# DESCRIPCIÓN DE LA BASE DE DATOS DE PATENTES PATSTAT

[1. Objetivo del documento 2](#_Toc520162243)

[2. Algunos conceptos importantes 2](#_Toc520162244)

[3. Estructura de la Base de Datos y Descripción de Tablas 6](#_Toc520162245)

[3.1. Tabla tls201\_appln 9](#_Toc520162246)

[3.2. Tablas para análisis de temáticas 9](#_Toc520162247)

[3.2.1. tls202\_appln\_title 9](#_Toc520162248)

[3.2.2. tls203\_appln\_abstr 10](#_Toc520162249)

[3.3. Tablas de relaciones entre proyectos (prioridades y relaciones técnicas) 10](#_Toc520162250)

[3.3.1. tls204\_appln\_prior 10](#_Toc520162251)

[3.3.2. tls205\_tech\_rel 10](#_Toc520162252)

[3.3.3. tls216\_appln\_contn 10](#_Toc520162253)

[3.4. Tablas de personas 10](#_Toc520162254)

[3.4.1. tls206\_person 10](#_Toc520162255)

[3.4.2. tls906\_person 11](#_Toc520162256)

[3.4.3. tls226\_person\_orig 11](#_Toc520162257)

[3.4.4. tls207\_pers\_appln 11](#_Toc520162258)

[3.4.5. tls227\_pers\_publn 11](#_Toc520162259)

[3.5. Tablas de publicaciones y citas 12](#_Toc520162260)

[3.5.1. tls211\_pat\_publn 12](#_Toc520162261)

[3.5.2. tls214\_npl\_publn 12](#_Toc520162262)

[3.5.3. tls212\_citation 12](#_Toc520162263)

[3.5.4. tls215\_citn\_categ 12](#_Toc520162264)

[3.5.5. tls228\_docdb\_fam\_citn 12](#_Toc520162265)

[3.6. Tablas de Categorías 12](#_Toc520162266)

[3.6.1. tls209\_appln\_ipc 12](#_Toc520162267)

[3.6.2. tls210\_appln\_n\_cls 13](#_Toc520162268)

[3.6.3. tls222\_appln\_jp\_class 13](#_Toc520162269)

[3.6.4. tls223\_docus 13](#_Toc520162270)

[3.6.5. tls224\_appln\_cpc 13](#_Toc520162271)

[3.6.6. tls229\_appln\_nace2 13](#_Toc520162272)

[3.6.7. tls230\_techn\_field 13](#_Toc520162273)

[3.7. Otras tablas auxiliares 14](#_Toc520162274)

[3.7.1. tls801\_country 14](#_Toc520162275)

[3.7.2. tls901\_techn\_field\_ipc 14](#_Toc520162276)

[3.7.3. tls902\_ipc\_nace2 14](#_Toc520162277)

[3.7.4. tls904\_nuts 15](#_Toc520162278)

## 1. Objetivo del documento

El presente documento describe la base de datos MySQL que contiene el corpus de patentes. Se trata de hacer una versión más digerible que el documento “oficial” de casi 300 páginas, que puede descargarse por ejemplo de aquí:

<http://www.tsc.uc3m.es/~jarenas/DataCatalog_v5.10.pdf>

El volcado de los datos se ha realizado haciendo uso del proyecto de Github: <https://github.com/simonemainardi/load_patstat>

Y con un poco de paciencia y prueba y error, ya que el tamaño de la base de datos completa es de unos 350 GB. Por este motivo, no todas las tablas se han volcado, aunque las que están pendientes se podrían volcar en caso de que alguien las considerase necesarias.

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\_Pa\_PATSTATS'

dbSERVER = 'vanir.tsc.uc3m.es'

Por algún motivo, en algún momento decidí llamar a la base de datos ‘PATSTATS’, y así se ha quedado en la base de datos, pero el acrónimo habitual y que debemos usar para referirnos a ella es PATSTAT (sin la S final)

