Universidad Zaragoza

Trabajo Fin de Máster
Máster en Ingeniería Informática

DESARROLLO DE UN ALMACÉN DE DATOS DE PUBLICACIONES CIENTIFICAS CON VISUALIZACIÓN DE DATOS E INFORMES

-

DEVELOPMENT OF A DATA WAREHOUSE OF ACADEMIC PUBLICATIONS WITH DATA VISUALIZATION AND REPORTING

Autor

Jorge Eliecer Higuera Muñoz

Director

Sergio Ilarri Artigas

ESCUELA DE INGENIERÍA Y ARQUITECTURA

2017

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD

TRABAJOS DE FIN DE GRADO / FIN DE MÁSTER

Escuela de Ingeniería y Arquitectura Universidad Zaragoza

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD

(Este documento debe acompañar al Trabajo Fin de Grado (TFG)/Trabajo Fin de Máster (TFM) cuando sea depositado para su evaluación).

D./Dª. Jorge Eliecer Higuera Muñoz,

con nº de DNI Y3726255Q en aplicación de lo dispuesto en el art. 14 (Derechos de autor) del Acuerdo de 11 de septiembre de 2014, del Consejo de Gobierno, por el que se aprueba el Reglamento de los TFG y TFM de la Universidad de Zaragoza,

Declaro que el presente Trabajo de Fin de (Grado/Máster) Máster, (Título del Trabajo)

Desarrollo de un almacén de datos de publicaciones científicas con visualización de datos e informes.

Development of a data warehouse of academic publications with data visualization and reporting.

es de mi autoría y es original, no habiéndose utilizado fuente sin ser citada debidamente.

Zaragoza, 01 Septiembre de 2017.

Fdo:

Agradecimientos

Quiero agradecer en primer lugar a la vida por darme mi familia que siempre está en mis pensamientos y la misma que día a día me dan palabras de ánimo y de fuerza para conseguir el objetivo que sea; mi familia que, sin verla, la siento a mi lado cada vez que respiro. Gracias a la vida por las oportunidades que he tenido y las que he intentado aprovechar siempre.

Quiero agradecer a España por recibirme y hacerme feliz durante todo este tiempo, por enseñarme a vivir alejado de mi familia y por mostrarme que existen otras familias y sobre todo por hacer más fuertes mis fortalezas y cada día tratar de mejorar mis debilidades.

Gracias a la universidad por admitirme, por dejarme ser parte del máster y por enseñarme además de lo académico, la importancia del cariño al estudio y a la dedicación, aunque las cosas no siempre nos parezcan agradables.

Mil y un millón más de gracias a INYCOM por apoyarme, por darme la mano, el brazo, "todo" al momento de subir mi primer escalón laboral en España, gracias por confiar en mí, gracias por todo el tiempo, por todas mis ausencias, por mis mil trámites y gracias por darme el mejor equipo de trabajo que nunca he tenido hasta ahora (BI-MAZ). Estoy seguro que he compensado todo lo que he recibido con mi trabajo, mi dedicación.

Gracias a mi director de proyecto Sergio Ilarri por compartir conmigo sus conocimientos, por dejar en este trabajo un pequeño trozo de su vida y por enseñarme que todo puede ser mejor. Además, quiero agradecerle por su paciencia y pedir perdón por mis mil errores de ortografía, gramática, redacción, estilo… que a pesar de la corrección dos mil quinientos, mantenía una sonrisa amable.

¡Gracias a mis compañeros, a todos los profesores, a las señoras de la cafetería que siempre me dicen "majo", a los administrativos, a todos… y a los que están leyendo esto… también gracias!

Resumen

DESARROLLO DE UN ALMACÉN DE DATOS DE PUBLICACIONES CIENTIFICAS CON VISUALIZACIÓN DE DATOS E INFORMES

Se plantea como trabajo de fin de master: Diseñar e implementar un almacén de datos que permita evaluar el rendimiento científico de investigadores. Y así mismo poder analizar algunas métricas de los lugares de publicación de estos artículos.

Lo que se ha buscado desarrollar, es un proyecto que permita analizar de manera sencilla los indicadores de los artículos, autores y lugares de publicación de acuerdo a ciertas métricas establecidas. La herramienta está dirigida a usuarios convencionales que podrán manejar los *dashboards* desarrollados y también para usuarios expertos que podrán realizar cualquier tipo de consulta-análisis directamente en la base de datos y también podrán crear nuevos paneles de visualización de información.

Lo que hace especial esta herramienta y diferente a las que ya existen, es que permite con solo una descarga de información de diferentes fuentes, integrar los datos de dichas fuentes y mostrar los indicadores en un mismo lugar. Se han utilizado técnicas de emparejamiento (*matching*) exacto, reemplazo de palabras clave y confianza-similitud basadas en la herramienta *Data Quality Services* de Microsoft. Teniendo en cuenta que es una herramienta en la cual se ha basado el emparejamiento, es posible que, si se utilizara esta o alguna otra herramienta especializada, se puedan obtener mejores resultados. Sin embargo, se ha implementado una carga de ficheros de configuración donde el usuario puede hacer emparejamiento manual. Por ejemplo, si la herramienta muestra dos nombres de autores diferentes, pero se sabe que es el mismo autor, pero escrito de forma diferente (Sergio Ilarri artigas – Ilarri, A. Sergio) se puede rellenar un fichero de configuración de nombres para que un proceso unifique estos nombres. Igualmente, con los nombres de los artículos y los nombres de los lugares de publicación. Se puede resumir el desarrollo de este proyecto en cinco fases principales:

Análisis de las fuentes de información: Se ha hecho un esfuerzo por seleccionar las fuentes de información más importantes y a su vez las más adecuadas para el proyecto teniendo en cuenta la información que nos otorga cada una en los ficheros de exportación.

Descarga de los ficheros de información: Ficheros de artículos y lugares de publicación.

Diseño del almacén de datos: Modelo estrella bajo la metodología de Kimball.

Proceso ETL: Carga de las tablas de *Staging*, auxiliares y finalmente las estrellas de Impacto Artículo e Impacto Lugar de Publicación.

Visualización de información: Los informes presentados a los usuarios.

ÍNDICE

I. Introducción

1.1 Contexto

En la actualidad se habla de que el activo más importante de las empresas no son las adquisiciones físicas o monetarias, sino la información y el conocimiento. “El recurso "conocimiento" es la clave para el éxito sostenible de empresas y economías. Es la empresa inteligente la que va sobrevivir en la competencia global.” [1].

Actualmente no se encuentra ninguna herramienta desarrollada a la medida para poder consultar, analizar y comparar métricas de las publicaciones científicas y los lugares de publicación. Una herramienta que permita explorar esta información integrando los datos de todas las fuentes de información de manera amigable y visual basado en las principales fuentes de datos de publicaciones científicas y artículos, capaz de unificar todas las fuentes de información en un solo almacén de datos. Este sistema está orientado a dos tipos de usuarios: los usuarios convencionales que interactúan con los paneles de visualización desarrollados y los usuarios expertos que disponen de acceso directo a la base de datos para hacer consultas y también pueden implementar nuevos paneles de visualización. El proyecto está basado en un almacén de datos estrella (*Data Warehouse*) optimizado, diseñado e implementado bajo las características más importantes utilizadas hoy día para el desarrollo de almacenes de datos siguiendo la metodología de Kimball.

Según los ficheros de prueba descargados para el proyecto, nos encontramos en un momento en que se publican un gran número de artículos científicos en el ámbito de ciencias de la información, creciendo en los últimos años. La anterior afirmación está basada en los resultados mostrados por la herramienta luego de su desarrollo, esto se puede apreciar en la figura 1

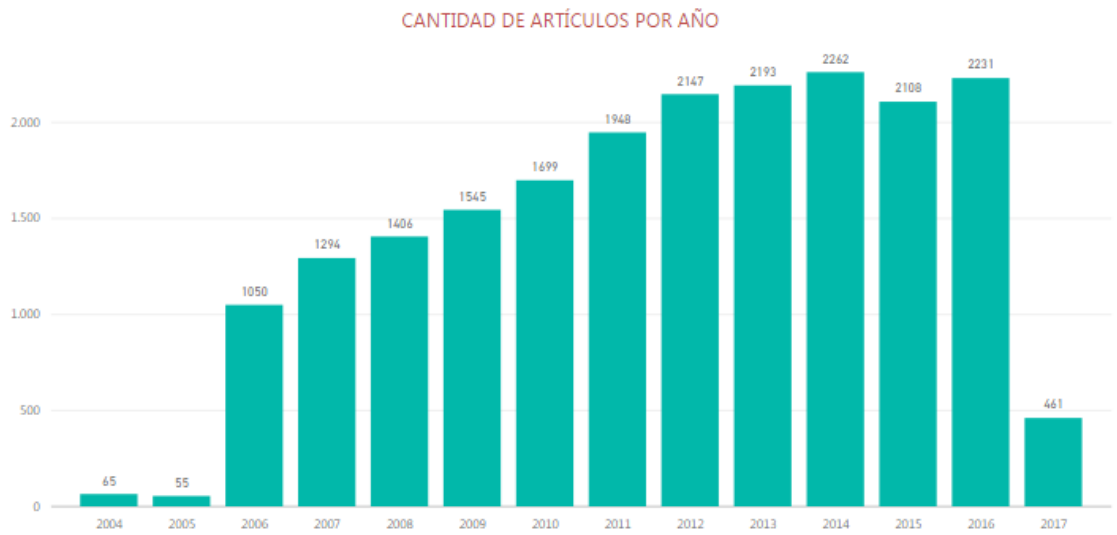


Figura 1: Cantidad de artículos por año

Este trabajo fin de master para su desarrollo ha tenido en cuenta publicaciones científicas del ámbito tecnológico. No obstante, de las mismas fuentes de información se pueden extraer ficheros de información de cualquier ámbito académico permitiendo el análisis por autores, lugares de publicación y artículos de dichos ámbitos, tal y como está definido en el objetivo principal del proyecto.

