# DESCRIPCIÓN DE LA BASE DE DATOS DEL PROYECTO DE INDICADORES DE PROYECTOS DE INVESTIGACIÓN

## 1. Control de Versiones

Versión actual: 3.0

Fecha: 12 de julio de 2018

Autor: Jerónimo Arenas García

## 2. Objetivo del documento

El presente documento describe la base de datos MySQL sobre la que se realiza el análisis basado en tópicos de los proyectos de investigación. En la presente versión del documento se han procesado un total de 119.519 proyectos de investigación en el intervalo temporal 2004-2016 proporcionados por la FECYT.

La importación de los datos requirió una primera etapa de procesado supervisado de los ficheros excel proporcionados, durante la cuál se unificó la denominación de las cabeceras, y se realizó cierta tarea de preprocesado, como la unificación de la duración de proyectos (en años), la introducción del Título de Convocatoria en los ficheros en los que no estaba presente, revisión de nombres de columnas, supresión de columnas con todos los datos perdidos o iguales a cero, etc.

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\_FECYT'

dbSERVER = 'vanir.tsc.uc3m.es'

## 3. Estructura de la Base de Datos

La base de datos contiene las tablas siguientes:

* **proyectos**: 119.519 entradas correspondientes a otros tantos proyectos.
* **investigadores**: 206.457 entradas correspondientes a todos los investigadores disponibles identificados por NIF.
* **investigadorproyecto**: cruza las referencias de proyectos y los NIF de investigadores, para saber qué investigador trabajó en qué proyecto, y el papel que desempeñó en el mismo. 586.797 entradas. Nótese que se incluyen en esta tabla las referencias de 89.755 proyectos, ya que no todos los proyectos incluyen el listado de investigadores participantes.
* **organismos**: Datos de 12.930 organismos identificados por CIF.
* **convocatorias:** Para cada título de convocatoria se incluyen los siguientes datos: Actuación, Programa y Subprograma Estatal, Entidad Convocante y Unidad Gestora. (52 entradas)
* **provincias:** Código numérico asignado a cada provincia del estado español (52 entradas)
* **CCAAs:** Código numérico asignado a cada CCAA (19 entradas)
* **areasANEP:** Código numérico asignado a cada área (26 entradas)
* **retos:** Código numérico asignado a cada reto (8 entradas)
* **TFEs:** Código numérico asignado a cada TFE (6 entradas)
* **orgtipos**: Código asignado a cada tipo de organismo (14 entradas)
* **orgsectors**: Código asignado a los sectores de organismos (5 entr)

El siguiente diagrama EER recoge de forma concisa el listado de campos de cada tabla:



Figura 1. Diagrama EER de la Base de Datos de Indicadores

## 4. Presencia de campos por convocatoria

A modo de resumen global acerca de la disponibilidad de datos, la siguiente tabla muestra la información disponible para cada convocatoria y título de convocatoria. Los campos no disponibles se muestran en rojo o en azul, utilizando el color azul para aquellos campos que UC3M considera sería de mayor utilidad poder conseguir.



Figura 2. Disponibilidad de campos según convocatoria y título de convocatoria. El color verde indica disponibilidad, el rojo ausencia de dicho campo, y el amarillo disponibilidad parcial.

De forma adicional, conviene resaltar que los datos reto y Tecnología Facilitadora Esencial (TFE) únicamente están disponibles para determinadas convocatorias, y en ocasiones con datos perdidos. Esto se ha marcado en la tabla en color azul, indicando el número de proyectos de la convocatoria para los que no se dispone de dicha información. En los casos en los que el fondo de la casilla es verde y se indica un código, dicho código corresponde al reto único para todos los proyectos de la convocatoria.

## 5. Descripción de campos de la Tabla de Proyectos

En las secciones restantes del documento vamos a describir brevemente el contenido de los diferentes campos que componen las tablas de la base de datos.

### 5.1. Campos de importación directa

* REFERENCIA: Identificador único del proyecto. Tenemos 119.519 proyectos
* TITULO CONVOCATORIA: Identificador del tipo de convocatoria, por ejemplo, SEIDI\_EXPLORA, SEIDI\_RETOS\_INVESTIGACION, ISCIII\_PI, etc ... Cada tipo de convocatoria puede tener proyectos en uno o varios años. El TITULO\_CONVOCATORIA sirve como llave para conocer el PROGRAMA, SUBPROGRAMA, ENTIDAD y UNIDAD correspondientes a cada proyecto, aunque estos valores están recogidos en la tabla convocatoria.

