Universitat Oberta de Catalunya

Patricia Lázaro Tello
plazarotello@uoc.edu

PRA1: Análisis y diseño del *data warehouse*

Patricia Lázaro Tello

Análisis de los requerimientos

El sistema debe dar respuesta a las siguientes preguntas:

- ¿Cuántos residuos, y de qué tipo, genera cada país cada año?
- ¿Cuánto han reducido el consumo energético de combustibles fósiles cada país cada año?
- ¿Cómo es el balance de la transición energética (consumo de energía renovable propia vs. consumo de combustibles fósiles) en cada país en cada año?
- ¿Cuánto cuida cada país sus áreas terrestres y marinas protegidas?, es decir, ¿cuál es la evolución de estas áreas en cada país a lo largo del tiempo?
- ¿Cuánto invierte cada Comunidad Autónoma en protección ambiental cada año?
- ¿Cuánto contribuye cada país a los Objetivos de Desarrollo Sostenible a lo largo del tiempo?

Estas preguntas pueden ser desglosadas en dimensiones y hechos:

1. Hechos
   - Residuos: cantidad
   - Consumo energético: cantidad
   - Áreas protegidas (terrestres y marinas): extensión (km^2)
   - Inversión en protección ambiental: cantidad
2. Dimensiones
   - Localización: países y C.C.A.A.s
   - Información temporal: años
   - Objetivos de Desarrollo Sostenible
   - Residuos: tipo
   - Consumo energético: origen y tipo
   - Áreas protegidas: tipo
   - Inversión en protección ambiental: tipo y ámbito

Y podrá aportar información a los siguientes clientes:

- Organizaciones ecologistas
- Población general
- Administraciones de C.C.A.A.s
- Administraciones de Estados

Patricia Lázaro Tello
plazarotello@uoc.edu

Análisis de las fuentes de datos

A continuación se procede a revisar las fuentes de datos proporcionadas, especificando qué información contienen y en qué formato. Se seleccionarán también los datos que deben ser cargados.

- **02002.xls**: inversión (en €) en protección ambiental por tipo de equipo e instalación, ámbitos medioambiental y sector de actividad económica, agrupado por comunidad autónoma.

Nombre	Tipo	Ejemplo
Periodo	Texto	2015
Sector	Texto	C. Industria manufacturera
Tipo	Texto	INVERSIÓN EN EQUIPOS E INSTALACIONES INDEPENDIENTES
Ámbito	Texto	Protección del aire y el clima
Andalucía	Texto	6390920
Aragón	Texto	0
Asturias, Principado de	Texto	..
Balears, Illes	Texto	1096
Canarias	Texto	692088
Cantabria	Texto	26000
Castilla y León	Texto	1632806
Castilla-La Mancha	Texto	3220619
Cataluña	Texto	4480883
Comunitat Valenciana	Texto	12605792
Extremadura	Texto	5854
Galicia	Texto	970966
Madrid, Comunidad de	Texto	560660
Murcia, Región de	Texto	1364692
Navarra, Comunidad Foral de	Texto	587085
País Vasco	Texto	800628
Rioja, La	Texto	0
Total nacional	Texto	42490008

  - Registros: 154
  - Atributos: 22
  - Solo hay un valor para el atributo *Sector*

Patricia Lázaro Tello
plazarotello@uoc.edu

- Los valores del atributo *Periodo* tienen espacios delante y/o detrás del año
- Existe el valor ".." entre los valores de los atributos de las C.C.A.A.s.

- **Countries.json**: nombres en español, inglés y francés de los países, junto con la abreviatura (código ISO 3166-1-alpha-2 y alpha-3) y el prefijo de los teléfonos.

  Tiene un único atributo "*countries*" de tipo *array*, que contiene una lista de objetos:

Nombre	Tipo	Ejemplo
nombre	Texto	Albania
name	Texto	Germany
nom	Texto	Albanie
iso2	Texto	DE
iso3	Texto	DEU
phone_code	Texto	355

  - Registros: 246
  - Atributos: 6
  - El atributo *iso2* es una cadena de texto de 2 caracteres.
  - El atributo *iso3* es una cadena de texto de 3 caracteres.
  - Los valores del atributo *phone_code* son de tipo texto y contienen números.