Se puede considerar este trabajo de fin de máster como un sistema de gestión del conocimiento en el cual existen tres conceptos básicos que deben estar perfectamente diferenciados para garantizar la solidez y coherencia del mismo: Información (fuentes de información científica: información de publicaciones y fuentes de información de los lugares de publicación), Conocimiento (el proceso de Extracción, Transformación y Carga que convierte los datos de los ficheros planos en información) y Representación del Conocimiento (visualización y análisis del conocimiento) [2]. Se debe hacer uso de estos tres conceptos bajo una metodología que así lo permita, para así mismo poder arrojar un resultado esperado en el desarrollo de este trabajo de fin de máster.

1.2 Formulación del problema

¿Cómo a partir de un diseño e implementación de un almacén de datos de publicaciones científicas y a través de una herramienta de análisis y visualización de información, se puede analizar y/o medir los indicadores de los autores, artículos y lugares de publicación?

1.3 Motivación

En el contexto de la gestión de datos, se afirma que el activo más importante de una empresa o en cualquier contexto donde se plantee el análisis de datos, es la información. Con base en lo anterior se busca desarrollar un sistema que pueda extraer, relacionar y enfocar toda la información y conocimientos de publicaciones científicas y explotarlos. Se busca demás, gestionar de tal forma la información que pueda ser utilizada eficientemente en el análisis de las métricas e indicadores planteados e incluso podría ser una herramienta para la ayuda de toma de decisiones. Entre las principales características de este trabajo fin de master están:

- Si bien se encuentran herramientas para el análisis de publicaciones científicas como SeGeDa [3], Kampal [3], Publish or Perish [4], esta herramienta desarrollada como trabajo de fin de master, permite unificar la información y los datos de todas las fuentes de datos planteadas:

  - Para Publicaciones: Scopus [6], Google Scholar [7], Web of Science [8], Publish or Perish [5]. Science Direct [9]
  - Para lugares de publicación: CORE Conference Ranking [10], CORE Journal Ranking [11], JCR Ranking [12], SJR Journal Ranking [13], GGS Ranking [14]

  Para almacenarla en un mismo almacén de datos y visualizarla a partir de este.

- El diseño del almacén de datos hace que las consultas sean rápidas puesto que se ha eliminado la redundancia en las tablas y la integridad de la información está garantizada. Lo anterior gracias a la ETL que se encarga de cargar y validar toda la información que se carga en las tablas del proyecto.
- El proyecto soporta cualquier tipo de consulta directamente a la base de datos si el usuario es experto, asimismo la herramienta de *Power BI* facilita la creación de nuevos paneles de visualización.
- Para los usuarios convencionales, es una herramienta *óptima* puesto que se han desarrollado paneles de visualización con las métricas más importantes que se utilizan para medir las publicaciones científicas.
- El sistema cuenta con ficheros de configuración donde el usuario puede hacer ajustes en la información que no pudo ser solucionada por las ETL´s. Estos son cargados de forma automática por la herramienta y asimismo la información es actualizada.
- Este proyecto de fin de máster puede ser utilizado para diferentes propósitos académicos. Es una herramienta útil para un investigador para hacer seguimiento de sus indicadores. Presta asesoría a un departamento de investigación para hacer memorias de investigación. Es una herramienta donde una universidad puede hacer seguimiento al desempeño de sus investigadores. Es una fuente de datos donde se pueden ver publicaciones de temas específicos y apoyar la creación de nuevas publicaciones.

"… en una economía cada vez más global, la innovación, la tecnología y los activos intangibles en general tienen una importancia fundamental para mantener la competitividad." [3] es por eso la importancia de que el recurso de conocimiento pueda ser explotado de una mejor manera y ser utilizado en pro del desarrollo, inversión, investigación, etc.

"El análisis de las propiedades de conocimiento nos van a servir para complementar el concepto de conocimiento y para estudiar sus implicaciones estratégicas." [4]

De acuerdo con lo anterior se justifica la necesidad de desarrollar e implementar el almacén de datos de publicaciones científicas, su carga automática y finalmente el diseño y desarrollo de la visualización de la información como un apoyo a cualquier persona o entidad del ámbito académico o científico. Este desarrollo permite hacer un análisis de los indicadores de las publicaciones científicas propias o ajenas de diferentes fuentes de información.

1.4 Objetivo del proyecto

El objetivo general de este Trabajo de Fin de Máster es diseñar e implementar un almacén de datos que permita evaluar el rendimiento científico de investigadores (publicaciones con sus citas, factor de impacto, etc.). Diseñar e implementar paneles (dashboards) de visualización de datos de interés.

Finalmente, se plantea diseñar e implementar la generación de informes.

Para ello se hará toma y análisis de los requerimientos y el potencial interés de los usuarios que podrían hacer uso del sistema. Asimismo, se analizarán qué fuentes de datos que se utilizaron en el proyecto y se unificarán en un mismo almacén de datos para su explotación.

1.5 Objetivos específicos

- Realizar la toma de requerimientos para el proyecto y análisis de los mismos; priorización de tareas a realizar.
- Analizar de las fuentes de datos:

- Analizar la extracción de cada fichero descargado de las fuentes de información.
- Analizar y elegir el formato de descarga de acuerdo a la estructura del fichero y la información disponible de cada fuente.

- Definir las fuentes de información que se utilizarán en el proyecto.
- De acuerdo a la información disponible de las fuentes de publicaciones científicas y los lugares de publicación, diseñar e implementar el almacén de datos.

- Implementar conceptos científicos para el diseño e implementación del almacén de datos para la optimización de los recursos y reducir posibles costes de tiempo y ejecución del sistema en general siguiendo la metodología de Kimball.

- Diseñar e implementar los procesos ETL (Extracción-Transformación y Carga) para la carga de datos.

- Diseñar e implementar los procesos ETL para la carga de las tablas de staging.
- Diseñar e implementar los procesos ETL y los scripts necesarios (procedimientos almacenados, cursores, seguimiento) para la carga de las tablas auxiliares.
- Diseñar e implementar los procesos ETL y los scripts necesarios (procedimientos almacenados, cursores, seguimiento) para la carga de las tablas del almacén de datos (DATAWAREHOUSE); tablas de dimensiones y hechos.

- Diseñar e implementar posibles vistas y/o esquemas con cálculos generales para hacer más liviano el proceso de ejecución del proceso.
- Diseñar e implementar los paneles-hojas de visualización de datos.

- Panel de información de Autores.
- Panel información Artículos.
- Panel de información de Lugares de Publicación.

- Validar la consistencia de la información mostrada.

1.7 Marco referencial y teórico

Para el desarrollo e implementación del proyecto del almacén de datos de publicaciones científicas, como proyecto para optar al grado de Máster en Ingeniería en Informática, ha sido necesario aplicar los conocimientos adquiridos durante el master, particularmente los relacionados con las asignaturas de “Administración y dirección estratégica de empresas” y “Gestión de la innovación en tecnologías de la información”, porque es en estas asignaturas donde se aprende a planear, desarrollar, gestionar y documentar un proyecto de software de acuerdo a parámetros y especificaciones dadas las características del mismo. También se han aplicado los conocimientos adquiridos en la asignatura de “Manipulación y análisis de grandes volúmenes de datos”, especialmente esta asignatura puesto que en el desarrollo del proyecto se ha contemplado prácticamente el total del contenido de la misma.