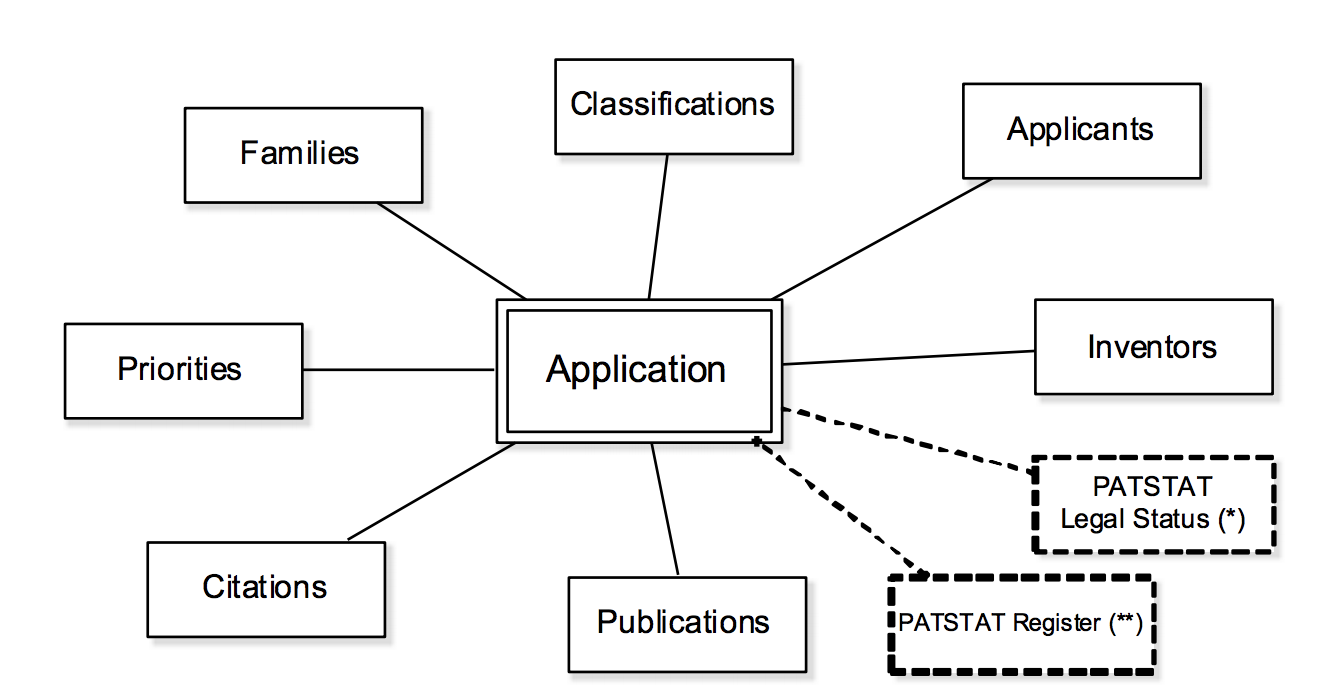
## 2. Algunos conceptos importantes

Lo que sigue es una descripción laxa, y probablemente sujeta a bastantes errores acerca del procedimiento de patentes, que tiene su reflejo en la propia estructura de la base de datos. Si alguien se lee con detalle el documento completo y se siente con ganas de actualizar esta sección, ¡adelante!

La siguiente figura recoge las entidades más importantes de la base de datos. Aunque realmente todo debería girar en torno a publicaciones, que son las que tienen autores y demás, en realidad se ha optado en PATSTAT porque el elemento central de la BD sea la “aplicación” (solicitud), pudiendo tener cada aplicación una o varias publicaciones asociadas[[1]](#footnote-1).

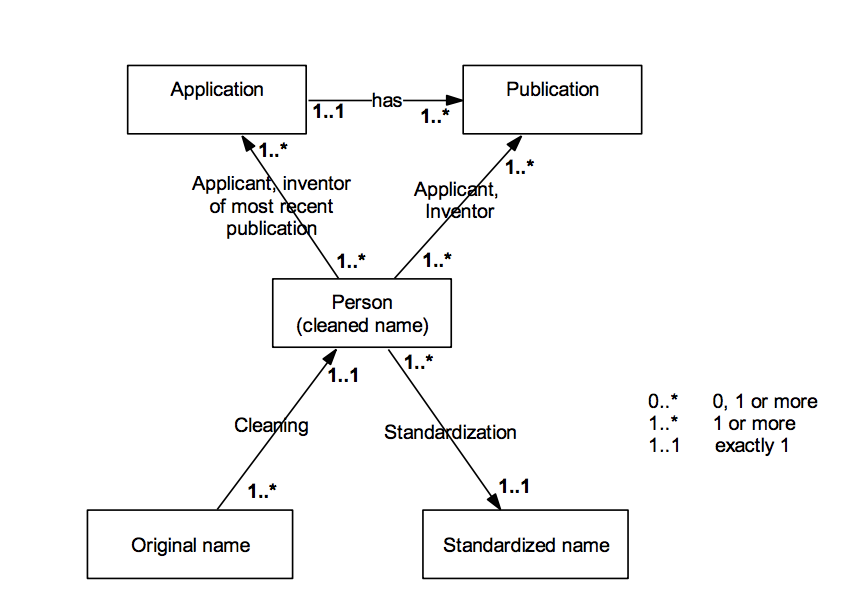
Durante la vigencia de una patente, se emiten varias publicaciones. Una aplicación tiene al menos una publicación; de lo contrario, se trataría como confidencial y no sería accesible en ninguna base de datos. Existen excepciones a esta regla, como comentaré más adelante: aplicaciones “artificiales” que no contienen ninguna publicación.

Las aplicaciones que cubren la misma invención o similar se agrupan en familias. Existen varias definiciones familiares; en consecuencia, hay varios tipos de familias. Cada familia contiene una o más aplicaciones como miembros de la familia. Para cada tipo de familia se aplica que cada aplicación pertenece exactamente a una familia de ese tipo. Quizá el caso más claro es el de una misma invención para la que se solicita la patente en varias oficinas de forma simultánea o próxima en el tiempo. Todas esas patentes quedarían agrupadas en una única familia.