- **DataGeneric.xml**: datos de la generación, método de disposición y tratamiento de residuos municipales (incluidos los de las unidades familiares).

Nombre	Tipo	Ejemplo
COU	Texto	AUS
VAR	Texto	MUNICIPAL
TIME_FORMAT	Texto	P1Y
UNIT	Texto	TONNE
POWERCODE	Numérico	0
TIME	Numérico	1992
OBSVALUE	Numérico	0.004
OBS_STATUS	Texto	E

  - Registros: 985
  - Atributos: 8
  - El atributo *COU* hace referencia al código ISO-3166-1-alpha-3 de los países.

También existen valores que hacen referencia a conjuntos, como "*OECD*".

- El atributo *VAR* hace referencia al método u origen de disposición, tratamiento o generación de los residuos.
- El atributo *TIME_FORMAT* representa el formato de tiempo; todos los datos se encuentran en el formato "P1Y", que equivale a años (*1992*).
- El atributo *UNIT* representa las unidades de *OBSVALUE*.
- El atributo *OBS_STATUS* hace referencia a las características de la observación. Un valor de "*E*" representa un valor estimado, "*B*" representa un salto, "*I*" representa datos incompletos.
- El atributo *OBSVALUE* separa los decimales con '.'.
- El atributo *POWERCODE* representa la potencia de 10 por la que hay que multiplicar *OBSVALUE* para que esté en las unidades estándares (Kgs, etc). Por ejemplo, $Tons \times 10^3 = Kgs$.

- **env_bio.tsv**: km^2 de área terrestre y marina protegida por cada estado miembro de la Unión Europea, durante los años 2011-2019. Para el área terrestre protegida incluye el %. Se incluye también el área total del estado miembro.

Nombre	Tipo	Ejemplo
areaprot	Texto	AREA_KM2
geo\time	Texto	AT
2019	Texto	83944
2018	Texto	0
2017	Texto	762860 s
2016	Texto	11
2015	Texto	30667
2014	Texto	0
2013	Texto	43167
2012	Texto	315
2011	Texto	:

- Registros: 120
- Atributos: 11
- El atributo *areaprot* define la semántica de los datos del registro (si son km^2 de área terrestre protegida, km^2 de área marina protegida, % de área total del estado miembro o km^2 de área total).

- Aunque el atributo se llama *geo\time*, hace referencia al código ISO-3166-alpha-2 de los países, salvo los valores especiales **EU28** y **EU27_2020**, que hacen referencia a los 27 o 28 países miembros de la UE.
- En los atributos *2011-2019* aparecen valores especiales:
  - *":"* → dato no disponible
  - *Número seguido de "s"* → medida estimada, no real

- **ODS.xlsx**: definición de los Objetivos de Desarrollo Sostenible y relación entre las acciones individuales y el Objetivo principal al que afecta.

Tabla: **ODS**

Nombre	Tipo	Ejemplo
Objetivo	Numérico	3
Nombre	Texto	Hambre cero
Descripción	Texto	Reducir la desigualdad en y entre los países

Tabla: **Ambito_VAR_Flow-ODS**

Nombre	Tipo	Ejemplo
Código	Texto	BULKY
Ambito/VAR/Flow	Texto	Bulky waste
ODS principal	Numérico	13

- Tabla **ODS**:
  - Registros: 17
  - Atributos: 3
- Tabla **Ambito_VAR_Flow-ODS**:
  - Registros: 50
  - Atributos: 3

- **WorldEnergyBalancesHighlights_final.xlsx**: balance energético por la Agencia Internacional de Energía (IEA). Incluye un desglose por años (*1971-2019*), por país, por producto y por flujo.

Nombre	Tipo	Ejemplo
Country	Texto	Australia

Nombre	Tipo	Ejemplo
Product	Texto	Heat
Flow	Texto	Industry (ktoe)
1971	Texto	14
2019 Provisional	Texto	..