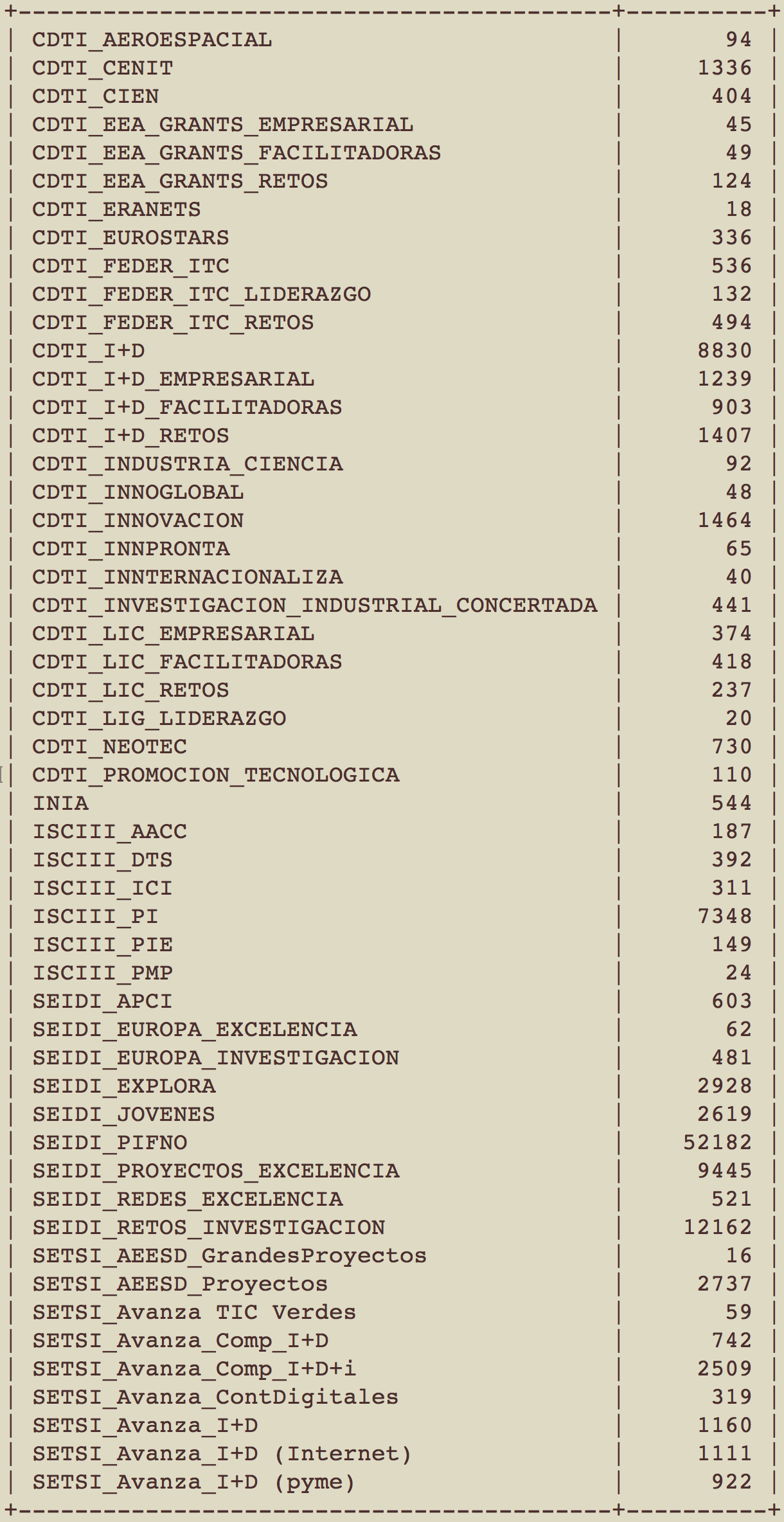


Tabla 1. Número de Proyectos según Título de Convocatoria

* CONVOCATORIA: Año de la convocatoria a la que concurrió el proyecto:

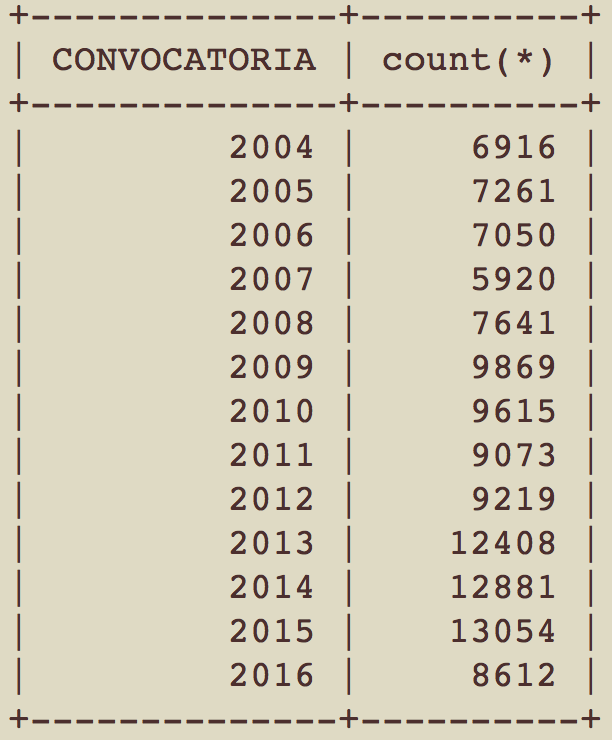


Tabla 2. Número de Proyectos según Convocatoria

* TIPO DE PROYECTO: Individual, Coordinado (==En cooperación) o Multicéntrico.
* STATUS: Aprobado o Denegado (67.770 Y 51.749 proyectos, respectivamente)
* CODIGOS NAB 2007: Clasificación temática que en principio no utilizaremos
* PROGRAMA: El acrónimo del área encargada de la evaluación (p. ej, TEC, TIN, etc ... ). Según se observa en la tabla, este campo está perdido para algo más del 16% de los proyectos.

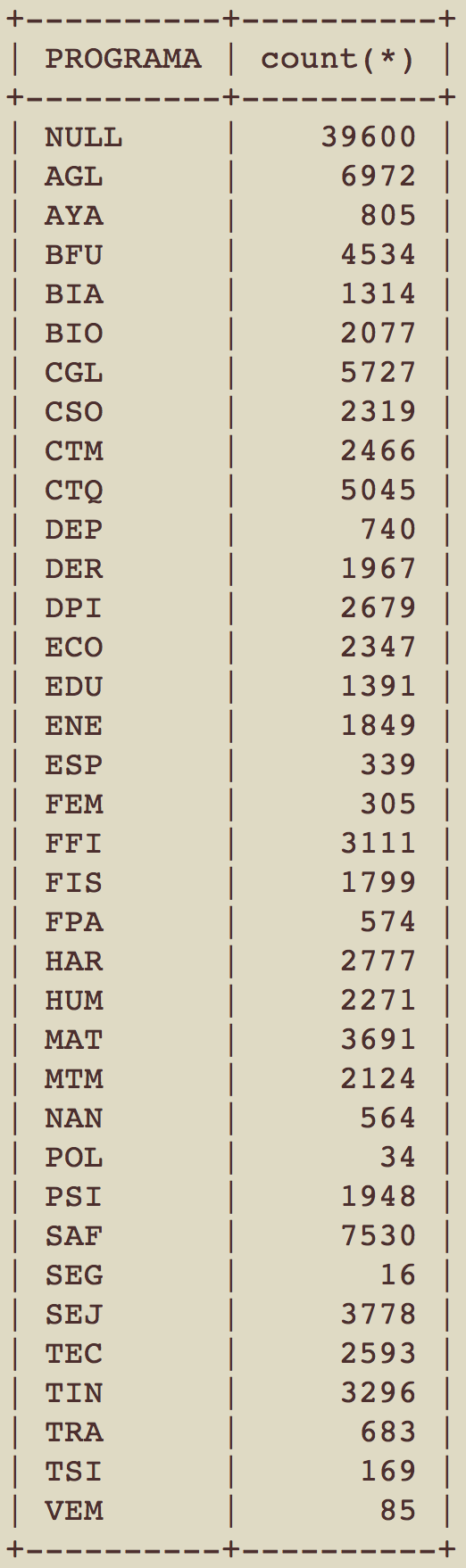


Tabla 3. Número de Proyectos según Programa.

* RETO: Reto en el que se enmarca el proyecto

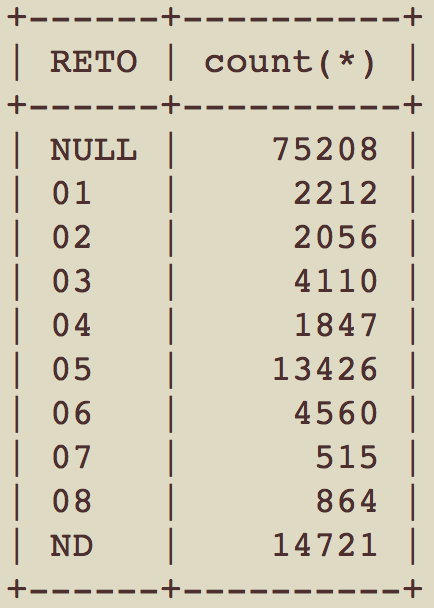
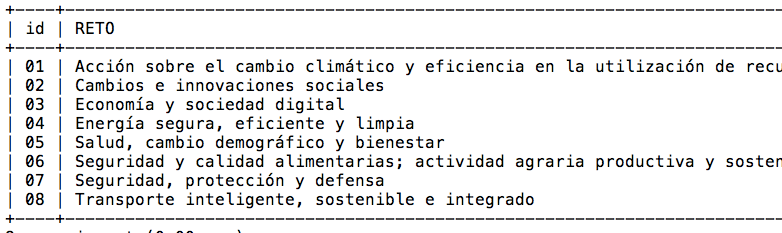
 

Tabla 4. Reparto de Proyectos según RETO.