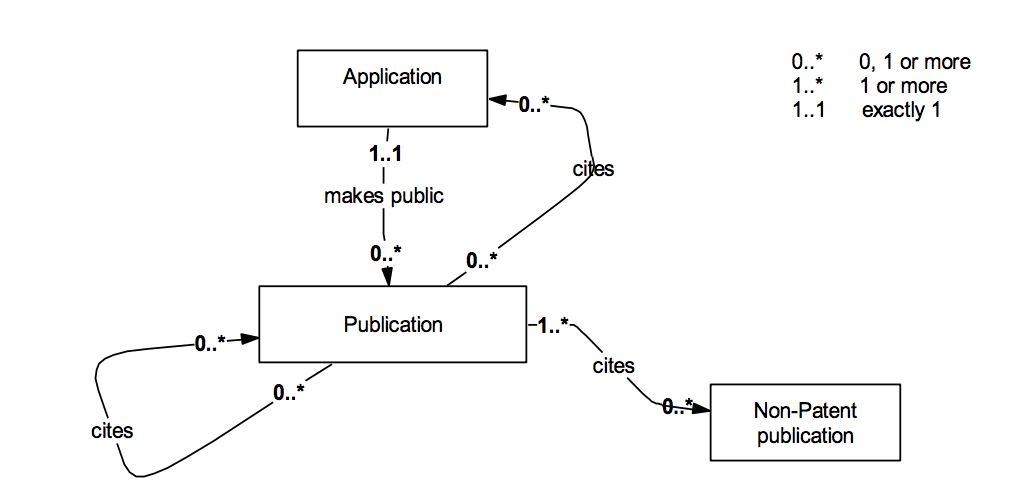


Otros elementos relevantes son:

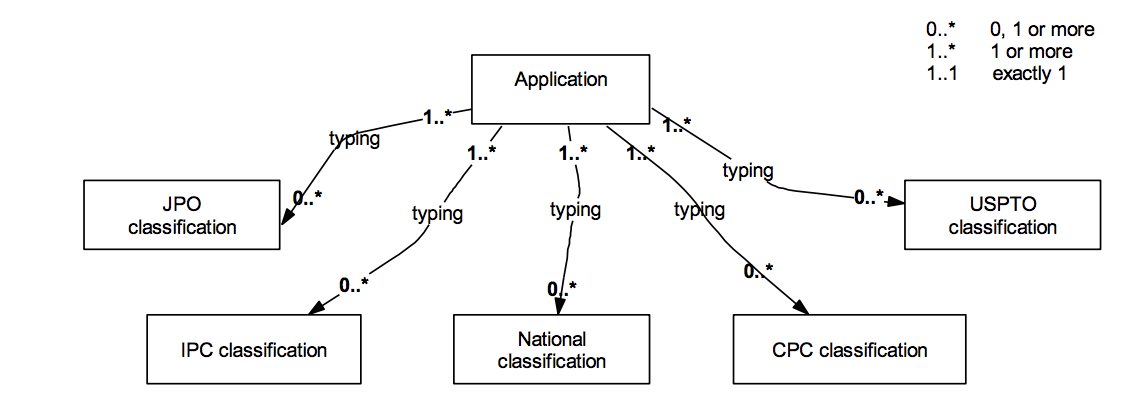
* Personas: La BD tiene una tabla de personas, y las que figuran en esa tabla son los autores de las publicaciones de las patentes. Es decir, la relación es indirecta, en principio no hay autores de “solicitudes”, pero se toma en su lugar los autores de las publicaciones asociadas a una solicitud. Hay que mencionar que los inventores pueden ser tanto personas físicas como personas jurídicas. He encontrado un proyecto github que trata de clasificar las personas según ese criterio. Cada aplicación puede tener una o varias personas asociadas, y una persona puede ser solicitante pero también inventor.



* La base de datos incluye, asociada a cada solicitud, un título y un abstract. El abstract consiste en un resumen de un párrafo de la invención que se muestra en la primera página de una publicación. Por diseño, en PATSTAT los resúmenes no están relacionados con la publicación individual (aunque es de una publicación de donde se toma), sino con la aplicación de la publicación. Los resúmenes pueden estar en cualquier idioma. PATSTAT contiene solo 1 resumen por aplicación. Se prefieren resúmenes en inglés
* Además de publicaciones, se incluye una tabla de publicaciones no patentes (*non-patent literature*, fundamentalmente artículos de investigación). También se incluye una tabla de citas. Las citas aparecen en distintos momentos durante la vida de la patente, por ejemplo identificadas en la solicitud inicial, o marcadas como relevantes por los examinadores. El elemento citado puede ser tanto otra publicación asociada a patentes como un artículo de non-patent literature. En raras ocasiones se cita una solicitud de patente, que no se puede identificar dentro de la base de datos, y esto se resuelve en PATSTAT incluyendo en la tabla “solicitudes artificiales”, sin publicaciones asociadas, y por tanto ni personas ni título o abstract. Existe literatura de investigadores estudiando el efecto de estas solicitudes en el análisis a partir de PATSTAT (en realidad hay mucha literatura sobre muchos aspectos de esta base de datos).