La gestión de proyectos es una parte fundamental de la Ingeniería de Software; un proyecto bien gestionado tiene una alta probabilidad de éxito, pero es claro que una mala gestión es sin duda el camino al fracaso de un proyecto de software debido a tardanzas de entregas, planeación errónea de costos, falta de personal y muchos factores más que se tienen que gestionar para el éxito de un proyecto de software en general.

1.8 Marco tecnológico

En este capítulo se describen las principales herramientas y tecnologías empleadas para el desarrollo del proyecto. Se pueden diferenciar en dos grupos: Herramientas principales para el desarrollo del proyecto y herramientas de apoyo.

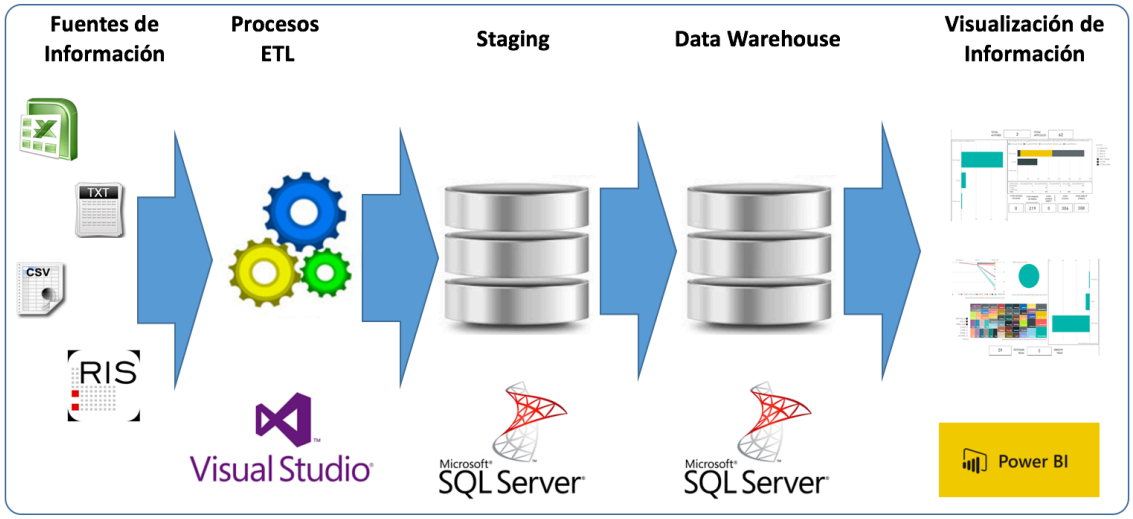


Figura 2: Principales herramientas para el desarrollo del proyecto

Figura 3: Herramientas de apoyo para el proyecto

1.8.1 Herramientas de desarrollo

- *SQL Server:* Es el gestor y herramienta de Base de datos. Es allí donde se alojan todas las tablas, modelos del sistema y procedimientos almacenados de la base de datos.
- *SSIS (SQL Server Integration Services)*: Entorno de desarrollo de los Procesos ETL (Extracción, Transformación y Carga)
- *Visual Studio*: *Framework .Net* para desarrollo de los scripts utilizados en los paquetes de los procesos ETL y procedimientos de Base de Datos
- *NotePad++*: Editor de texto. Empleado como auxiliar para el planteamiento de consultas y desarrollo de scripts.

1.8.2 Herramientas de visualización

- *Power BI*: Herramienta de visualización y transformación de datos de la suite de Microsoft.
- *SSRS (SQL Server Reporting Services)*: Herramienta de Visualización y generación de informes de la suite de Microsoft.

1.8.3 Herramientas de gestión del proyecto / Control de versiones / Documentación

- *GIT Lab*: Repositorio del proyecto.
- *Microsoft Office*: suite ofimática. Empleada para la escritura del documento académico entregable del proyecto, realización de diagramas y tablas mostradas en el documento.

1.9 Estructura de la memoria

De acuerdo a las especificaciones general para la entrega de proyectos de fin de master, este documento está dividido en cinco capítulos principales:

- Introducción: Planteamiento del problema, objetivo general, objetivos específicos, metodología de desarrollo, descripción de herramientas y tecnologías.
- Desarrollo del Proyecto: De acuerdo a los requerimientos del desarrollo, se deben cumplir los objetivos descritos, así como el planteamiento de la arquitectura, el diseño y la implementación.
- Gestión del Proyecto: Se menciona el cómo se ha abordado el desarrollo del proyecto, la metodología utilizada y la planificación y seguimiento.
- Conclusiones: Se escriben los resultados del trabajo de fin de máster, dificultades técnicas o académicas que pudieron surgir durante el desarrollo del proyecto y lecciones aprendidas.
- Bibliografía: El listado de las fuentes citadas en el proyecto.
- Anexos: Es necesario incorporar una sección de anexos donde se pueden consultar documentos técnicos, manuales de usuario y demás documentación para ampliar la información descrita durante el documento. En total son 13:

  A. Definición de las métricas.
  B. Manuales de descarga de información.
  C. Manual configuración Instancia base de datos.
  D. Manual de Restauración de las bases de datos.
  E. Manual de configuración de rutas del proceso.
  F. Script carga estrella Impacto Artículo.
  G. Script carga estrella Lugar de Publicación.
  H. Configuración DQS (*Data Quality Services*) similitud y confianza.

I. Propiedades Parametrización Búsqueda Aproximada DQS.

II. Contexto Tecnológico

En este capítulo se analiza la tecnología empleada para el desarrollo del trabajo de fin de master en cada una de las etapas de almacenamiento, procesos ETL y visualización, y adicionalmente se citan los cuadrantes mágicos de Gartner para respaldar la elección de las herramientas. Se analizan también los conceptos teóricos que definen un almacén de datos, la definición del modelo escogido y la metodología que se ha seguido para su desarrollo. Además, se estudia el proceso de la minería de datos y se plantean posibles aplicaciones para el proyecto.

2.1 Almacenes de datos

Los almacenes de datos son colecciones de datos orientadas a un dominio, integradas, variables en el tiempo y no son volátiles [5]. Generalmente son útiles para ayudar a la toma de decisiones.

Los objetivos de los almacenes de datos son:

- Facilitar el acceso a la información.
- Presentar información consistente.
- Seguridad en el almacenamiento de los datos.
- Ser adaptable y resistente al cambio.
- Soporte para la adaptación incremental.
- Apoyo para la toma de decisiones.

El desarrollo del almacén de datos y almacenamiento se ha realizado con SQL Server – *Management Studio* de Microsoft. Según el cuadrante mágico de Gartner para el año 2017, Microsoft es un líder de acuerdo al almacenamiento, accesibilidad, procesado de los datos y soporte analítico.

Figura 4: Cuadrante mágico para herramientas de solución de gestión de datos para Analitics

Fuente: Gartner (2017)

Microsoft trabaja constantemente para que el procesamiento y el análisis de datos sean más sencillos y accesibles. Esto lo hacen a través de su herramienta SQL Server donde ofrece un gran número de soluciones y servicios para el almacenamiento de datos, incluso grandes volumnes de datos y soluciones avanzadas de análisis.

SQL Server ofrece a las organizaciones una amplia gama de capacidades para estas puedan hacer grandes análisis de forma local o en la nube, donde incluye analítica de datos en tiempo real y almacenamiento en memoria. Ofrece también integraciñon con Hadoop a través de PolyBase y análisis con bases de datos R. Todo en su herramienta de almacenamiento de datos, el cual utilizamos para este trabajo de fin de master.

2.1.1 Tecnologías para almacenes de datos

Las jerarquías, las dimensiones y los cubos son la base para el procesamiento analítico en línea (OLAP), Al representar toda la información de estas maneras, se presenta al usuario una manera intuitiva de navegar en un conjunto complejo de datos. No obstante, ofrecer al usuario una manera intuitiva de consultar los datos, puede no ser suficiente si lo que se quiere es obtener una gran velocidad en las consultas.

Uno de los objetivos más importantes del OLAP es ofrecer al usuario tiempos de respuesta óptimos en las búsquedas de información, información consistente y en algunos casos que los agregados de información sean calculados anticipadamente. Estos valores pre-calculados hacen parte del buen desempeño del OLAP.

Inicialmente se consideraba que la única solución para una aplicación OLAP era un modelo no relacional. Después, otras compañías descubrieron que las estructuras de bases de datos (modelos en estrellas y copos de nieve), los almacenamientos de agregados y los índices se podían utilizar también en estas aplicaciones.

Así que la tecnología OLAP relacional fue llamada ROLAP y la tecnología OLAP multidimensional fue llamada MOLAP. Normalmente la tecnología MOLAP tiene mejor desempeño que la tecnología ROLAP en cuanto a las consultas y tiempos de respuesta, pero no es tan escalable como la ROLAP.

