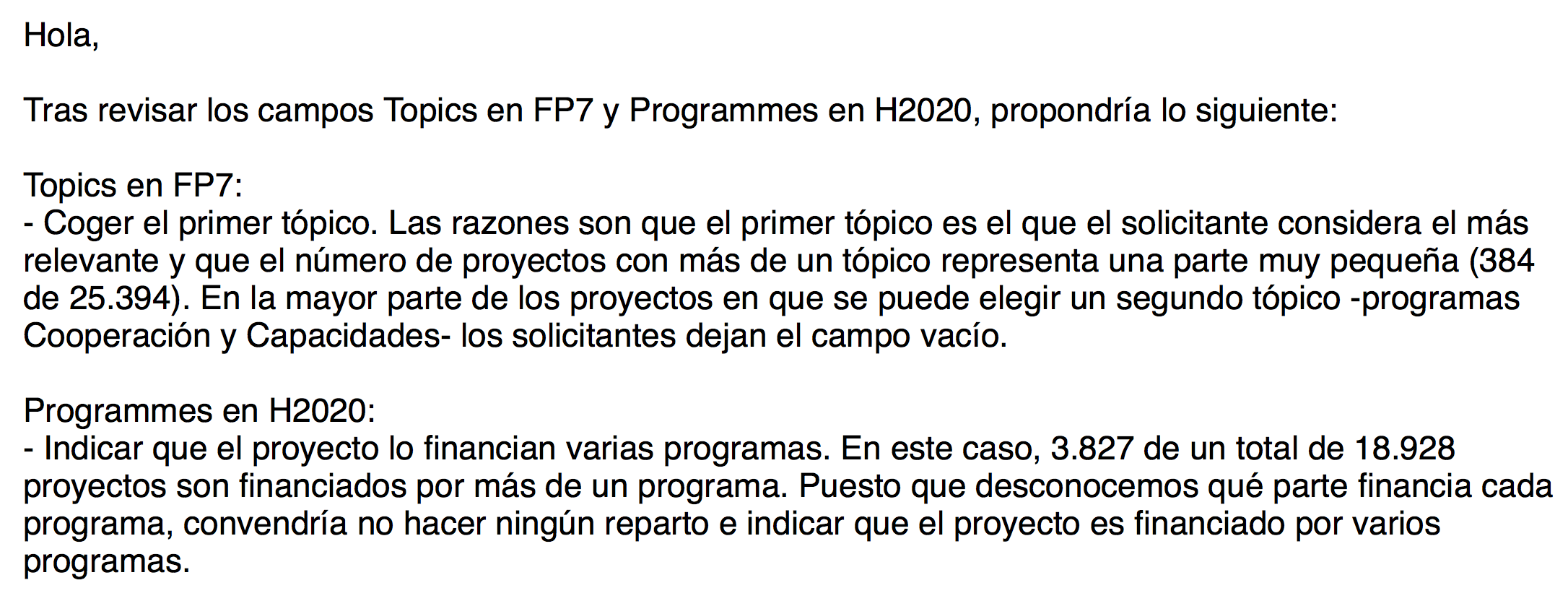
Next, we outline the different columns available in this table:

* 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. This a numeric field.
* 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:

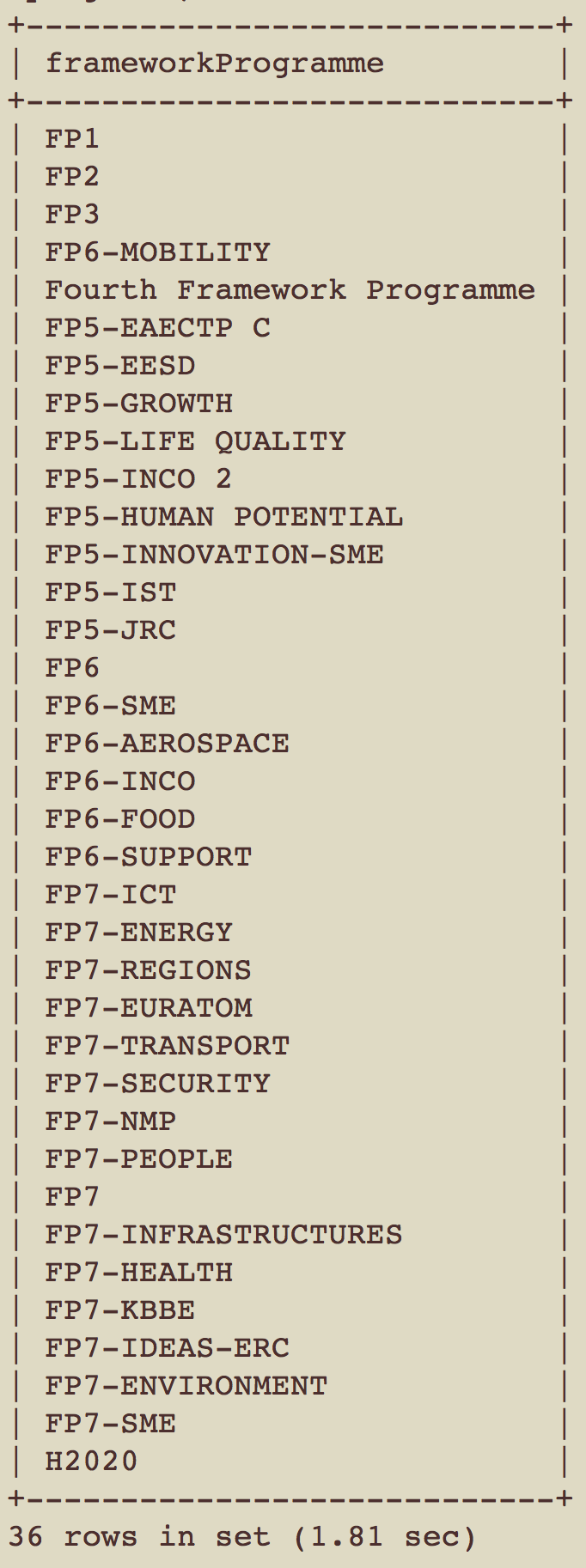


* programme / topics / call : Identifiers of the programme, topic and call associated to a project. This is only widely available for FP7 and H2020, but even in this case, I doubt these fields can be of direct applications given the large number of values they take. Maybe they can still be exploited by pruning the values in some way.

Joseba feedback:



* 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: Subject identification codes for the project. Several codes can be present, in which case they are separated by colons. In fact, sometimes codes can be repeated. It would be interesting to study the original XML files to get an explanation for this.
* 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.

### 4. Description of auxiliary tables for the project data

#### 4.1. 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.

#### 4.2. 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)

#### 4.3. Table ‘programmes’

It contains information on the different programs funding the calls of FP6, FP7 and H2020. The codes and a program rcn (Record Cordis Number) are available for every program, and titles and shorttitles (in several languages) are only available for some of them.

#### 4.4. Table ‘topics2020’

It contains information on the different topics for H2020. The codes and a topic rcn (Record Cordis Number) are available for every program, and titles (in english) are only available for some of them (2090 topics with titles given only for 1133 topics). The table has been named explicitly after H2020 to emphasize that it only contains topics for the H2020 framework programme, even though there are other topics in the dataset for projects belonging to other programmes.

#### 4.5. Table ‘sicCodes’

It contains information on the different SICs (subject identification codes). 71 different SICs are included in the table, with titles and descriptions available in different languages.

## 5. Organization Information in CORDIS dataset

### 5.1. Available organization information

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 file for a particular framework programme.



In general, it can be seen that the information seems more complete in the file for FP7 and H2020. In the table above some fields have been highlighted in yellow. This is so because the information in the organization files provided by CORDIS have been splited in two separate tables:

* organizations: contains an entry per organization, and it includes all available information pertaining to the organization itself but not to a particular call or project. Therefore, it includes all available information in the fields “id”, “name”, “shortName”, “activityType”, “country”, “street”, “city”, “postCode” and “vatNumber”.
* orgzproject: It contains information involving a particular organization and project. As so, it contains the role of the organization, ecContribution, etc …

### 5.2. Disambiguation of organizations

It should be noted that the original files contained an entry per pair “organization + table”, with a total number of rows of 522.341, and 127.184 unique names for the organizations.