* AREA ANEP: Clasificación temática que posiblemente sí utilicemos.

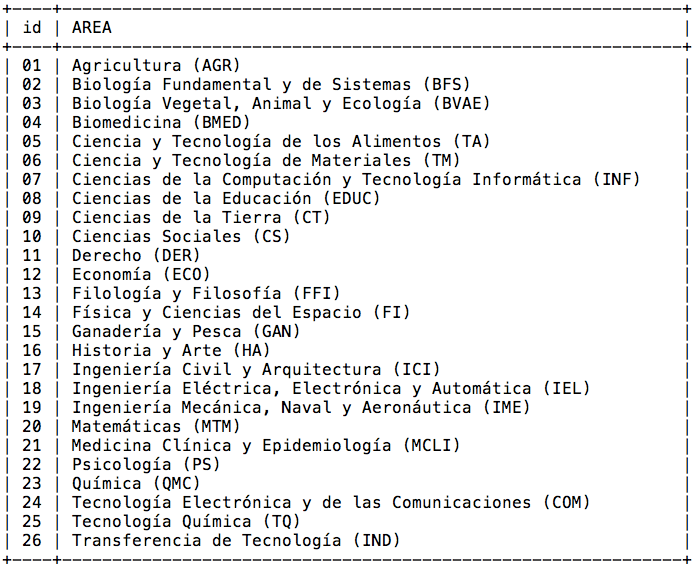
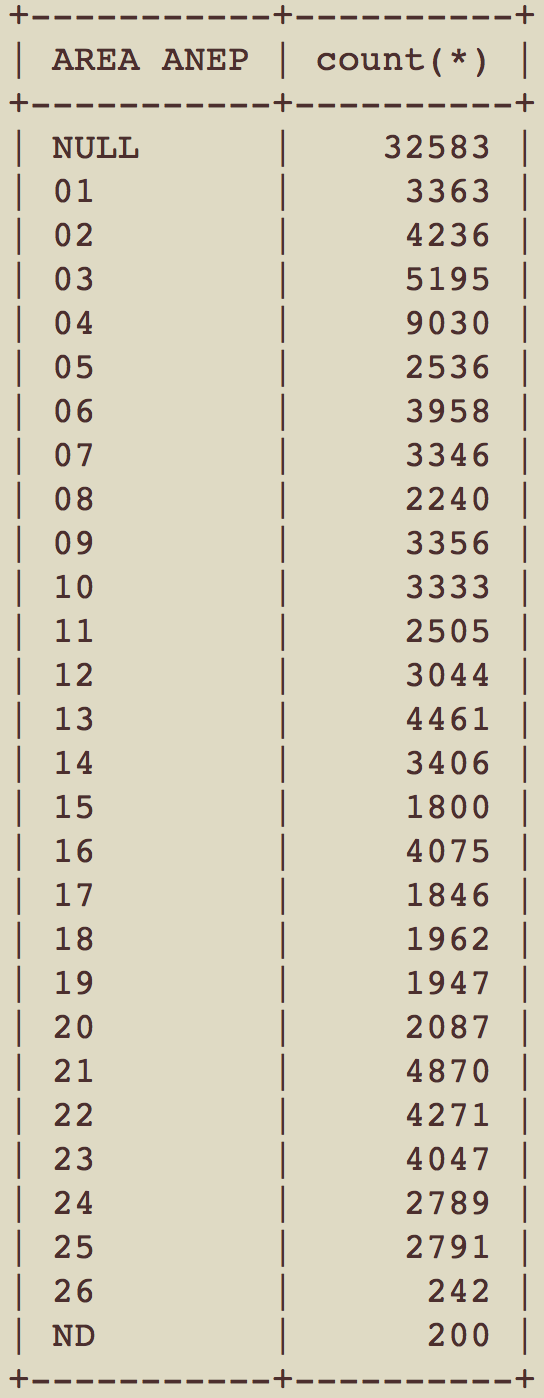


Tabla 5. Reparto de Proyectos según AREA ANEP. Desde un punto de vista práctico, el dato no está disponible para los casos en los que el valor queda a NULL, o toma valor “ND”.