MOLAP

La arquitectura MOLAP utiliza bases de datos multidimensionales para dar soluciones de análisis a los usuarios. Su mayor premisa es que considera que la mejor manera de almacenar los daros es de forma multidimensional para ser analizada y visualizada en varias dimensiones de análisis.

Esta tecnología utiliza una arquitectura de dos niveles: el motor analítico y las bases de datos multidimensionales que son las encargadas de la administración, acceso y disponibilidad del dato.

La arquitectura MOLAP lee datos pre-compilados, es decir, para hacer lectura del modelo antes se debe hacer una compilación. Esta arquitectura también tiene limitaciones para calcular agregados de datos que no han sido calculados y almacenados previamente.

ROLAP

La arquitectura ROLAP accede directamente al almacén de datos (*Data Warehouse*) para dar soluciones de análisis OLAP. Sus premisas y ventajas son:

- Ofrece mejores prestaciones haciendo consultas en bases de datos relaciones.
- Provee herramientas que pueden garantizar la no existencia de duplicados.
- Puede garantizar integridad referencial. Por ser un modelo más fácil de leer.
- Facilita el proceso de normalización.

La clave es que utiliza el modelo relacional en estrella y un CUBO relacional.

La base de datos relacional se encarga del almacenamiento de los datos, y el motor ROLAP se encarga de dar la funcionalidad analítica. Este motor utiliza bases de datos relaciones para la administración, acceso y disponibilidad del dato.

La arquitectura ROLAP es capaz de calcular ágilmente agregados de los datos o utilizar datos pre-calculados si existen. Se accede directamente al almacén de datos y soporta técnicas para mejorar prestaciones, como pueden ser la partición de la base de datos, soporte a modelos no normalizados y múltiples *joins*.

HOLAP

HOLAP es la arquitectura más reciente. Es una arquitectura híbrida entre la arquitectura ROLAP y la arquitectura ROLAP, para brindar soluciones con las mejores características de cada una: el desempeño superior de la MOLAP y la fácil escalabilidad de la ROLAP.

Un tipo de HOLAP mantiene los registros de detalle de la información en arquitectura ROLAP, y los cálculos de los agregados en una arquitectura MOLAP a parte.

2.1.2 Modelo de datos

El modelo de datos es una estructura abstracta que almacena y documenta la información de uno o varios procesos de negocios dados. Un modelo de datos es diseñado y desarrollado con el objetivo de almacenar la información de una manera acorde al negocio, adicionalmente para mejorar la comunicación y la interpretación de la información entre las diferentes plataformas de una empresa y sus usuarios a través de interfaces de usuario.

Según la ANSI (Instituto Nacional Estadounidense de Estándares), un modelo de datos podría interpretarse como un esquema lógico (descripción semántica de tablas y columnas), conceptual (contiene las reglas de los datos definidas en los requerimientos) y físico (el almacenamiento de información como tal).

Modelo en estrella como subconjunto del modelo relacional.

El modelo estrella es el modelo elegido para el diseño y desarrollo del trabajo de fin de master. Los modelos en estrella son estructuras dimensionales desarrolladas para una base de datos relacional "*Data Warehouse*".

Este modelo está compuesto por las tablas de hechos que contienen las métricas o medidas de un evento específico y las tablas de dimensiones que definen el proceso de negocio; las tablas de dimensiones son las que tienen las claves primarias, y las tablas de hechos tienen las claves foráneas que apuntan a las primarias de las dimensiones.

Lo que hace característico este modelo es que las tablas de hechos están relacionadas con claves foráneas con las tablas de dimensiones, es esta relación la que da forma de estrella al modelo.

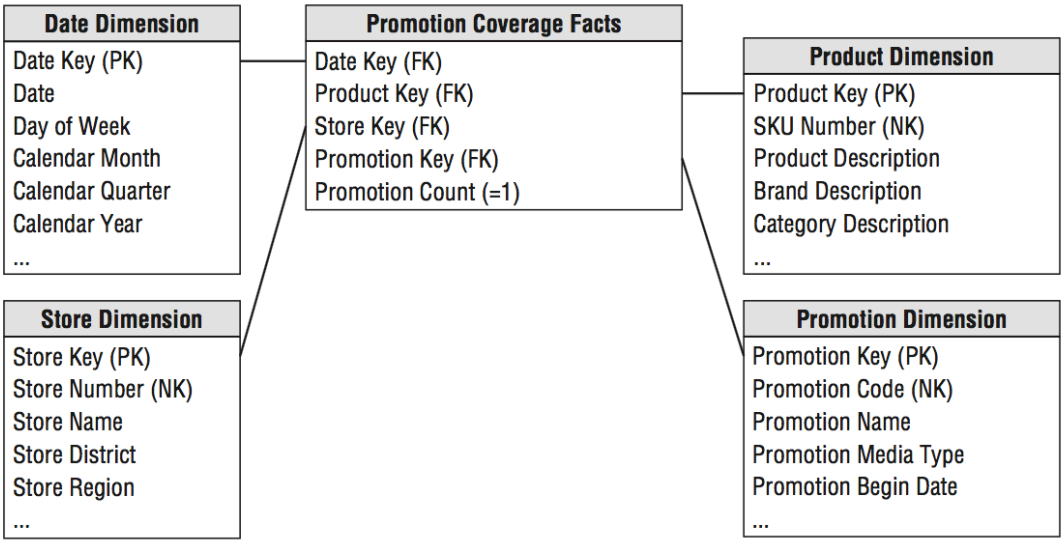


Figura 5: Ejemplo del modelo en estrella. *Promotion Coverage* [6]

Para almacén de datos del trabajo de fin de máster siendo un modelo en estrella. Se han utilizado diversos elementos de dicho modelo, como se describen a continuación.

Tablas de hechos (fact table)

Son las tablas primarias en un almacén de datos multidimensional, almacenan medidas numéricas del proceso de negocio. Un hecho es denominado por un indicador del negocio. Por ejemplo, la cantidad de citas de una publicación científica. Normalmente la tabla de hechos es la tabla central de un modelo en estrella. Las tablas de hechos contienen las llaves foráneas de las dimensiones que definen su nivel de detalle.

Para este proyecto de fin de master se han plateado dos tablas de hechos:

1. Impacto_Articulo: Hace referencia a toda la información y métricas de las publicaciones científicas (artículos). Se puede consultar el detalle de esta tabla de hechos en el apartado de anexos.
2. Impacto_lugar_Publicacion: Hace referencia a toda la información y métricas de los lugares de publicación. Esta tabla de hechos está segmentada en dos y puede consultarse el detalle en el apartado de anexos:
    a. Impacto_Revista: Las métricas de las revistas.
    b. Impacto_Congreso: Las métricas de los congresos.

    Adicionalmente hay tablas para los indicadores CORE que no tienen definida una política clara para mostrar el indicador, como por ejemplo las métricas de SJR que se deben mostrar las del año del artículo o la última disponible. Para CORE se mostrarán las métricas de todos los años y el usuario podrá filtrar el año que desee ver.

Tablas de dimensión (Dimension table)

Son las tablas que contienen los descriptores del negocio, a diferencia de las tablas de hechos, estas suelen tener pocas tuplas, estas, generalmente se utilizan para hacer filtros de información a la hora de hacer consultas. Por ejemplo, filtrar la información por un autor en específico.

Las tablas de dimensiones más representativas del proyecto son:

1. Dim_Autor: Contiene la información de todos los autores de las publicaciones científicas de las fuentes descargadas.
2. Dim_Datos_articulo: Contiene la información adicional de un artículo que no va en la tabla de hechos. Por ejemplo, título, volumen, paginas, etc.
3. Nombre_Lugar_Publicacion: Contiene todos los nombres de los lugares de publicación de las publicaciones científicas de las fuentes de datos descargadas.
4. Dim_Datos_CORE: Contiene toda la información de las métricas CORE.

La definición de estas y todas las tablas de dimensiones puede ser consultada en el apartado de anexos.

Tablas puente (Bridge table)

Estas tablas puente son utilizadas para dimensiones que toman múltiples valores. Forman grupos para un conjunto de valores que se relacionan con un único id con la tabla de hechos. Son utilizadas para las relaciones de muchos a muchos que no pueden ser "plasmadas" en una tabla de hechos. Por ejemplo, el problema de tener muchos autores para una publicación se soluciona con tablas puente, creando la tabla de "Grupo_Autores" donde su id hace referencia a uno o más autores (los que hacen parte de una publicación).