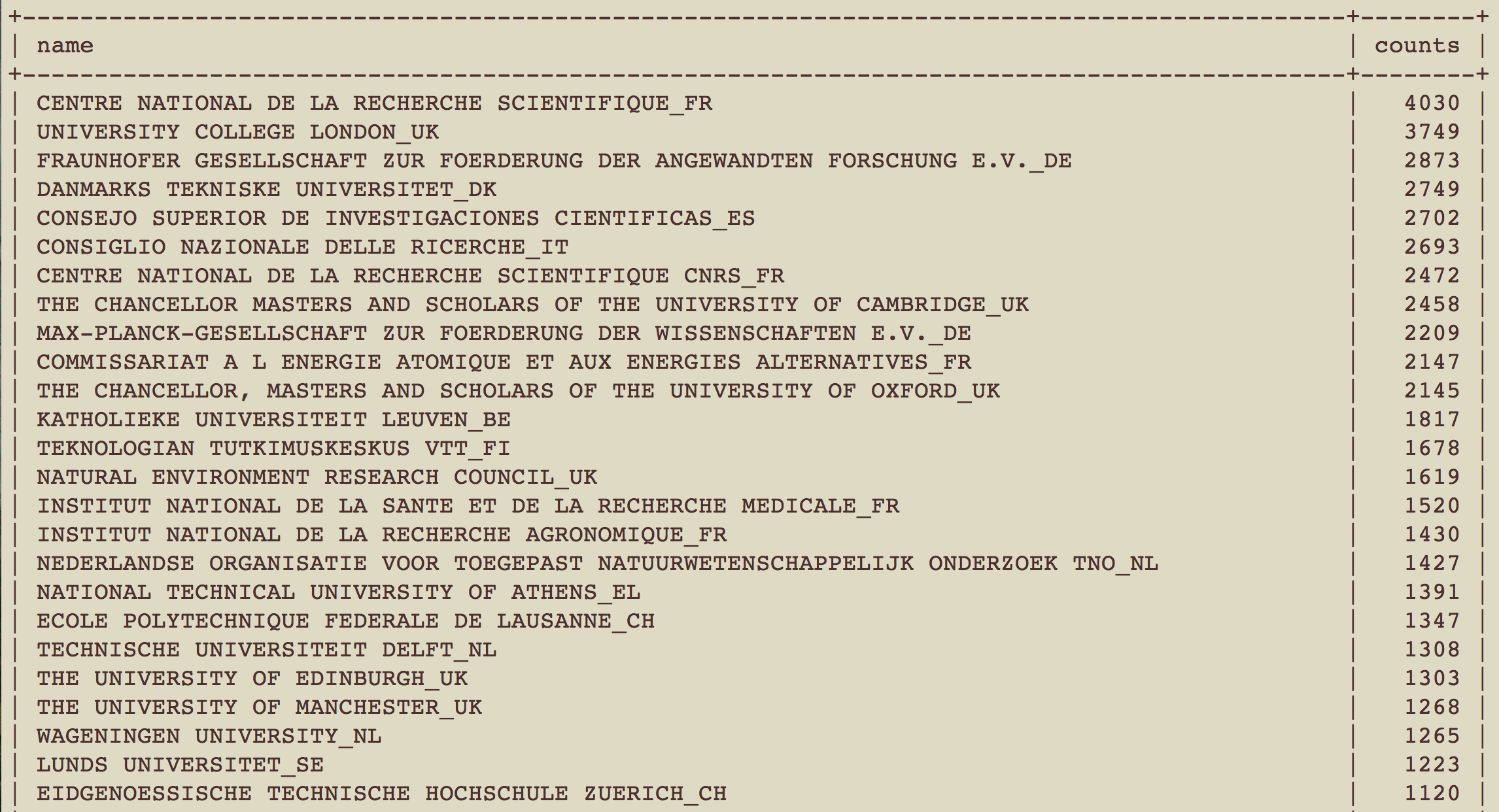
CORDIS does not provide any field to uniquely identify an organization. It could be claimed that fields “id” and “vatNumber” could be used for that purpose, but these fields are missing in most of the cases. The name of the organization could be used for that, but then again some organizations usually appear in the database with different names.