- Registros: 6048
- Atributos: 55
- Los valores del atributo *Country* son nombres completos de países y agrupaciones de países (como "*World*" o "*Non-OECD Total*").
- Los valores del atributo *Flow* incluyen las unidades entre paréntesis
- Hay un atributo para cada año, desde 1971 a 2019 (que se llama *2019 Provisional*). No se han añadido todos los atributos para evitar una tabla muy larga.
- Los valores de los atributos de años pueden ser números positivos o negativos (-19937), con y sin decimales (-1858205,464). El valor también puede ser "*c*" para hacer referencia a datos confidenciales, o ".." para los datos no disponibles.

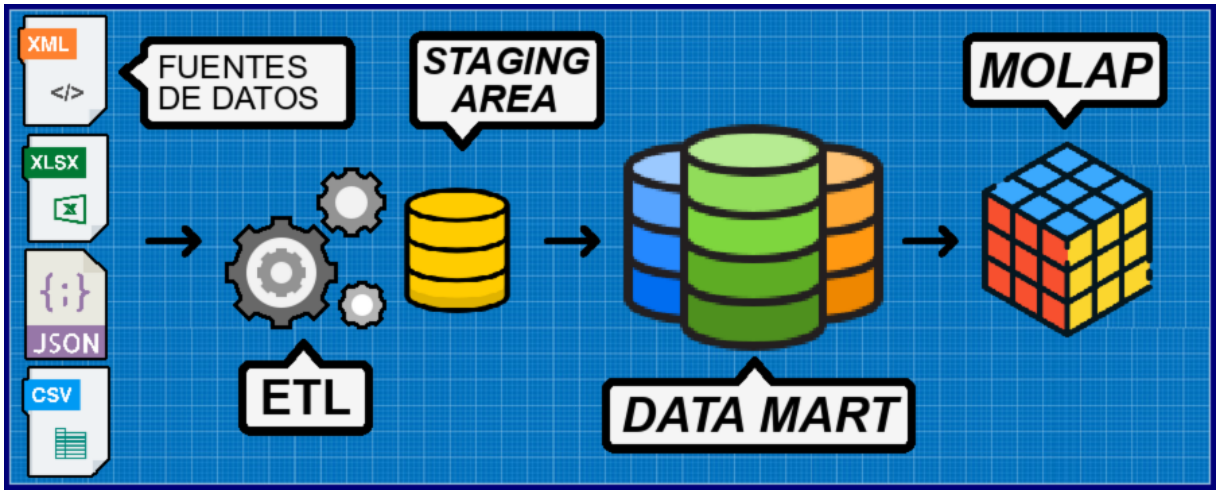
Volumen estimado

Archivo	Atributos	Registros	Datos
02002.xlsx	22	154	3388
Countries.json	6	246	1476
DataGeneric.xml	8	985	7880
env_bio1.tsv	11	120	1320
ODS.xlsx	$3+3$	$17+50$	201
WorldEnergyBalancesHighlights_final.xlsx	55	6048	332640
TOTAL	108	7620	346905

Análisis funcional

#	Requerimiento	Prioridad	Exigible/ Deseable
1	Se extraerá de forma adecuada la información de las fuentes de datos	1	E
3	Se creará un almacén de datos para el análisis del impacto ambiental y el consumo energético derivados de la actividad económica	1	E
4	Se cargará la información en el almacén de datos	1	E
5	Se creará un modelo OLAP para consultas automáticas de usuarios	2	E
6	Se documentará el proceso de carga inicial y el proceso ETL	3	D
7	Se automatizará el proceso de carga en el almacén de datos	4	D
8	Se creará un proceso de carga de datos incremental	4	D

En cuanto a la arquitectura de la factoría de información, se puede resumir en la siguiente figura:



- *Staging area* o área intermedia: al tener múltiples fuentes de datos, se recomienda utilizar un área intermedia para simplificar el proceso ETL de la Factoría de

Información.