2.2 Metodología Kimball

La metodología de Raph Kimball se basa en lo que denomina ciclo de vida del negocio (*Business Dimensional Lifecycle)* [6]. Enfocado en cuatro principios:

1. Centrarse en el negocio: Enfocarse en los requerimientos del negocio que se va a analizar identificándolo claramente, hacer un esfuerzo en desarrollar relaciones sólidas con el mismo.
2. Construir la infraestructura de información: diseñar y construir la base de datos especializada en el negocio, integrada, entendible por los usuarios y diseñarla de tal forma que ofrezca grandes prestaciones al momento de gestionar los datos y que cumpla claramente con los requerimientos del negocio.
3. Realizar entregas incrementales: desarrollar el almacén de datos de manera incremental, con entregas periódicas; tener en cuenta la importancia de los requerimientos del negocio para asimismo planificar el orden de las entregas-incrementos.

4. Hacer entrega de la solución completa: Consiste en generar el desarrollo total de los requerimientos, tener preparado también todo el entorno, y el almacén de datos, consistente, accesible y en perfecta ejecución. En este principio se tienen en cuenta las herramientas de explotación, *reporting*, publicación, documentación, etc.

La construcción del almacén de datos puede llegar a ser una tarea muy compleja, puesto que se deben tener en cuenta todas las partes del negocio y toda la información que interactúa en el proceso. Esta metodología, al estar enfocada solo a un proceso del negocio hace que pueda ser más fácil de segmentar el "todo" y por ende más fácil su análisis, diseño y desarrollo.

La metodología de Kimball, además, contempla cuatro pasos para el proceso de diseño del almacén de datos [6]:

1. Seleccionar el proceso de negocio: El proceso de negocio son las actividades operacionales de la organización. Un proceso de negocio podría ser el registro de matrículas en una universidad, donde cada registro generado de este proceso es trasladado a la tabla de hechos. Es importante la elección del proceso de negocio puesto que se deben cubrir los requerimientos dados, teniendo en cuenta las métricas, dimensiones y hechos.
2. Declarar el grano: La definición del grano es un paso esencial puesto que define el nivel de detalle de los hechos. Si el grano es pequeño, el nivel de detalle de la información será mayor, lo contrarío si el grano es grueso. El grano debe ser definido antes de definir las dimensiones puesto que estas deben ser consistentes al grano.
3. Identificar las dimensiones: Las dimensiones son las que nos dan la información de cómo, cuándo, el qué, por qué, dónde; es decir, las dimensiones proporcionan la información misma del proceso de negocio. Generalmente las tablas de dimensiones cuentan con una gran cantidad de columnas y sus cambios generalmente son lentos.
4. Identificar los hechos. En las tablas de hechos se encuentran las medidas, lo que se está realmente midiendo. Estas métricas son valores numéricos, para poder ser agregados, y van relacionados "uno a uno" con el tamaño del grano.

2.3 Procesos ETL

Los procesos ETL '*Extract, Transform and Load*' (Extracción, transformación y carga) son los que se encargan de recolectar los datos de las diferentes fuentes de datos (archivos planos, documentos de texto, bases de datos, etc.) y hacer algún tipo de transformación (agregados, limpieza, organización, etc.), para finalmente ser cargados en una base de datos para ser explotados o para ser fuente de un nuevo proceso.

Implícitamente, el proceso de integración de datos, tienen una serie de reglas de negocio, o requisitos sobre los datos que se extraen para convertirlos en datos que serán cargados. Algunas fuentes requieren manipulación en la estructura del dato. Por ejemplo, el tipo de dato, el tamaño, etc. No obstante, las transformaciones más comunes aplicadas en la integración de datos son:

- Convertir datos en códigos. Por ejemplo, 0 que podría definir falso y 1 que podría definir verdadero.
- Hacer reemplazos generales. Por ejemplo, la sigla cll. Podría reemplazarse por la palabra “calle”.
- Seleccionar un número determinado de columnas de la fuente de información. En muchos casos hay una fuente con más columnas de las que realmente se necesitan, así que sólo se seleccionan algunas.
- Hacer operaciones. Por ejemplo, sumar el total de citas que tiene un autor en las diferentes fuentes de información.
- Trasponer o pivotar los datos de la fuente de información.
- Dividir una columna en partes. Por ejemplo, si en una columna viene el nombre con sus apellidos, el proceso ETL podría dividir esa columna en tres: nombre, apellido1 y apellido2.

En el mercado existen varias herramientas para la integración de datos. Entre las más importantes están:

- Cognos Decisionstream – IBM.
- Data Integrator - herramienta de Sap Business Objects – SAP.
- IBM Websphere DataStage antes Ascential DataStage – IBM.
- Kettle (ahora llamado Pentaho Data Integration) – Pentaho.
- Integration Services – Microsoft.

La herramienta elegida para el desarrollo del trabajo de fin de máster es *Integration Services (SSIS)* de Microsoft. Se ha elegido esta herramienta porque con ella se pueden cubrir todos los requerimientos del trabajo con respeto a la transformación e integración de datos. Es una de las herramientas más importantes del mercado y contiene un variado conjunto de herramientas para la creación de paquetes ETL que van desde la carga de archivos planos hasta el envío de correos automáticos dando información a los usuarios. Adicionalmente se ha elegido esta herramienta porque se contaba con experiencia de trabajo en la misma.

Se ha consultado el cuadrante mágico de Gartner sobre herramientas de integración de datos del 2017 para corroborar que elegir *Integration Services* es una buena opción.

Figura 6: Cuadrante mágico para herramientas de integración de datos

Fuente: Gartner (2017)

Como se observa en la figura, la herramienta de Microsoft es una de las mejores con respecto a la integración de datos; según la evaluación, los proveedores líderes de integración de datos deben gestionar el tipo de dato que sea, ya sea estructurado o no estructurado, ofrecer trasmisión de datos en horarios programados o en tiempo real sin problemas, ofrecer implantaciones en entorno local, en la nube o híbridas y, como es de esperar, deben entregar datos fiables. Todo lo atenrior es ofrecido por la herramienta elegida para el trabajo de fin de master.

2.4 Analítica y BI

Analítica y BI (*Bussines Intelligence*) se describen como las técnicas y herramientas para el análisis de información. Existen dos categorías: herramientas de análisis y herramientas de *reporting*; el objetivo final es brindar al usuario información para que éste pueda analizar, y con respecto a esta tomar algún tipo de decisión. El BI es el análisis de información para la ayuda de la toma estratégica de decisiones: consiste en extraer datos de diferentes fuentes de información, analizarlas y agregarlas para finalmente mostrar dicha información en visualizaciones de datos.

Al igual que las herramientas de integración de datos, también existen herramientas representativas para la analítica y el BI:

- Tableau.
- Qlink.
- Pentaho.
- Sales force.
- Power BI y Reporting Services de Microsoft.

Para el desarrollo del proyecto de fin de máster, se ha elegido la herramienta de Microsoft Power BI, puesto que es una herramienta relativamente nueva de Microsoft, en que muchas empresas están apostando y que cada vez tiene más seguidores. Facilita todo el proceso de la carga de información y además ofrece un módulo de transformación de datos directamente en la herramienta. Este módulo es utilizado en el trabajo de fin de máster. Por ejemplo, al momento de cálculos entre las columnas cargadas del almacén de datos directamente desde *Power BI*. Las visualizaciones que ofrece Power BI son bastante dinámicas y lo más importante es que son muy amigables con el usuario siendo su manejo muy intuitivo.

Al investigar el informe de Gartner del 2017 para herramientas de analítica y BI, se aprecia que Microsoft se sitúa por segundo año consecutivo en la cabeza de líderes para herramientas de analítica e inteligencia de negocio con su herramienta Power BI. Así que la elección de la herramienta para el desarrollo del trabajo de fin de máster ha sido un acierto.

Figura 7: Cuadrante mágico para plataformas de analítica e inteligencia de negocio

Fuente: Gartner (2017)

El informe apunta a que las herramientas de inteligencia de negocio sean cada vez más fáciles para los usuarios y que éstas dependan cada día menos del departamento de tecnología, además de incorporar cada vez un mayor número de funcionalidades. Por ejemplo, aporta funcionalidades básicas para el usuario como podría ser exportar un informe en PDF.

Gartner considera que las empresas de hoy deben apostar en sus proyectos por herramientas de analítica modernas con el fin de aprovechar la innovación del mercado además que este tipo de herramientas permiten que el flujo de trabajo sea mucho más ágil y preciso.

III. Trabajo realizado

3.1 Requisitos – Alcance y Limitaciones

Los requisitos del proyecto están directamente sincronizados con el cumplimiento de los objetivos específicos.

En resumen, se debe diseñar y desarrollar un sistema que permita el análisis de las métricas de publicaciones científicas teniendo en cuenta los autores, artículo y lugares de publicación. Esto consolidado en el análisis de las fuentes de datos, desarrollo de los procesos ETL, explotación de información y visualización.