* TFE: Tecnología Facilitadora Esencial del proyecto. No disponible para casi el 90% de los proyectos de la base de datos.
* DURACION: Duración del proyecto en años (tiene decimales). Valor mínimo es 0,08333 (1 mes) y máximo es de 6 años. Duración media de 2,73 años. Se han excluido del cálculo de dicho valor medio un total de 8.727 proyectos para los que consta una duración de 0 años.
* TITULO: Disponible para todos los proyectos, si bien únicamente toma 110.819 valores diferentes. Las repeticiones se deben a los proyectos coordinados, y de hecho el número real de repeticiones es incluso mayor, ya que en varias convocatorias se añade al final del título el número de proyecto coordinado (e.g., 1/4, 2/4, 3/4 y 4/4).
* KEYWORDS: Con distintos separadores. Disponible para 78.091 proyectos (únicamente 70.722 valores diferentes).
* PALABRAS CLAVE: Con distintos separadores. Disponible para 105.423 proyectos (únicamente 97.721 valores diferentes).
* ABSTRACT: Abstract del proyecto. Disponible para 78.123 proyectos (75.851 valores sin diferentes).
* RESUMEN: Resumen en español. Disponible para 112.482 proyectos (104.727 valores únicos).
* TOTAL CONCEDIDO: Importe total concedido. Es el único campo de los económicos fiable. La suma de este campo es de 18.500 millones de euros.
* TOTAL SOLICITADO
* PERSONAL CONCEDIDO
* PERSONAL SOLICITADO
* CIFEMPRESA: CIF de la empresa, Universidad o centro de investigación que solicita la ayuda. Es el organismo al que pertenece el IP del proyecto.
* CENTRO: Nuevamente, el del IP del proyecto, en el momento de la solicitud.
* PROVINCIA: La del centro, en el momento de la solicitud.
* CCAA: La del centro, en el momento de la solicitud.

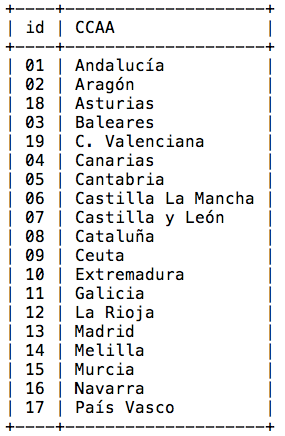
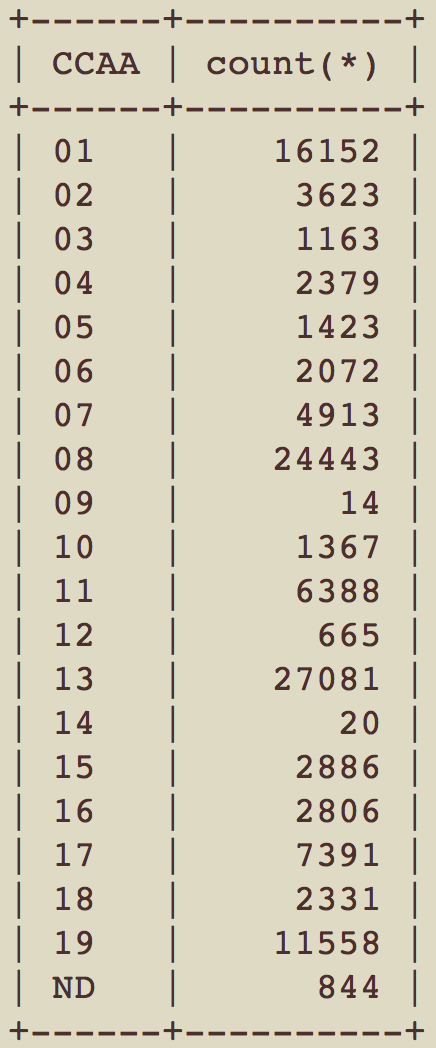


Tabla 6. Reparto de Proyectos según Comunidad Autónoma.

* INVEST\_FEMENINO: Número de mujeres declaradas en el equipo de investigación.
* INVEST\_MASCULINO: Número de hombres declarados en el equipo de investigación.
* INVEST\_EXTR: Número total de investigadores extranjeros en cada proyecto.
* INVEST\_TOTAL: Número total de investigadores.
* INVEST\_EDADES: Cadena de Texto con los años de nacimientos de los investigadores del proyecto, separados por comas.
* SEXO IP: Sexo del IP del proyecto. 29.562 (NULL, proyectos para los que no se dispone del equipo investigador), 28.867 (Femenino), 60.854 (Masculino), 236 (ND)
* EDAD IP: Año de nacimiento del IP del proyecto
* EXTR IP: IP con nacionalidad española o no. 29.751 (NULL, proyectos para los que no se dispone del equipo investigador), 86.908 (Español), 2.860 (Extranjero)
* NCOORDINADOS: Número de proyectos que se coordinan con el proyecto en cuestión. Toma valor no nulo para 27.091 proyectos (número total de proyectos coordinados en la base de datos), con un valor medio de 2,64 (i.e., en media los proyectos coordinados involucran 3,64 subproyectos) y un valor máximo de 26 (i.e., un proyecto con 27 subproyectos)
* PCOORDINADOS: String que contiene las referencias de los proyectos coordinados con el proyecto en cuestión.