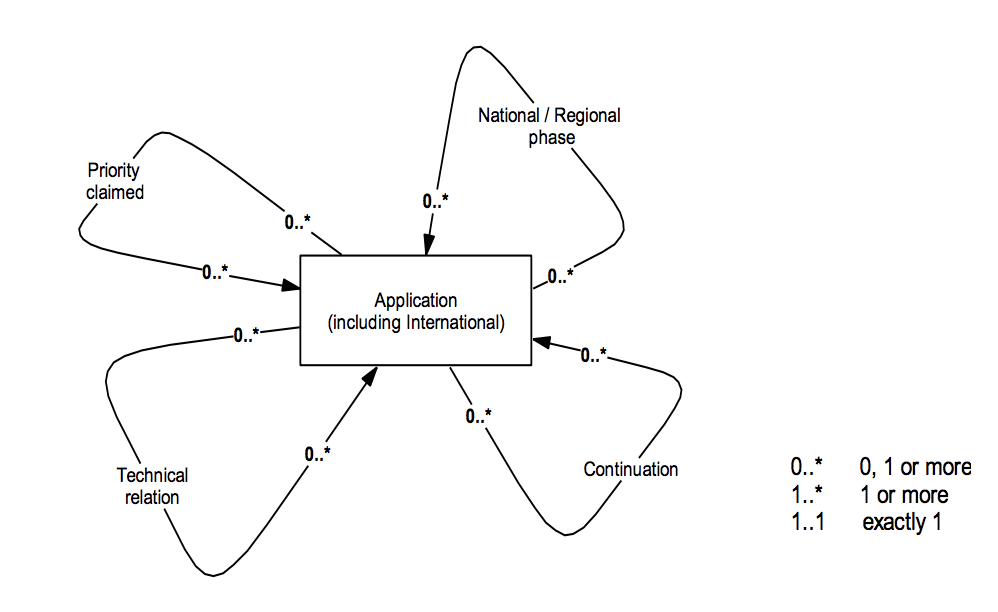
* Clasificaciones: Se incluyen distintos criterios tipo códigos FORD, UNESCO, etc … Aquí, la clasificación que tienen más cobertura es la de IPC (<http://www.wipo.int/classifications/ipc/en/>), pero hay otras (no todas precargadas).



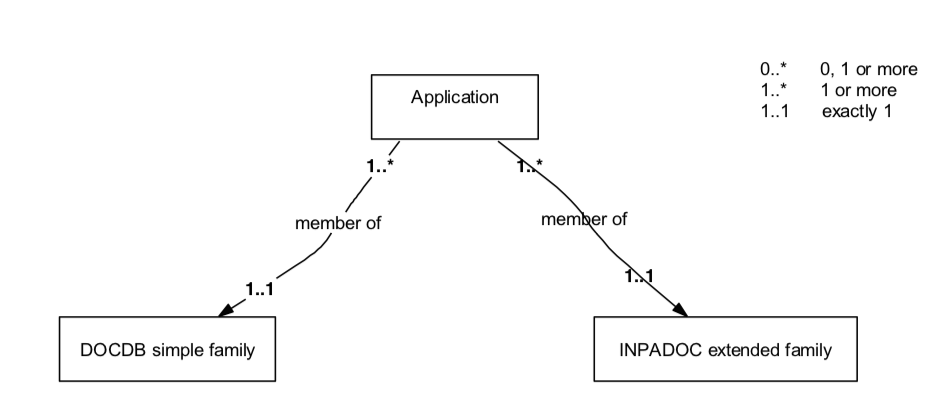
Además se dispone de una clasificación por campo técnico, que puede tomar 35 valores diferentes. La relación entre la clasificación según IPC y según campo técnico puede consultarse en caso necesario aquí:

<http://www.wipo.int/export/sites/www/ipstats/en/statistics/patents/xls/ipc_technology.xls>

* Información de relación entre aplicaciones: La BD incluye directamente algunas tablas que son grafos sobre las solicitudes. Estas relaciones pueden ser al menos de dos tipos:
  + Por prioridad: “In patent, industrial design rights and trademark laws, a priority right or right of priority is a time-limited right, triggered by the first filing of an application for a patent, an industrial design or a trademark respectively. The priority right allows the claimant to file a subsequent application in another country for the same invention, design, or trademark effective as of the date of filing the first application. When filing the subsequent application, the applicant must claim the priority of the first application in order to make use of the right of priority. The right of priority belongs to the applicant or his successor in title.”
  + Por relación técnica: detectada por algún evaluador. La documentación sugiere que estas relaciones no son muy empleadas, al menos en la actualidad.



* Familias: Cada aplicación pertenece a una familia de DOCDB (un base de datos de documentos mantenida por la oficina de patentes europea), y a una familia extendida de INPADOC.



La definición de familias es importante, porque además probablemente nos permita reducir el número de documentos con los que trabajar en algunos proyectos . Os copio lo que dice al respecto la documentación de PATSTAT:

*“Applications which cover the same or similar invention are grouped into families. There are several definitions of families; consequently there are several types of families. The EPO uses these types of families:*

* + *Simple family, also called DOCDB family or Espacenet patent family:  
    All applications which are member of the same simple family do have the same priorities. The technical content of these family members is regarded as (almost) identical, so their publications are sometimes called “equivalent”.*
  + *Extended family, also called INPADOC family:  
    All applications which are member of the same extended family are directly or directly linked to the same root priority application. Usually the applications are related to the same technical invention, but their individual content may differ.*

*Every application belongs to exactly 1 simple family and to exactly 1 INPADOC family. The extended family is potentially “broader” than the simple family: Each extended family contains the applications of 1 or more simple families.”*