- *Data mart* del impacto ambiental y consumo energético derivados de la actividad económica: se utiliza un *data mart* en lugar de un almacén de datos corporativo por estar centrado en una única temática.
- *MOLAP*: a partir de los datos del *data mart* se creará un cubo multidimensional para automatizar las consultas de los usuarios.

Patricia Lázaro Tello
plazarotello@uoc.edu

Diseño del modelo conceptual

A continuación se detallan los hechos (*facts*) con sus medidas y las dimensiones (*dimensions*) que se presentaron en **Análisis de los requerimientos**.

Tabla de hechos	Descripción
FACT_Residuos	Residuos generados y tratados
FACT_AreasProtegidas	Extensión de las Áreas Protegidas
FACT_ConsumoEnergetico	Balance energético desglosado, incluidas importaciones y exportaciones de energía
FACT_InversionProteccionAmbiental	Inversiones en Protección Ambiental

FACT_Residuos

Métrica	Descripción
cantidad	Cantidad de residuo generado y/o tratado

Dimensión	Descripción
localización	País en el que se ha generado/tratado el residuo
tiempo	Año en que se ha generado/tratado el residuo
ods	Objetivo de Desarrollo Sostenible al que contribuye la disposición del residuo
tipo_residuo	Tipo de residuo (con sus unidades asociadas)

FACT_AreasProtegidas

Métrica	Descripción
cantidad	Extensión del área protegida

Dimensión	Descripción
localización	País en el que se encuentra el área protegida
tiempo	Año en que se ha medido la extensión del área protegida
ods	Objetivo de Desarrollo Sostenible al que contribuye la preservación del área protegida
tipo_area	Tipo de área protegida (con sus unidades asociadas)

FACT_ConsumoEnergetico

Métrica	Descripción
cantidad	Cantidad de energía consumida

FACT_ConsumoEnergetico

Dimensión	Descripción
localización	País en el que se ha consumido la energía
tiempo	Año en que se ha consumido la energía
ods	Objetivo de Desarrollo Sostenible al que contribuye la energía consumida
origen_energia	Origen de la energía consumida
tipo_energia	Tipo de energía consumida (con sus unidades asociadas)

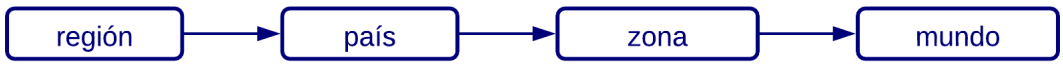
FACT_InversionProteccionAmbiental

Métrica	Descripción
cantidad	Cantidad de dinero (€) que se ha invertido

Dimensión	Descripción
localización	C.C.A.A./región o país que ha llevado a cabo la inversión
tiempo	Año en que se ha realizado la inversión
ods	Objetivo de Desarrollo Sostenible al que contribuye la inversión
tipo_inversion	Tipo de protección ambiental en que se ha invertido
ambito_inversion	Ámbito de protección ambiental en que se ha invertido

Algunas consideraciones a tener en cuenta respecto a las dimensiones, sus atributos y sus jerarquías de agregación:

- Localización: su nivel atómico es *Región* (equiparable a Comunidad Autónoma).



- Objetivos de Desarrollo Sostenible: posee un identificador numérico y una descripción textual del objetivo.
- “tipo_residuo”: tipo de residuo y las unidades en que se mide.
- “tipo_area”: tipo de área protegida y las unidades en que se mide.
- “tipo_energia” tipo de energía y las unidades en que se mide.