### 5.2. Campos de contenido textual (título, keywords, resúmenes de proyectos y lemas) provenientes de procesado de los datos disponibles en los ficheros excel

* TITULO\_ESP: Título del proyecto en español, habiendo realizado la traducción (con la API de google) para aquellos proyectos de los que únicamente se dispone de título en inglés.
* PALABRAS\_ESP: Palabras clave del proyecto en español, habiendo realizado la traducción (con la API de google) para aquellos proyectos de los que no se dispone de palabras claves en castellano, y verificado el idioma de las palabras disponibles en los campos “PALABRAS CLAVE” y “KEYWORDS”. Este campo no está disponible para 15.519 proyectos para los que no se dispone de versión en inglés o español del campo (entre ellos, los 6.822 proyectos de SETSI 2004-2012).

* RESUMEN\_ESP: Resumen del proyecto en español, habiendo realizado la traducción (con la API de google) para aquellos proyectos de los que no se dispone de resumen en castellano, y verificado el idioma de los resúmenes disponibles en los campos “RESUMEN” y “ABSTRACT”. Este campo no está disponible para 8.199 proyectos para los que no se dispone de versión en inglés o español del campo (entre ellos, los 6.822 proyectos de SETSI 2004-2012).
* TITULO\_ENG: Título del proyecto en inglés, habiendo realizado la traducción (con la API de google y de DeepL Pro) para aquellos proyectos de los que únicamente se dispone de título en español.
* PALABRAS\_ENG: Palabras clave del proyecto en inglés, habiendo realizado la traducción (con la API de google) para aquellos proyectos de los que no se dispone de palabras claves en inglés, y verificado el idioma de las palabras disponibles en los campos “PALABRAS CLAVE” y “KEYWORDS”. Este campo no está disponible para 15.519 proyectos para los que no se dispone de versión en inglés o español del campo (entre ellos, los 6.822 proyectos de SETSI 2004-2012).
* RESUMEN\_ENG: Resumen del proyecto en inglés, habiendo realizado la traducción (con la API de google) para aquellos proyectos de los que no se dispone de resumen en inglés, y verificado el idioma de los resúmenes disponibles en los campos “RESUMEN” y “ABSTRACT”. Este campo no está disponible para 8.199 proyectos para los que no se dispone de versión en inglés o español del campo (entre ellos, los 6.822 proyectos de SETSI 2004-2012).
* LEMAS\_SETSI: Lemas proporcionados por la SETSI.
* LEMAS\_UC3M: Lemas en castellano obtenidos por la UC3M. Conservan el orden de aparición en el texto original, en caso de que se quieran utilizar para entrenar un Word Embeddings, por ejemplo. También conservan indicadores para señalar los límites entre frases.

El formato de este campo es el que sigue:

TITULO LEMATIZADO + ‘\*\*\*\*\*’ + PALABRAS CLAVE LEMATIZADAS + ‘\*\*\*\*\*’ + RESUMEN LEMATIZADO

En cada uno de los fragmentos, la separación entre frases se indica con el carácter de nueva línea ‘\n’.

Los n-gramas pueden identificarse fácilmente porque incluyen el carácter “\_” para separar los elementos del ngrama.

* LEMAS\_UC3M\_ENG: Lemas en inglés obtenidos utilizando la herramienta proporcionada por UPM. Conservan el orden de aparición en el texto original, en caso de que se quieran utilizar para entrenar un Word Embeddings, por ejemplo. También conservan indicadores para señalar los límites entre frases.

El formato de este campo es el que sigue:

TITULO LEMATIZADO + ‘\*\*\*\*\*’ + PALABRAS CLAVE LEMATIZADAS + ‘\*\*\*\*\*’ + RESUMEN LEMATIZADO

En cada uno de los fragmentos, la separación entre frases se indica con el carácter de nueva línea ‘\n’.

Los n-gramas pueden identificarse fácilmente porque incluyen el carácter “\_” para separar los elementos del ngrama.

### 5.3. Campos de asignación temática de los proyectos inferida mediante topic modeling

Para su potencial uso durante proyectos posteriores se ha añadido un indicador acerca de la pertenencia de cada proyecto al ámbito BIO, TIC o ENE. Dicho indicador se encuentra disponible de un proyecto anterior (Indicadores 2017) y se obtuvo mediante aplicación de un conjunto de reglas proporcionadas por FECYT, análisis basado en modelos de tópicos y etiquetado manual. No se encuentra disponible para todos los proyectos.

Se debe considerar que un valor “1” permite asegurar la pertenencia del proyecto a un ámbito concreto, mientras que un valor “0” no asegura la no pertenencia, ya que el etiquetado manual no fue exhaustivo sobre toda la base de datos. Finalmente, un valor NULL simplemente indica que el proyecto pertenece a proyectos proporcionados con posterioridad al análisis temático, y por lo tanto el indicador no se encuentra disponible.