## 3. Estructura de la Base de Datos y Descripción de Tablas

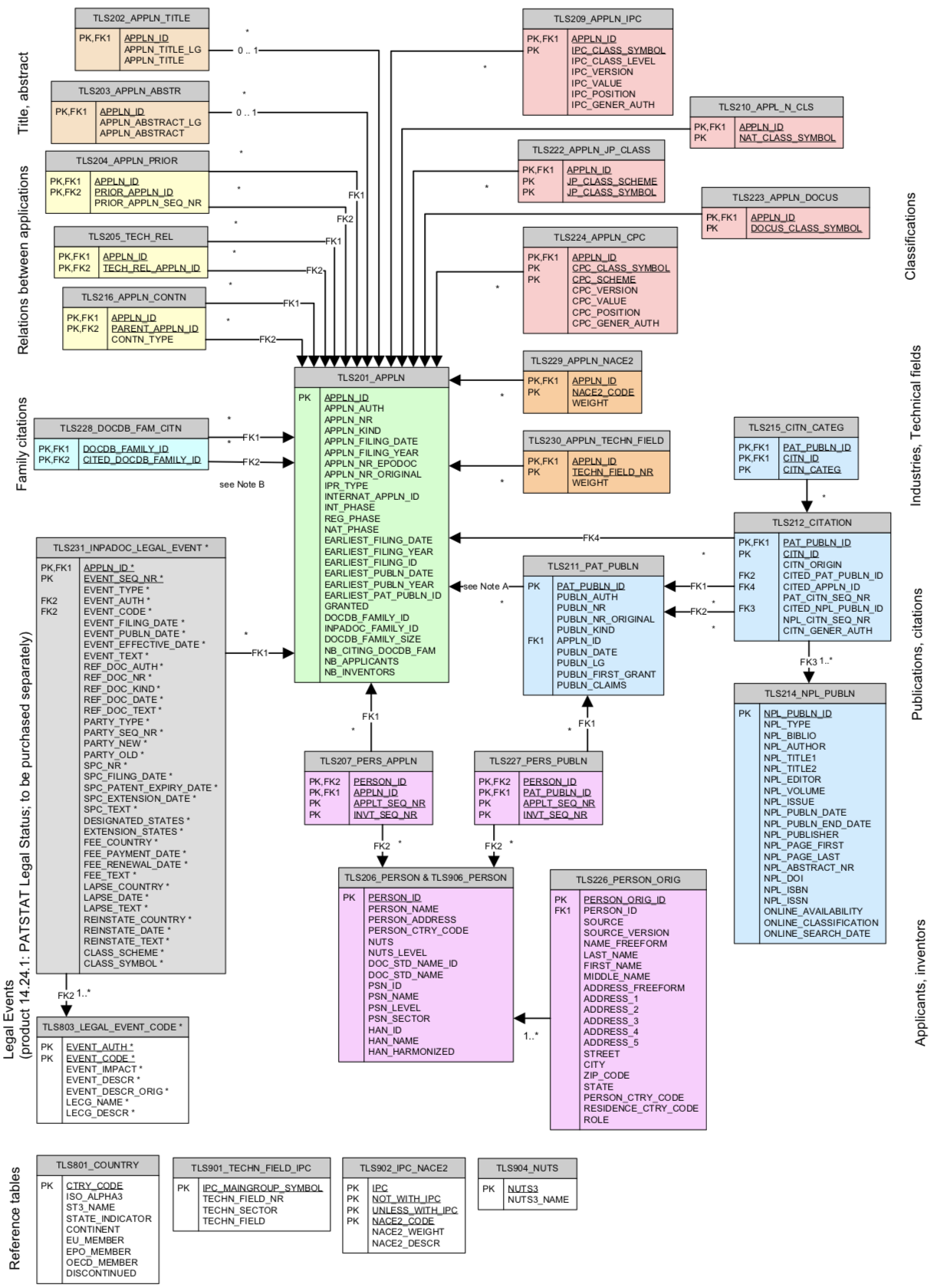
El diagrama EER es extenso, por lo que no vamos a hacer un análisis detallado de las tablas y sus campos. Sin embargo, se incluye la siguiente información:

* Status de la tabla: Si se encuentra disponible o no (en caso de que una tabla no esté disponible y la necesitéis, decídmelo, por favor)
* Propósito general de la tabla
* Descripción de algún campo concreto que sea relevante.

La descripción detallada puede encontrarse en la documentación de PATSTAT:

Las siguientes variables surrogadas (i.e., sin significado físico) identifican de forma única a los distintos agentes. Se pueden identificar porque acaban en ‘\_ID’, en contraste con las que acaban en ‘\_NR’ que sí tienen sentido físico.

* APPLN\_ID
* INTERNAT\_APPLN\_ID
* PRIOR\_APPLN\_ID
* TECH\_REL\_APPLN\_ID
* PERSON\_ID
* DOC\_STD\_NAME\_ID
* PAT\_PUBLN\_ID
* CITN\_ID
* NPL\_PUBLN\_ID
* CITED\_PAT\_PUBLN\_ID
* PARENT\_APPLN\_ID
* DOCDB\_FAMILY\_ID
* INPADOC\_FAMILY\_ID



### 3.1. Tabla tls201\_appln

Número de entradas: 91.738.066

Estado de la tabla: Disponible

Es la tabla central de la base de datos, y contiene una entrada por cada solicitud, utilizando como llave primaria el índice surrogado ‘appln\_id’. Existen 4 rangos de valores de este índice asociados a otros 4 tipos de solicitudes:

* Rango 1: appln\_id entre 1 y 900.000.000: Corresponden a solicitudes que tienen un documento en DOCDB, y por tanto al menos una publicación asociada.
* Rango 2: appln\_id entre 900.000.001 y 930.000.000: Corresponden a “solicitudes artificiales” creadas en PATSTAT en correspondencia con la reivindicación de priority de otra solicitud. Es necesario crear “solicitudes artificiales” cuando no existe una publicación en DOCDB para dicha priority.
* Rango 3: appln\_id entre 930.000.001 y 960.000.000: Corresponden a “solicitudes artificiales” que se crean en PATSTAT debido a que se citan publicaciones de patentes que no aparecen en DOCDB. Por ese motivo, se crea una “publicación artificial”, y dado que todas las publicaciones deben ir asociadas a una solicitud de patente, también una “solicitud artificial”.
* Rango 4: appln\_id entre 960.000.001 y 999.999.999: Similar al caso anterior, pero para aquellos casos en los que se cita directamente es una solicitud de patente, que no aparece en la base de datos.

El número total de patentes no artificiales (Rango 1) es de 82.147.124.

El campo appln\_filing\_date proporciona la fecha en que la aplicación se recibió físicamente en la autoridad de patentes correspondiente. El dato se puede considerar correcto para las patentes de Rango 1 y Rango 2 y algunas de Rango 4 (cuando la fecha de la patente se tiene en la cita, se utiliza). Para aquellos otros casos, se utiliza como fecha por defecto ‘9999-12-31’.

Dos campos que también pueden resultar relevantes son ‘earliest\_filing\_id’ y ‘earliest\_filing\_date’ que contienen el id de la primera solicitud directamente relacionada con cada solicitud (la propia solicitud, una prioridad declarada, la aplicación internacional, e incluso las solicitudes relacionadas por technical\_relation), y la fecha de entrada de dicha primera solicitud.

En este sentido, también tiene sentido analizar las familias de DOCDB, encontrándonos que el número total de familias es de 56.492.581, considerando únicamente las correspondientes a solicitudes de Rango 1.

### 3.2. Tablas para análisis de temáticas

#### 3.2.1. tls202\_appln\_title

Número de entradas: 72.486.891

Estado de la tabla: Disponible

Contiene los títulos de las solicitudes de patente, junto con el idioma correspondiente.

#### 3.2.2. tls203\_appln\_abstr

Número de entradas original en PATSTAT: 49.133.801

Estado de la tabla: No hemos volcado la información de PATSTAT, ya que por sí sola ocupa unos 80 GB y pienso que nadie la va a utilizar realmente.

Esta información sería relevante sobre todo para el proyecto de tendencias, pero el coste de lematizar es elevado, y en principio en el proyecto se acordó que SESIAD pasaría los textos lematizados. No nos han pasado lemas, pero sí unos BOW para 40.075.363 solicitudes, por lo que en lugar de volcar la información de PATSTAT se ha añadido un campo nuevo appln\_abstr\_bow, donde se ha volcado la información correspondiente. Los campos de la tabla original ‘appln\_abstr\_lg’ y ‘appln\_abstract’ siguen estando en la tabla, pero se encuentran vacíos.

Por otro lado, no he mirado en detalle por qué el número de BoW pasados es inferior al número de entradas original de la tabla, pero es posible que el motivo sea que se hayan descartado aquellos abstract que no estaban en lengua inglesa.

### 3.3. Tablas de relaciones entre proyectos (prioridades y relaciones técnicas)

#### 3.3.1. tls204\_appln\_prior

Número de entradas: 39.730.627

Estado de la tabla: Disponible

Tabla de prioridades declaradas.

#### 3.3.2. tls205\_tech\_rel

Número de entradas: 4.092.486

Estado de la tabla: Disponible

Tabla de relaciones técnicas entre proyectos.

#### 3.3.3. tls216\_appln\_contn

Número de entradas: 3.079.656

Estado de la tabla: Disponible

Tablas de “continuaciones” declaradas.

### 3.4. Tablas de personas

#### 3.4.1. tls206\_person

Número de entradas: 55.601.405

Estado de la tabla: Esta tabla no se ha volcado, porque es como la tls906\_person, que sí se ha volcado, y que contiene más información que esta.