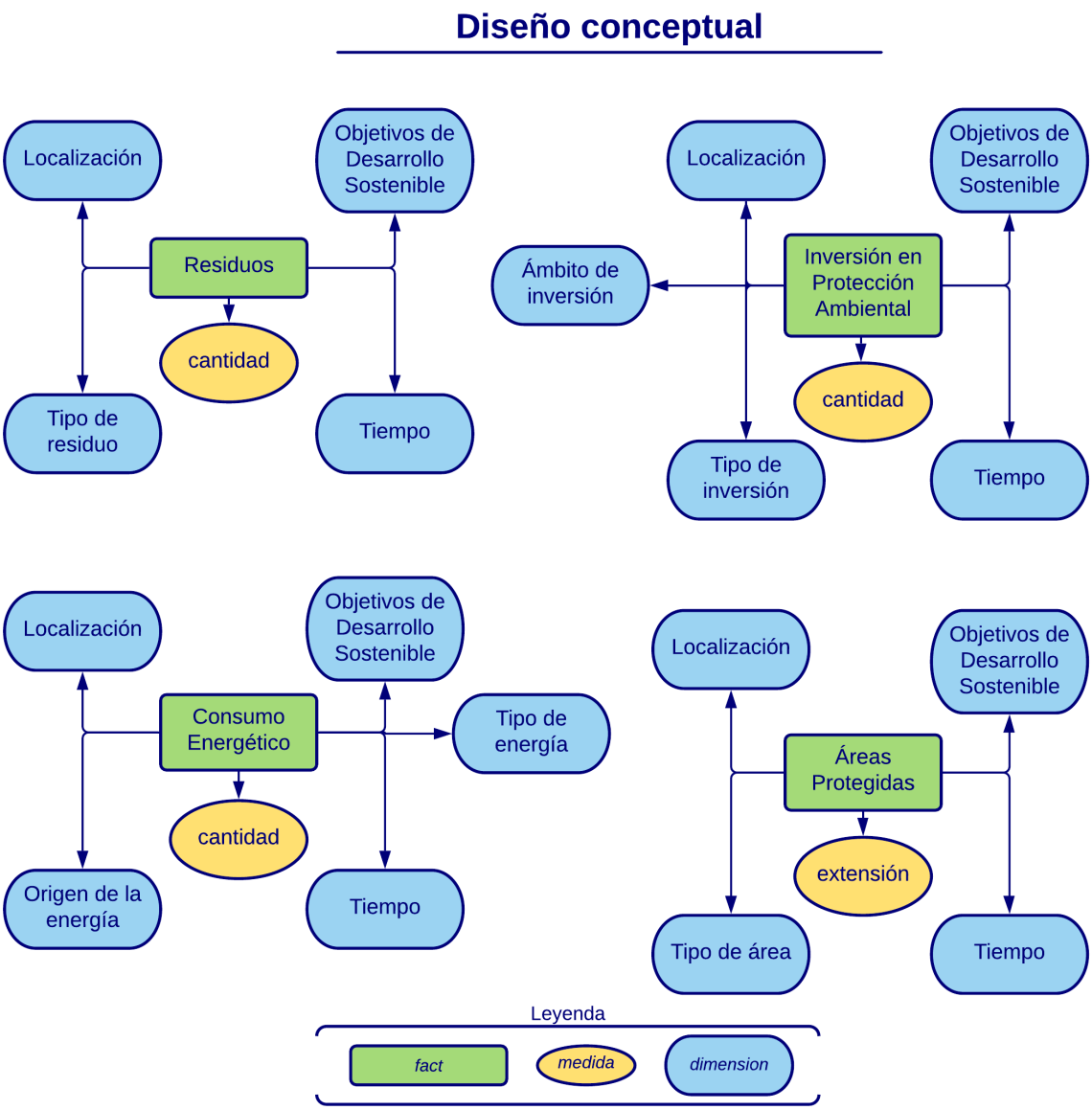


Figura 1: Diseño conceptual final del almacén de datos de impacto ambiental y consumo energético derivados de la actividad económica

Patricia Lázaro Tello
plazarotello@uoc.edu

Diseño del modelo lógico

A continuación se detallan las métricas y atributos de cada una de las tablas de hechos, y los atributos de cada una de las dimensiones. A su vez, se muestra la relación entre los hechos y las dimensiones.