A very simple disambiguation has been carried out, but being very conservative in the mechanisms applied. In the present dataset we are using the concatenation of the name of the organization and the country as the identifier, and have collapsed names that were different under a common vatNumber, “id”, or “organization Url”. As a result, the number of organizations has gone down from 127.184 to 118.581, but it seems that there are still a very large number of duplicates, so a large effort of consolidation would be advisable here.

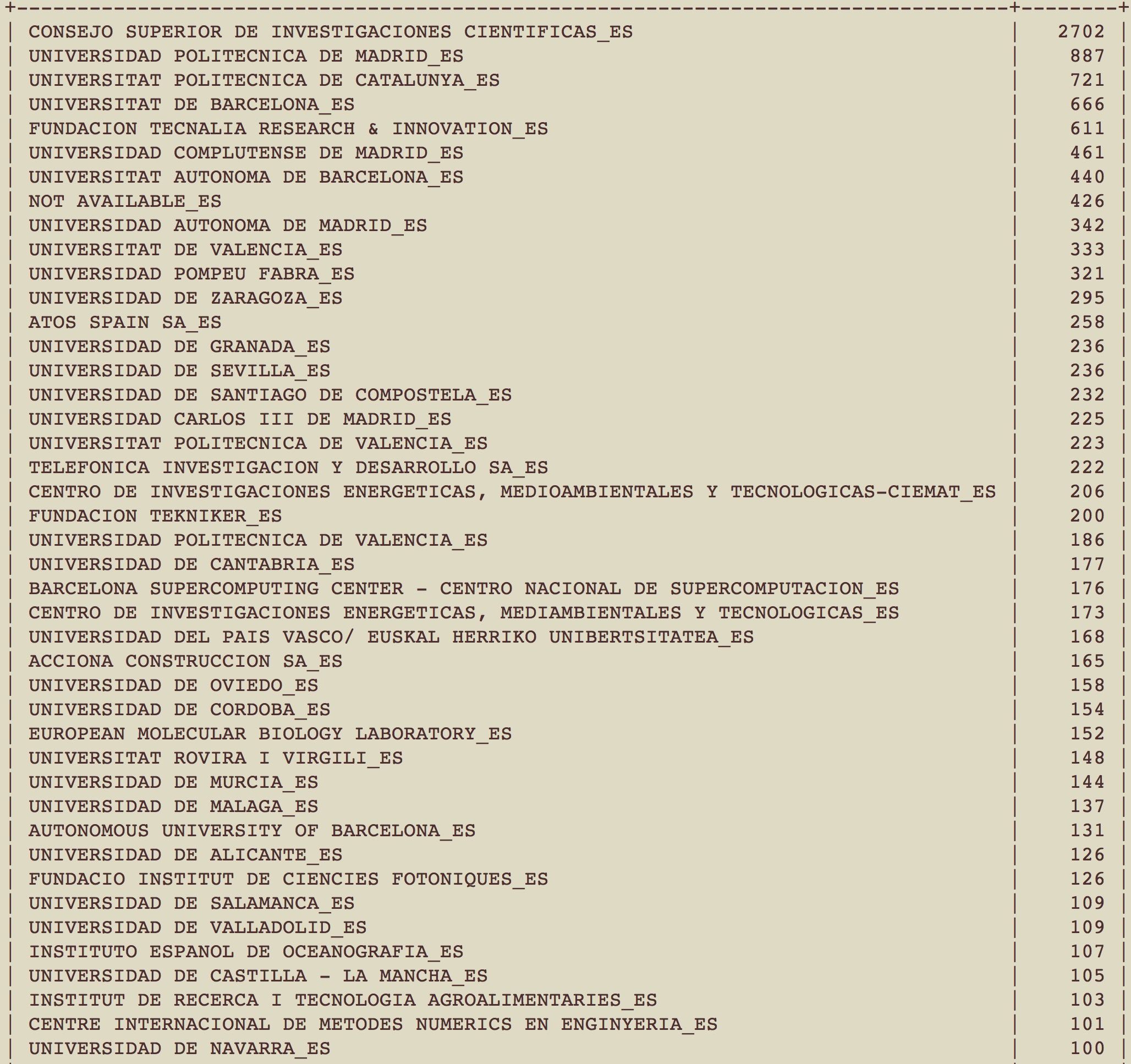
As an indication, in the organizations table there are only:

* 15.966 unique vatNumbers
* 38.670 unique URLs
* 45.200 unique ids

The table below shows the organizations that have a larger number of participations in the dataset:

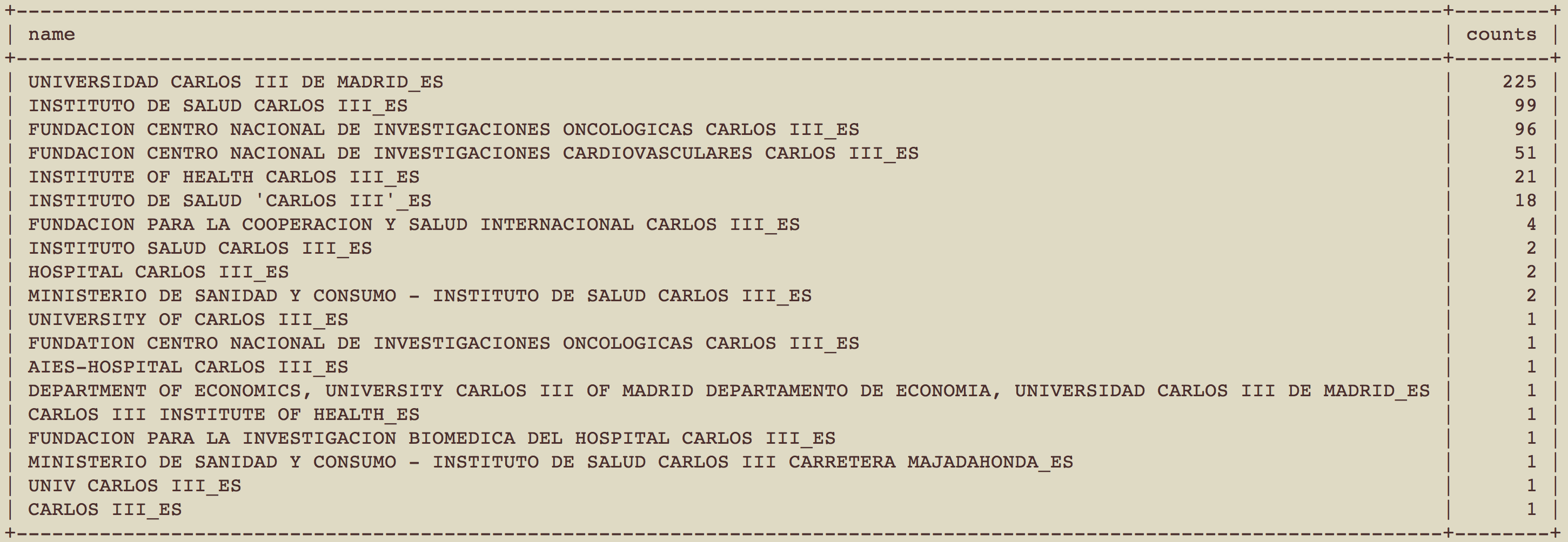


And now, the Spanish organizations that appear with more than 100 participations in the dataset (out of 39.858 spanish organizations in the dataset).