3.1.1 Alcance

- El sistema estará soportado en un modelo de Base de Datos relacional en estrella para la adecuada gestión de los datos de las publicaciones científicas y lugares de publicación.
- El sistema para su puesta en marcha tendrá una base de información de pruebas, descargada de los ambientes reales de extracción de información planteados en el proyecto.
- De acuerdo a los manuales de usuario, el usuario puede descargar la información que desee analizar de acuerdo a sus áreas de interés y ponerlo en las rutas de carga de información y el sistema automáticamente cargará la información y armará la estructura de información.
- El sistema será una herramienta de análisis de información de publicaciones científica y la información mostrada, las métricas, autores, lugares de publicación y demás depende de los datos cargados y no contempla una alimentación automática directamente de alguna fuente en especial.
- No está dentro del alcance del presente proyecto desarrollar un sistema experto en gestión de conocimiento y/o análisis de información. Tampoco se pretende desarrollar una herramienta de visualización de información científica, puesto que en el mercado ya existen muchas que pueden dan cobertura a esos. El desarrollo de una nueva herramienta de visualización no proporcionaría un valor añadido al resultado del proyecto.

- Las herramientas seleccionadas para el desarrollo del proyecto se han elegido teniendo en cuenta sus características y la documentación disponible y teniendo en cuenta que la demanda de perfiles que manejen estas herramientas es alta de acuerdo a los portales de búsqueda de empleo. Es preciso aclarar que se buscan perfiles con conocimientos en un sinnúmero de herramientas, pero se quiere hacer énfasis en que poseer conocimiento y/o experiencia en las herramientas utilizadas hace que se puedan tener mayores oportunidades de vinculación laboral.

3.1.2 Limitaciones

- El funcionamiento y pruebas iniciales del sistema se han realizado con ficheros de carga que son "descargados" de las fuentes reales y con estructura tal cual la da la fuente de datos. Es necesario seguir los manuales de descarga, puesto que cada fuente de información tiene más de una opción de descarga.

- Las fuentes de información de las cuales se hace el análisis son limitadas a las más representativas con respecto a las publicaciones científicas. Es posible incluir nuevas fuentes si es necesario, ya que el modelo de base de datos y la carga de ETL está adaptado para que la incorporación de una nueva fuente no tenga mayor impacto en el desarrollo.

3.2 Arquitectura

El trabajo de fin de master puede describirse en cuatro partes:

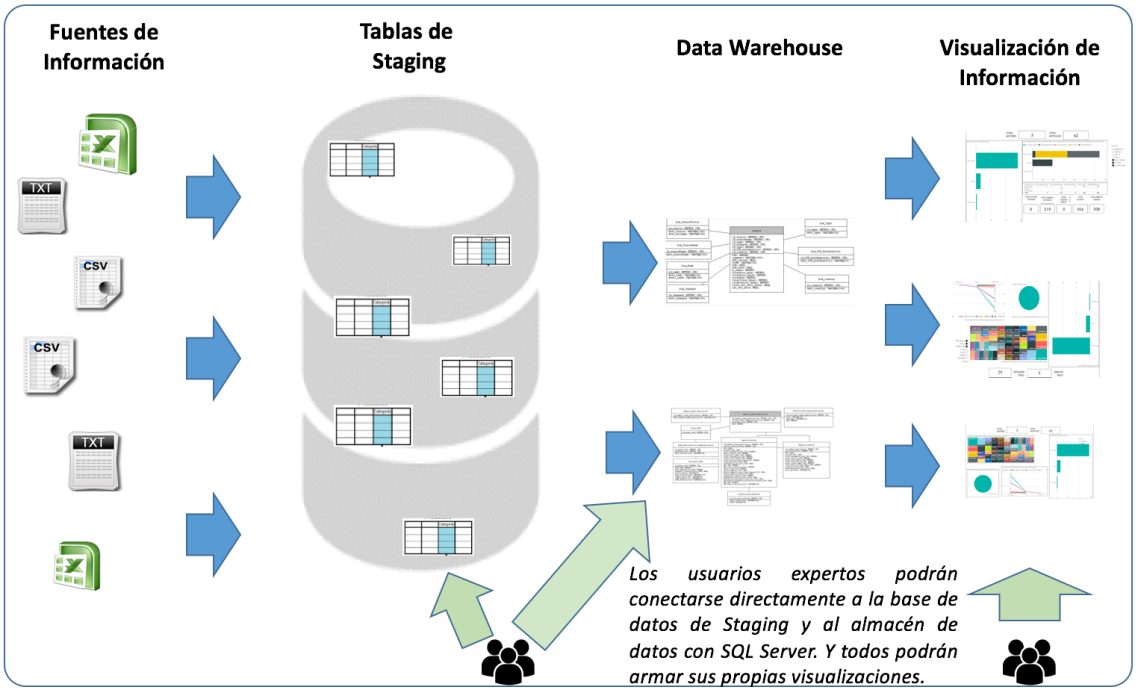


Figura 8: Arquitectura del proyecto

De acuerdo a la figura anterior los cinco segmentos planteados son los siguientes:

Fuentes de información: todos los ficheros de archivo plano que se utilizan como fuentes de datos, ficheros en formato .csv, .xlsx, .txt y. ris

Tablas Staging: son las tablas que son cargadas directamente de la lectura de los ficheros de información. En esta base de datos también se alojan las tablas auxiliares que son las mismas tablas de *staging* que tienen algún tipo de proceso para dejarlas preparadas (sin errores, sin caracteres extraños y sin datos incorrectos) para a partir de estas tablas alimentar el almacén de datos. Es posible crear accesos a la base de datos para usuarios que quisieran acceder a estas tablas.

Data Warehouse: Es el almacén de datos estructurado, formado por los dos modelos de estrella, el primero de Impacto Artículo y el segundo de Impacto Lugar de Publicación. Es posible crear accesos a la base de datos para usuarios que quisieran acceder a estas tablas.

Visualización de información: Son los paneles de información donde se muestran métricas de autores y lugares de publicación. Los informes presentados allí son solo una base de lo que el usuario puede hacer. La herramienta es muy intuitiva y permite al usuario hacer de forma sencilla diferentes visualizaciones de la información que desee ver.

Se contemplan usuarios que puedan acceder directamente a la base de datos para trabajar con las tablas. Estos usuarios sólo tendrían permisos de lectura. Podría contemplarse el caso de tener usuarios con el rol de administrador del proyecto para hacer modificaciones en la estructura del modelo o de los datos almacenados en el mismo.

3.3 Fases del desarrollo del proyecto

En este apartado se describirá todo el proceso que se ha realizado para el desarrollo del proyecto: las fuentes de datos analizadas (inclusive las que finalmente no se han contemplado para el desarrollo del proyecto), el diseño de la base de datos, el diseño y desarrollo de los procesos de ETL, las visualizaciones finales y, además, una sección que describe el problema del emparejamiento de datos.

3.3.1 Análisis y recolección de las fuentes de datos

Para el trabajo fin de máster, era necesario seleccionar las fuentes de datos tanto para los artículos, como las fuentes de datos de los lugares de publicación.

En general hubo complicaciones a la hora de cargar los ficheros descargados de las fuentes de datos por temas de que los ficheros traen errores de codificación, porque el tamaño de algún campo sea muy grande o fuera de lo normal. Estas características hacían que el proceso de carga fallara en algún momento. Otra complicación que surgió es que en los ficheros existen líneas que no respetan la estructura general del archivo. Por ejemplo, un fichero que esté delimitado por comas ‘,’ y tenga cuatro columnas: Nombre, año, país, comentario y hay líneas con estructura: “dato1”, “dato2” y directamente un salto de línea sin completar la estructura con comas ‘,’.

Una anécdota curiosa al momento de cargar uno de los ficheros de artículos es que existe un artículo con 222 autores. Inicialmente se ha considerado como un error, pero al revisar el fichero se vio que sí que estaba bien, así que se ha modificado el proceso de carga para aceptar ese tipo de valores tan grande.

Los errores en la carga se iban solucionando uno a uno a medida que iban apareciendo. En alguna ocasión se ha optado por cambiar el formato de exportación, puesto que era imposible cargar los datos.

3.3.2 Fuentes de análisis de artículos

En el apartado de anexos se encuentran los manuales para la descarga de información de cada una de las fuentes; para los ficheros de análisis de artículos se tuvieron en cuenta las siguientes fuentes:

Scopus

Scopus es una web que ofrece funciones de búsqueda de manera gratuita a usuarios que pueden o no ser estar inscritos en la plataforma. Permite, por ejemplo, buscar autores y ofrece métricas de autores y fuentes como por ejemplo el volumen del artículo, total de autores, lugar de publicación, página de inicio y final, entre otras.

Scopus es una de las fuentes elegidas para el proyecto, puesto que nos permite descargar una gran cantidad de registros de artículos en una sola descarga de acuerdo a parámetros como búsquedas por autor, título, etc. [8]

Adicionalmente, nos muestra las citas de los artículos.