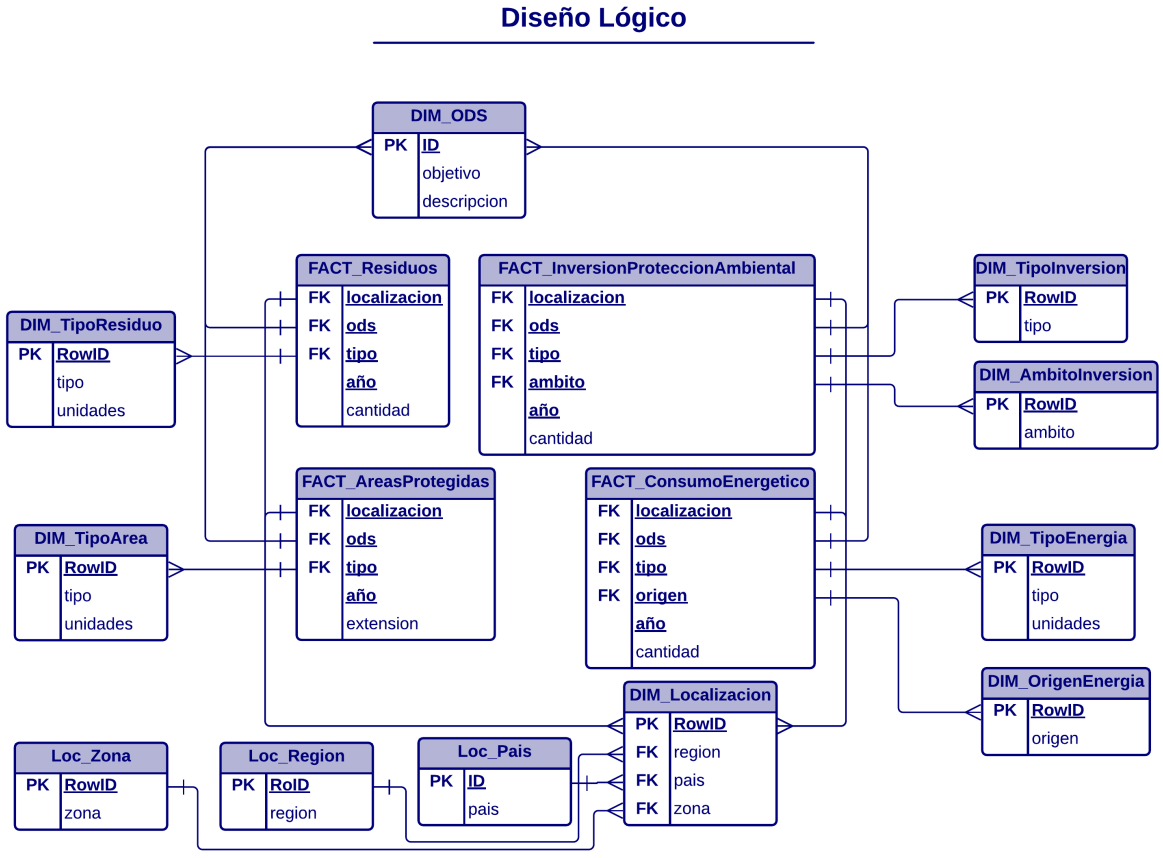


Figura 2: Diseño lógico final del almacén de datos de impacto ambiental y consumo energético derivados de la actividad económica

La dimensión **tiempo** hace referencia únicamente al año en que se tomó la medición; al no tener jerarquía de agregación ni atributos, se considera una **dimensión degenerada** y se añade como un atributo a cada tabla de hechos en la que participa, sin crearse una tabla de dimensión.

En la dimensión **Objetivos de Desarrollo Sostenible** existen atributos que solo se utilizarán para generar informes: *objetivo* y *descripción*. Estos atributos se han trasla-

dado a una **dimensión sombra**.

La creación de la dimensión sombra para los objetivos de desarrollo sostenible supone que la nueva dimensión reducida se considere una dimensión degenerada (solo tiene un atributo, el identificador del objetivo).

Por tanto, se ha utilizado el identificador del objetivo como clave en las tablas de hechos en lugar de utilizar un identificador numérico autogenerado (como puede ser *RowID*).

Existen otras dimensiones que podrían considerarse degeneradas, como **Tipo de Inversión**, **Ámbito de Inversión** y **Origen de Energía**, ya que solo tienen un identificador autogenerado y un atributo.