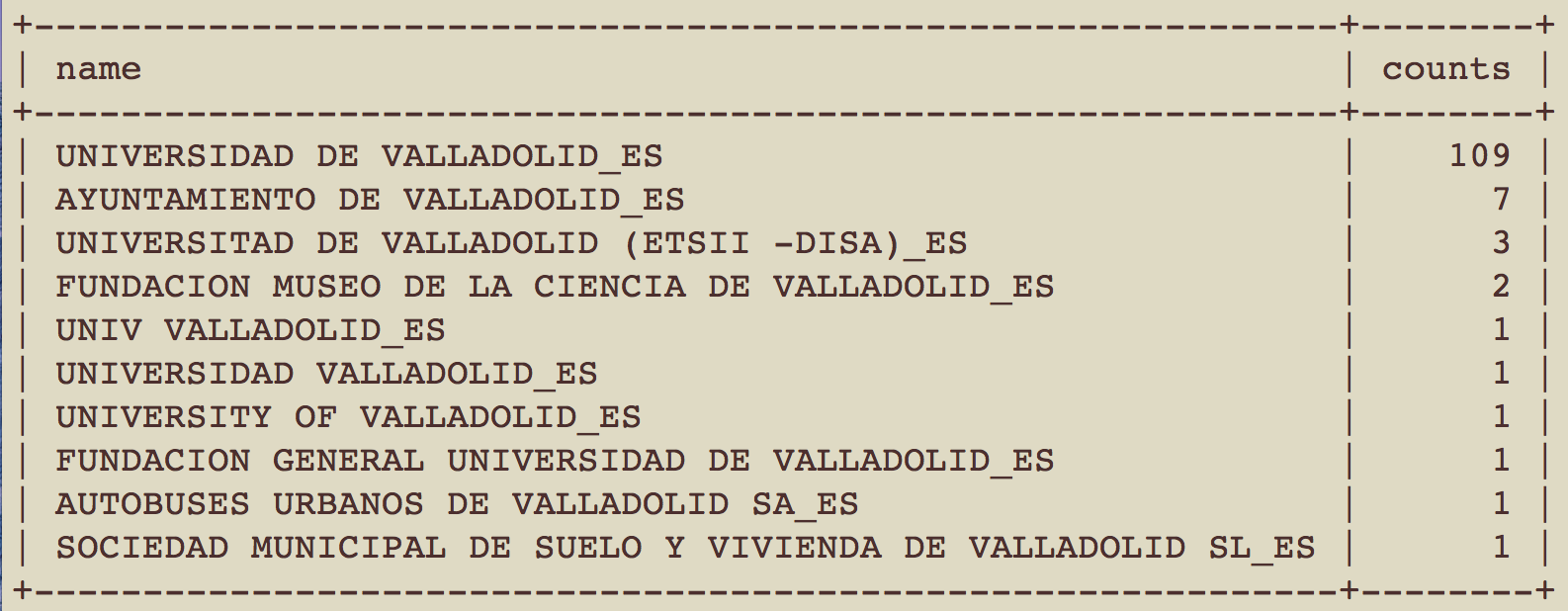


Even for this small subset of organizations we see that Universidad Autónoma de Barcelona appears twice with different names (in Catalán and English). The same is observed for other queries:

Example 1) Search for organizations that contain “Carlos III”



Example 2) Search for organizations that contain “Valladolid”



Clearly, some simple rules could be applied to greatly reduce the number of organizations in the dataset, and it is not difficult to think of a few simple rules for that, e.g.:

* Use translation in different languages of very common words (Universitat, University, Universidad)
* Use other fields that can be exploited to that end: “street” and or “postcode”, the domain of the contact person when available, etc.
* Focus on finding repeated organizations with a large number of participations, constraining the search by country

### 5.3. Data summary

Currently, we have a total of 118.581 organizations with a total of 503.637 participations.

Fieldnames in the tables are in most cases self-explanatory. We should just clarify that the orgzproject table includes also fields ‘activityType’ and ‘country’ that in principle could be found just in the organizations table. This has been done because these fields are expected to be of common use, and in this way we can exploit joins of the tables projects and orgzproject, without needing to include in the join the table organizations. This is important, since currently the organization name is being used as the primary key for the table organizations.

Pending tasks:

* Improve disambiguation of organizations
* Include an autoincrement index for the primary key of the table organizations, and use that in orgzproject instead of the full name.

### 5.4. Auxiliary 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’).

## 6. Researcher Information in CORDIS dataset

Very limited information is included in the dataset for researchers. We have only access for researchers in H2020 projects, and only for a very reduced number of funding schemes.

Nevertheless, this information has been included in the dataset following a scheme similar to that we have explained for organizations: we have included a table for researchers and another for the pairs research-project. As a result, we ended up with:

* 22.213 researchers
* 24.475 pairs researcher-project

In principle, the organization id of the researchers is provided in the files, but we have not incorporated it to the dataset because some researchers appear twice in the original files with different organization ids assigned to them.

It is also worth mentioning that the original files contained the project reference, so it has been necessary to replace that by the project rcn for coherence with the projects table.