#### 3.4.2. tls906\_person

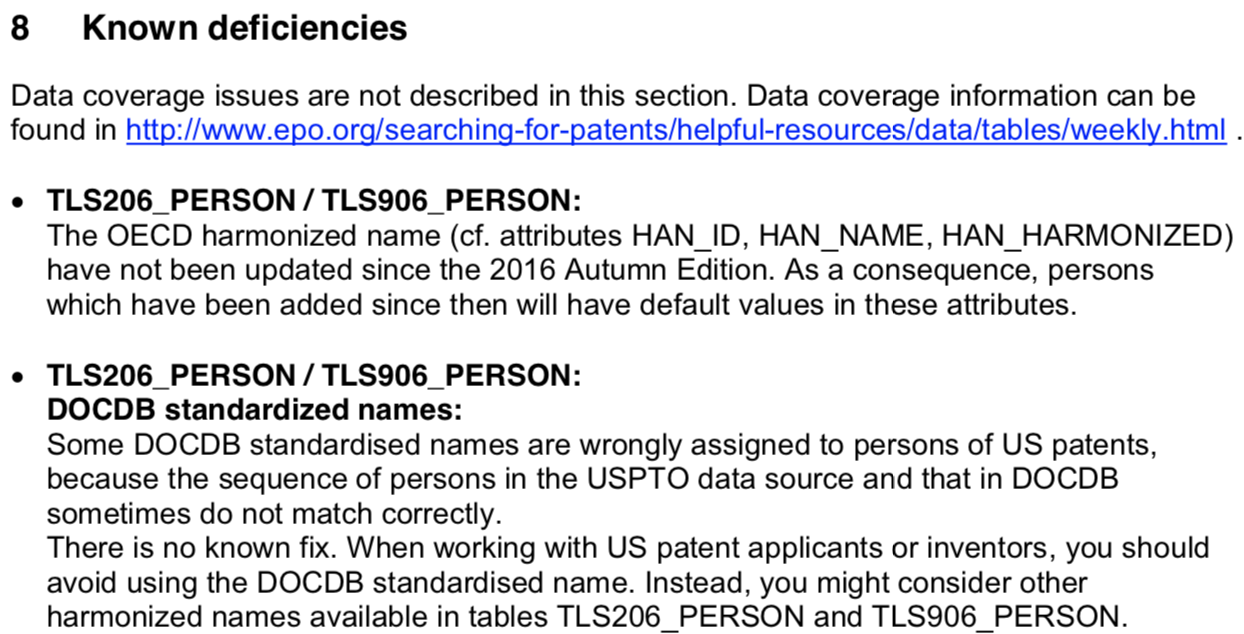
Número de entradas: 55.601.405

Estado de la tabla: Disponible

Contiene listado de personas que pueden ser personas físicas o jurídicas. Además, se incluyen distintas versiones normalizadas del nombre:

* Según DOCDB
* Según PATSTAT (Univ. Leuven)
* Según OECD

Existen algunas deficiencias que incluyo a continuación.



#### 3.4.3. tls226\_person\_orig

Número de entradas: 64.588.088

Estado de la tabla: No hemos volcado esta tabla por ser también de gran tamaño y no estar clara su necesidad.

Tabla que contiene las fuentes originales de las que se han tomado los datos que se han incluido en la 906.

#### 3.4.4. tls207\_pers\_appln

Número de entradas: 217.993.472

Estado de la tabla: Disponible

Relación entre personas y solicitudes. Incluye qué persona ha participado en qué solicitud de patente, y su posición en la lista ordenada de solicitantes y/o inventores.

#### 3.4.5. tls227\_pers\_publn

Número de entradas: 266.045.610

Estado de la tabla: Disponible

Similar a la anterior, pero respecto de publicaciones.

### 3.5. Tablas de publicaciones y citas

#### 3.5.1. tls211\_pat\_publn

Número de entradas: 106.202.165

Estado de la tabla: Disponible

Información bibliográfica de publicaciones relacionadas con patentes.

#### 3.5.2. tls214\_npl\_publn

Número de entradas: 34.971.774

Estado de la tabla: Disponible

Información bibliográfica de publicaciones de *non-patent-literature*.

#### 3.5.3. tls212\_citation

Número de entradas: 273.371.080

Estado de la tabla: Disponible

Incluyes relaciones de citación entre las publicaciones de patentes y la de otra literatura, así como relaciones de citas entre solicitudes de patentes y *non-patent literature.*

#### 3.5.4. tls215\_citn\_categ

Número de entradas: 45.450.583

Estado de la tabla: Disponible

Tipologías de las citaciones.

#### 3.5.5. tls228\_docdb\_fam\_citn

Número de entradas: 150.222.683

Estado de la tabla: Disponible

Contiene citas entre solicitudes de familias de DOCDB. Contiene una entrada por cada publicación de una familia A que ha citado a otra publicación de una familia B.

### 3.6. Tablas de Categorías

#### 3.6.1. tls209\_appln\_ipc

Número de entradas: 227.484.083

Estado de la tabla: Disponible

Categorías de *International Patent Classification.*

#### 3.6.2. tls210\_appln\_n\_cls

Número de entradas: 25.788.528

Estado de la tabla: No está disponible. Si alguien la necesita, que la pida.

Categorías proporcionadas por algunas agencias nacionales de patentes.

#### 3.6.3. tls222\_appln\_jp\_class

Número de entradas: 334.118.081

Estado de la tabla: No hemos volcado esta tabla por ser también de gran tamaño y no estar clara su necesidad.

Categorías de la Agencia Japonesa. Por la documentación parece que es más completa y se renueva de forma más continua que la de IPC. Pero sigue la estructura de la IPC.

#### 3.6.4. tls223\_docus

Número de entradas: 39.008.391

Estado de la tabla: Disponible

Categorías de la agencia de los USA. No sigue la categoría de IPC.

#### 3.6.5. tls224\_appln\_cpc

Número de entradas: 205.135.275

Estado de la tabla: No hemos volcado esta tabla por ser también de gran tamaño y no estar clara su necesidad.

No me he enterado muy bien en qué consiste esta información. Revisar documentación de PATSTAT.

#### 3.6.6. tls229\_appln\_nace2

Número de entradas: 105.145.274

Estado de la tabla: Disponible

Contiene categorías NACE. Cada entrada consiste en un appln\_id + un código NACE2, junto con el peso asignado a dicha asignación. De esta manera, todas las entradas existentes para un único appln\_id (y distintos NACE2) suman la unidad.