Sin embargo, se ha optado por mantener una tabla para cada dimensión dado que estos atributos serán cadenas de caracteres y ocuparán más espacio si se añaden como atributos en las tablas de hechos.

Respecto a la dimensión **Localización**, se han trasladado a **dimensiones sombra** los nombres de los países, regiones y zonas, y se han utilizado en la tabla de dimensión los códigos ISO-3166-1-alpha-2 de los países y un identificador numérico autogenerado para región y zona. Así se consigue reducir espacio en disco.

Patricia Lázaro Tello
plazarotello@uoc.edu

Diseño del modelo físico

A continuación se muestra para cada tabla definida en el diseño lógico sus campos, incluyendo el tipo y la longitud de los mismos, así como si es clave foránea y si forma parte de la clave primaria de la tabla.

DIM_ODS

Campo	Tipo	FK	PK
id	número (2 dígitos)		X
objetivo	cadena (50 caracteres)		
descripcion	cadena (500 caracteres)		

DIM_TipoResiduo

Campo	Tipo	FK	PK
id	número (4 dígitos)		X
tipo	cadena (150 caracteres)		
unidades	cadena (25 caracteres)		

DIM_TipoArea

Campo	Tipo	FK	PK
id	número (4 dígitos)		X
tipo	cadena (25 caracteres)		
unidades	cadena (5 caracteres)		

DIM_TipoEnergia

Campo	Tipo	FK	PK
id	número (4 dígitos)		X
tipo	cadena (50 caracteres)		
unidades	cadena (5 caracteres)		

DIM_OrigenEnergia

Campo	Tipo	FK	PK
id	número (4 dígitos)		X
origen	cadena (100 caracteres)		

Patricia Lázaro Tello
plazarotello@uoc.edu

DIM_TipoInversion

Campo	Tipo	FK	PK
id	número (4 dígitos)		X
tipo	cadena (100 caracteres)		

DIM_AmbitoInversion

Campo	Tipo	FK	PK
id	número (4 dígitos)		X
ambito	cadena (100 caracteres)		

DIM_Localizacion

Campo	Tipo	FK	PK
id	número (4 dígitos)		X
region	número (6 dígitos)	X	
pais	cadena (2 caracteres)	X	
zona	número (3 dígitos)	X	

Loc_Region

Campo	Tipo	FK	PK
id	número (6 dígitos)		X
region	cadena (100 caracteres)		

Loc_Pais

Campo	Tipo	FK	PK
id	cadena (2 caracteres)		X
pais	cadena (100 caracteres)		

Loc_Zona

Campo	Tipo	FK	PK
id	número (3 dígitos)		X
zona	cadena (100 caracteres)		

FACT_Residuos

Campo	Tipo	FK	PK
id_localizacion	número (4 dígitos)	X	X
id_ods	número (2 dígitos)	X	X
id_tipo	número (4 dígitos)	X	X
anyo	número (4 dígitos)		X
cantidad	número (3 dígitos, 3 decimales)		

FACT_AreasProtegidas

Campo	Tipo	FK	PK
id_localizacion	número (4 dígitos)	X	X
id_ods	número (2 dígitos)	X	X
id_tipo	número (4 dígitos)	X	X
anyo	número (4 dígitos)		X
extension	número (10 dígitos)		

FACT_InversionProteccionAmbiental

Campo	Tipo	FK	PK
id_localizacion	número (4 dígitos)	X	X
id_ods	número (2 dígitos)	X	X
id_tipo	número (4 dígitos)	X	X
id_ambito	número (4 dígitos)	X	X
anyo	número (4 dígitos)		X
cantidad	número (10 dígitos)		

FACT_ConsumoEnergetico

Campo	Tipo	FK	PK
id_localizacion	número (4 dígitos)	X	X
id_ods	número (2 dígitos)	X	X
id_tipo	número (4 dígitos)	X	X
id_origen	número (4 dígitos)	X	X
anyo	número (4 dígitos)		X
cantidad	número (10 dígitos)		