* Ind2017\_BIO: 17.735 proyectos están marcados como BIO
* Ind2017\_TIC: 18.664 proyectos están marcados como TIC
* Ind2017\_ENE: 3.274 proyectos están marcados como ENE

## 6. Descripción de campos de la Tabla de Investigadores

* NIF
* NOMBRE
* YEAR NACIMIENTO
* SEXO: 124.571 Hombres y 94.004 mujeres. Perdido para 2.134 investigadores

## 7. Descripción de campos de la Tabla de Organismos

* CIF
* ORGANISMO OBJ SOCIAL (nombre)
* SECTOR

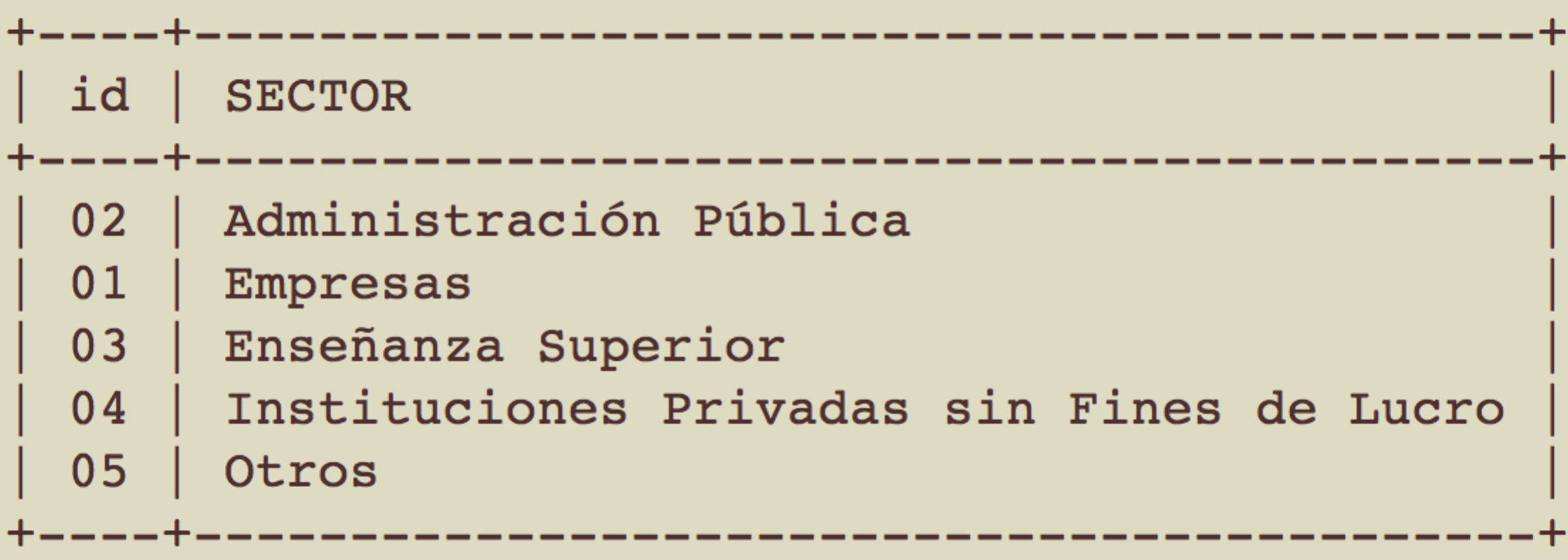
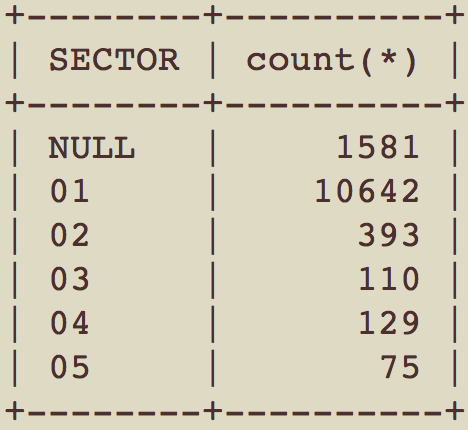