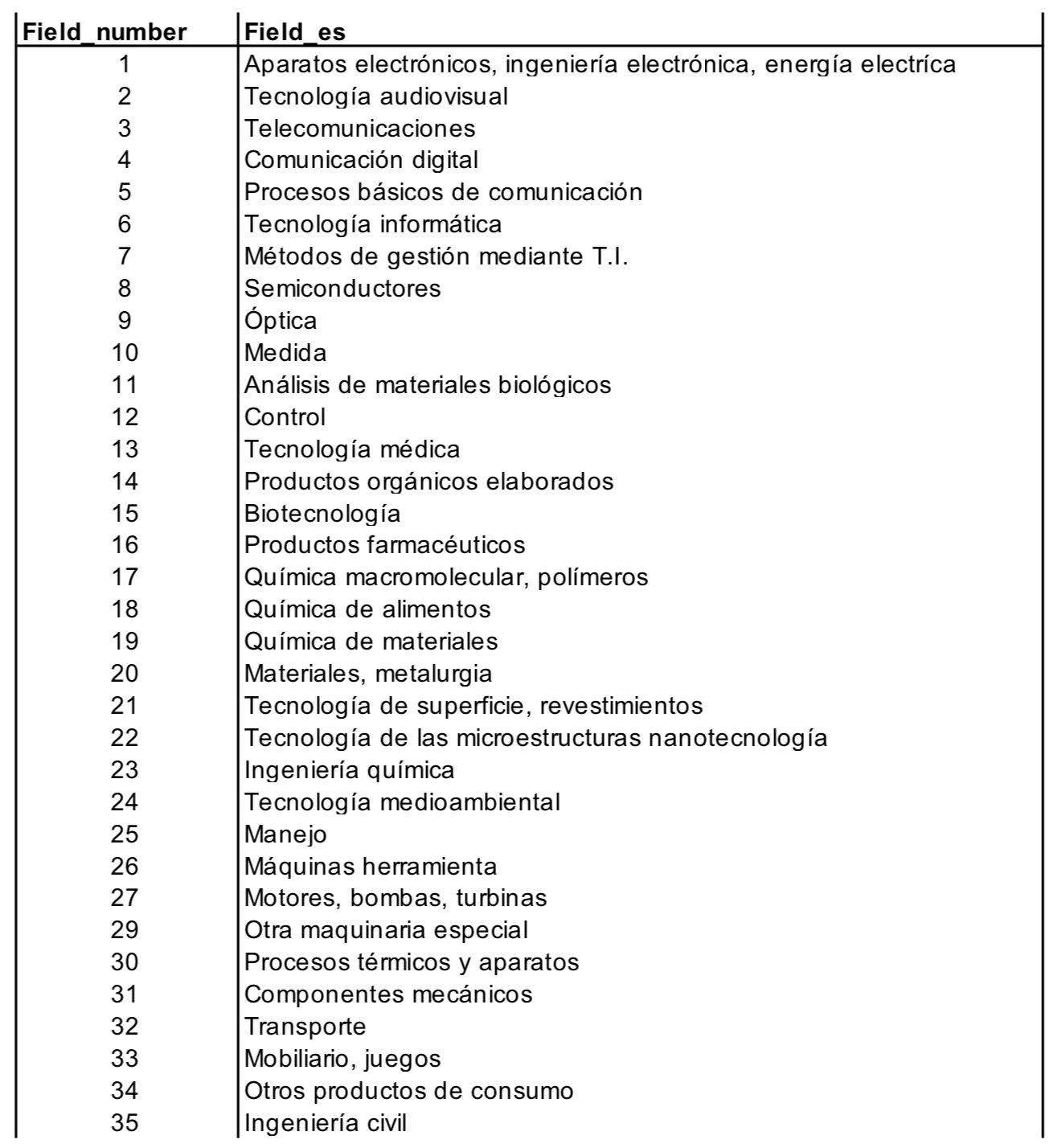
#### 3.6.7. tls230\_techn\_field

Número de entradas: 105.496.717

Estado de la tabla: Disponible

Contiene categorías técnicas. Cada entrada consiste en un appln\_id + un código de categoría, junto con el peso asignado a dicha asignación. De esta manera, todas las entradas existentes para un único appln\_id (y distintos códigos de categoría técnica) suman la unidad.

Se consideran las siguientes categorías:



### 3.7. Otras tablas auxiliares

#### 3.7.1. tls801\_country

Número de entradas: 236

Estado de la tabla: Disponible

Códigos de país.

#### 3.7.2. tls901\_techn\_field\_ipc

Número de entradas: 757

Estado de la tabla: Disponible

Relaciones existentes entre campos técnicos (tabla tls230) y las categorías IPC.

#### 3.7.3. tls902\_ipc\_nace2

Número de entradas: 850

Estado de la tabla: Disponible

Relaciones existentes entre códigos NACE2 (tabla tls229) y las categorías IPC.

#### 3.7.4. tls904\_nuts

Número de entradas: 1.484

Estado de la tabla: Disponible

Códigos de regiones europeas.

1. Prefiero usar incorrectamente “aplicación” porque es más cercano a la denominación en inglés que se emplea en los campos de la base de datos. [↑](#footnote-ref-1)