Google Scholar

Google Scholar es también un buscador de documentos científicos y adicionalmente identifica las citas de los mismos [9]. Se considera como una fuente de información de publicaciones científicas que compite con Scopus o Web Of Science, que también están contempladas en este proyecto.

Al igual que Scopus, Google Scholar es una de las fuentes elegidas en el proyecto. No obstante, esta fuente de información no permite hacer una descarga de grandes conjuntos de registros: sólo permite descargar, por ejemplo, la información de un artículo de un autor.

Web of Science

Web of Science facilita el acceso a un conjunto de datos de investigaciones científicas a través de un servicio en línea suministrado por Thomson Reuters. Presenta información de citas, artículos y lugares de publicación, abarcando todos los tópicos del ámbito académico [10].

Esta fuente de información también se ha tenido en cuenta para el desarrollo del proyecto puesto que igual que Scopus permite descargar información de gran cantidad de registros. Para el proyecto se ha descargado la información haciendo la búsqueda por autor. Además de proporcionar información básica de los artículos, también provee información de la cantidad de citas por artículo.

Researcher ID

En principio la fuente de Researcher ID se tuvo en cuenta, pero al final no se ha implantado en el proyecto porque la exportación desde esta fuente de información sólo deja descargar registros de uno en uno (artículo por artículo), pero el problema más grande que se ha tenido es que esta fuente de información descarga ficheros en diferentes formatos y estructuras de acuerdo a la información que tuviera de cada publicación [11], esto era un problema puesto que no podíamos controlar todos los tipos de estructuras y la diferente información por cada descarga.

Publish Or Perish

Publish Or Perish es una plataforma que ejecuta búsquedas por autor, revistas, referencias concretas y sus citas. Permite hacer las búsquedas de acuerdo al objetivo de cada persona y además ordenar por diferentes criterios [12].

Los ficheros de descarga de Publish Or Perish se han tenido en cuenta para el desarrollo del proyecto. Al igual que Scopus, esta fuente permite extraer gran cantidad de registros. También descarga la información con el dato de las citas por artículo. Especialmente con el fichero que exporta no se han tenido problemas de carga puesto que es un archivo.cvs y en todos los registros mantiene la misma estructura de los encabezados.

Science Direct

Es una plataforma web que con acceso basado en suscripción (para el proyecto, se accede desde una IP de la Universidad de Zaragoza, estableciendo conexión sin problemas), aloja una gran base de datos de investigación científica y médica, con más de 12 millones de piezas de contenido de un total de 3.500 revistas académicas y 34.000 libros electrónicos [13].

La Fuente de datos Science Direct es la fuente con más particularidades con respecto a las anteriores puesto que el formato de descarga es RIS (*Research Information Systems*) [5] se muestra en la Figura 9.

```
TY  - JOUR
AU  - Baldwin,S.A.
AU  - Fugaccia,I.
AU  - Brown,D.R.
AU  - Brown,L.V.
AU  - Scheff,S.W.
TI  - Blood-brain barrier breach following cortical contusion in the rat
T2  - Journal of Neurosurgery
PY  - 1996
VL  - 85
IS  - 4
SP  - 476-481
SN  - 0022-3085
AB  - Adult Fisher 344 rats were subjected to a unilateral impact to the dorsal cortex above the hippocampus at 3.5
      m/sec with a 2 mm cortical depression. This caused severe cortical damage and neuronal loss in
      hippocampus subfields CAU, CA3 and hilus.
KW  - cortical contusion
KW  - blood-brain barrier
KW  - horseradish peroxidase
KW  - hippocampus
KW  - rat
DO  - DOI:10xxxxxxxx
ER  -
```

Figura 9: Ejemplo de un artículo en formato RIS

Así que una vez cargado el fichero, este es analizado por un cursor que se ha desarrollado para que recorra todos los registros y de acuerdo al encabezado de cada línea, lo inserte en la columna correspondiente de la tabla. Por ejemplo, los encabezados “AU – “son autores, el encabezado “T1 – “es el título.

3.3.3 Fuentes de análisis de lugares de publicación

Entre las fuentes de información de los lugares de publicación se tuvo especialmente en cuenta SeGeDa (DATAUZ Unizar) [16]. Es una aplicación para la analítica de datos para todo el personal de la Universidad de Zaragoza englobada en el proyecto DATAUZ (Datos Abiertos y Transparencia de la Universidad de Zaragoza).

Los objetivos más importantes de esta herramienta son:

- Facilitar a los responsables universitarios la toma estratégica de decisiones.
- Apoyo informativo a las unidades universitarias.
- Ser fuente de datos en los procesos de acreditación en requerimientos de datos, indicadores, métricas universitarias.
- Dar respuesta ágil a las preguntas de negocio que puedan surgir.

Toda la documentación e información relacionada, puede ser consultada una vez el usuario se identifique en su página web [16].

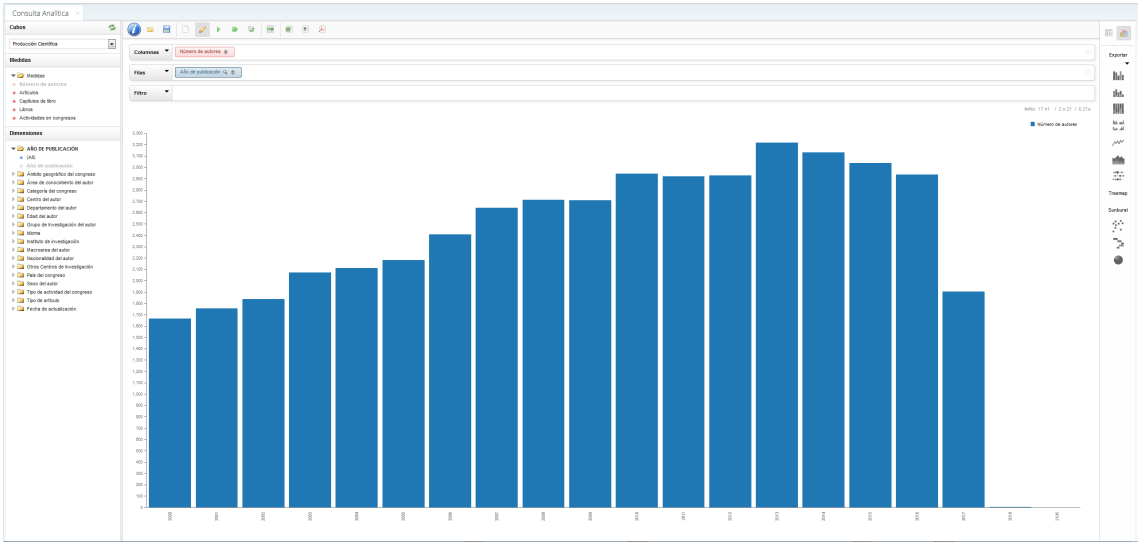


Figura 10: Ejemplo visualización en SeGeDa (Número de autores por año)

No fue posible utilizar esta fuente de información en el proyecto de trabajo de fin de master puesto que, si bien SeGeDa nos proporciona métricas que se pueden incorporar en el proyecto, la herramienta no permite obtener la información de los artículos y autores.

CORE Conference - Journal Ranking

La Asociación de Investigación y Educación en Computación de Australia (CORE), es una asociación de departamentos universitarios de informática en Australia y Nueva Zelanda. Antes de 2004 se conocía como la *Computer Science Association* (CSA).

Esta asociación ayuda y promueve la investigación en la informática y las tecnologías de la información en los centros de enseñanza superior y de investigación. Proporciona un foro de temas tecnológicos con el fin de estimular el debate de temas relacionamos y promover la cooperación con otras organizaciones. La descripción de las clasificaciones se puede consultar en el anexo de "Definición de las métricas".

- Conference Ranking

El Ranking CORE – Conferences, es una actividad continua que proporciona evaluaciones de las principales conferencias en las disciplinas de computación. Los rankings son administrados por el Comité Ejecutivo de CORE [10].

- Journals Ranking

Es una base de datos similar a la de conferencias, pero de revistas científicas. Ésta se ha iniciado con la lista de los indicadores ERA 2010 que también está incluida en el proyecto [11].

JCR Ranking

Journal Citation Reports (JCR), permite evaluar y comparar revistas usando datos de citas de aproximadamente 12,000 revistas científicas y técnicas y actas de conferencias de más de 3.300 editores en más de 60 países. JCR incluye prácticamente todas las especialidades en las áreas de ciencia, tecnología y ciencias sociales [17].

Los indicadores proporcionados por JCR son cargados en el proyecto, el usuario puede descargar la información de las revistas del área de interés o toda la información contenida en el portal. La herramienta consolida la información y la deja preparada para su visualización y/o análisis.