Tabla 7. Reparto de Organismos por Sector

* TIPO ENTIDAD

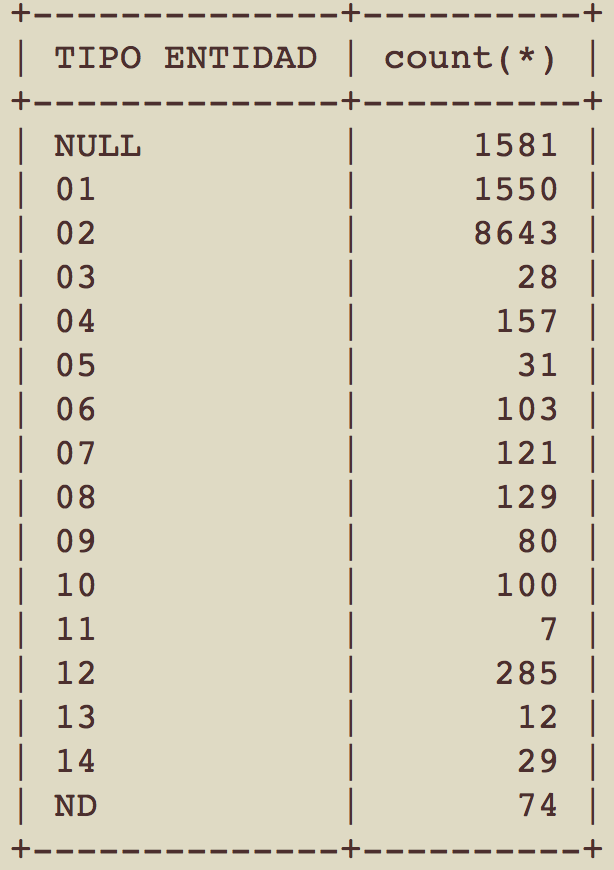
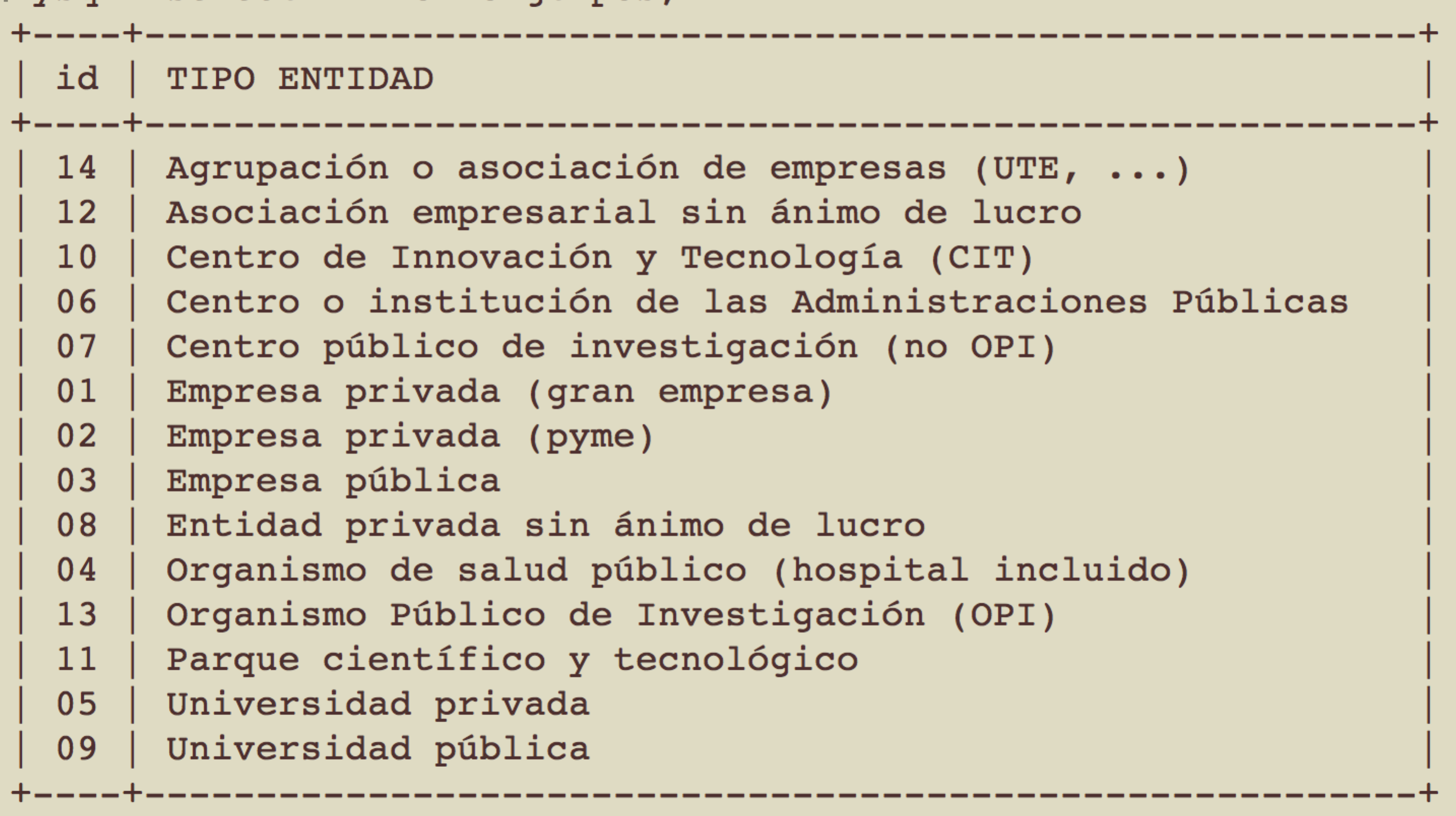
 

Tabla 8. Reparto de Organismos según Tipo de Entidad.