SJR Journal Ranking

SCImago *Journal Rank* (SJR) es un portal público que incluye las revistas e indicadores científicos a partir de la base de datos Scopus (también tenida en cuenta para la extracción de publicaciones científicas del proyecto) proporciona más de 21.500 títulos y más de 5.000 editoriales internacionales. Los indicadores mostrados en el portal se pueden utilizar para evaluar y analizar los dominios científicos [18]. En el proyecto son utilizados para evaluar el impacto de los lugares de publicación.

El portal permite exportar la información de las revistas de acuerdo a 27 áreas principales. Por ejemplo, Ingeniería y Computación. Para el proyecto, una vez descargado el fichero de información, se debe incluir en el nombre del mismo, el año y el área de conocimiento de acuerdo a como se especifica en el manual de usuario que se puede consultar en el apartado de anexos.

GGS Ranking

Es un portal también de información de métricas de los lugares de publicación de artículos científicos. El objetivo de este portal es desarrollar una clasificación unificada de las conferencias de las ciencias de la información. Es un comité conjunto de GII (*Group of Italian Professors of Computer Engineering),* GRIN (*Group of Italian Professors of Computer Science*) y SCIE (*Spanish Computer-Science Society*), que calcula la clasificación de acuerdo a un algoritmo automático basado en clasificaciones internacionales conocidas y existentes. Cada una de las tres sociedades del comité podrá someter cada una de las calificaciones generadas automáticamente para revisarla y corregirla. Esta calificación es actualizada en general cada dos años [19].

Las fuentes para calcular este ranking son CORE Conference, LiveSHINE y Microsoft Academic.

Mientras que el CORE Conference Ranking viene como un conjunto de lugares clasificados, LiveSHINE y Microsoft Academic simplemente reportan una serie de indicadores bibliométricos basados en citas sobre conferencias, especialmente el H-índex de la conferencia.

Los H-index se suelen considerar como indicadores robustos. Sin embargo, sufren un problema de dimensionalidad: es posible que las conferencias con un número muy elevado de artículos publicados tengan un H-index alto, independientemente de la calidad real de estos documentos. Lo contrario puede ocurrir para pequeñas conferencias que publican menos artículos.

Para reducir estas distorsiones, utilizando los datos disponibles en LiveSHINE y Microsoft Academic, se ha calculado un indicador secundario, denominado "citas medias", obtenido dividiendo el número total de citas recibidas por los documentos de la conferencia, entre el número total de documentos. Éste es un “lifetime impact factor” (FI) para la conferencia. Estos indicadores se basan en el promedio, y por lo tanto son sensibles a la presencia de valores atípicos. Esto sugiere que no deben usarse como indicadores primarios de clasificación. Sin embargo, pueden ayudar a corregir las distorsiones debidas a la naturaleza dimensional del índice H.

3.3.4 Diseño del almacén de datos

Para el diseño del almacén de datos se tuvo como principal requerimiento que sea un modelo ER (Entidad – Relación) y que sea un modelo en estrella diseñado siguiendo la metodología de Kimball [6]. Esta metodología se enfoca principalmente en el diseño de almacenes de datos y ha sido seleccionada porque da solución a un problema de negocio en un plazo corto de tiempo. Además, es una metodología muy versátil enfocada a los procesos de negocio. En el desarrollo de cada una de sus fases se da solución al objetivo planteado en el trabajo fin de máster. Plantea principalmente cuatro fases teniendo en cuenta que esta metodología es adaptable según las especificaciones del proyecto:

- **Seleccionar el proceso de negocio**: El proceso de negocio seleccionado es el análisis de publicaciones científicas, tanto de los artículos como de los lugares de publicación. Se ha tenido en cuenta que la extracción de información sea fácil. Se comprenden claramente los requerimientos del usuario y se debe asegurar que la solución planteada sea acorde con los datos disponibles.
- **Escoger el tamaño del grano:** Define el nivel de detalle disponible en el esquema multidimensional. Para el trabajo de fin de máster se decidió utilizar el grano conformado por un artículo/lugar de publicación como unidad mínima de medida. Así, las métricas mostradas serán conforme a los artículos y lugares de publicación (revistas o congresos). Por ejemplo, la cantidad de artículos publicados en cierto lugar de publicación, o la cantidad total de citas de un autor según sus artículos publicados.
- **Escoger las dimensiones**: Las dimensiones son las tablas que describen al usuario los datos que resultan del proceso de negocio. Se consideran para el proyecto las dimensiones de Autor, Datos_Articulo, Editorial_Articulo, Nombre_Lugar_Publicacion, Datos_Lugar_Publicacion, Datos_Revista, Datos_CORE. Son dimensiones que permitirán realizar análisis de las publicaciones científicas.
- **Identificar los hechos:** Los hechos son las métricas, normalmente son valores numéricos aditivos de acuerdo a la granularidad propuesta en el planteamiento de la solución. En el apartado de anexos se pueden consultar los detalles de los hechos y cada una de sus definiciones.

Se han planteado dos estrellas:

- Estrella de Impacto Artículo
- Estrella de Impacto Lugar de publicación.

Hechos principales del impacto artículo:

- Total de citas de cada fuente de información.
- Número de autores por publicación científica.
- H-index del autor.

Hechos principales del impacto de lugar de publicación:

- H-index del lugar de publicación (SJR).
- Indicador SJR para revistas y congresos.
- Número total de documentos publicados.
- Total de referencias.
- Factor de impacto JCR.
- FoR CORE.
- Calificación CORE.

3.3.5 Estrella de Impacto Artículo

En la estrella de Impacto artículo se encuentran todas las métricas e información relacionada a la publicación de un artículo, como las citas por cada fuente de información, los datos del artículo, los grupos de autores, la información de la editorial y adicionalmente la tabla nombre_Lugar_Publicacion. Esta tabla es especial debido a que es ésta la que permite hacer puente (*drill across*) con la otra estrella del modelo: Impacto_Lugar_Publicacion.

Para este diseño se ha aplicado el concepto de "Bridge Table" [6] para solucionar el problema de que un artículo puede tener muchos autores y muchos autores pueden estar relacionados con un artículo.

La versión final de dicha estrella se aprecia en la Figura 12.

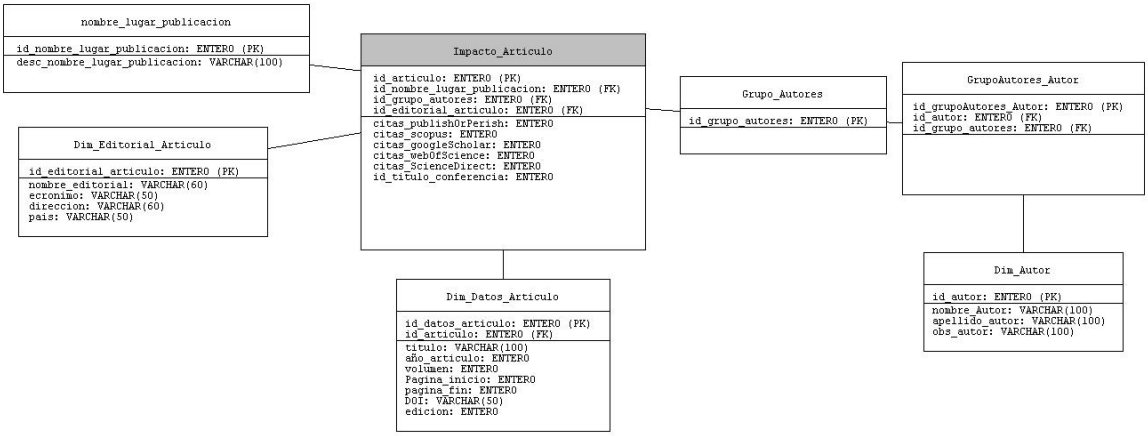


Figura 11: Estrella Impacto Artículo

Como se aprecia en la Figura 11, el concepto de "Bridge Table" se aplica en la relación entre las tablas Impacto_Articulo – Grupo_Autores –GrupoAutores_Autor – Dim_autor

3.3.6 Estrella Impacto Lugar de Publicación

En la estrella de Impacto Lugar de publicación se encuentran todas las métricas e información relacionada con los lugares de publicación de las publicaciones científicas. Encontramos las métricas SJR, JCR y CORE, teniendo en cuenta que en SJR tenemos métricas tanto para revistas y conferencias, las métricas JCR son para las revistas y las métricas CORE tanto para revistas como para conferencias.

De acuerdo a la solicitud del requerimiento hay que tener en cuenta que las métricas SJR y JCR dependen del año de publicación del artículo científico siguiendo la siguiente lógica:

Se busca en primer lugar las medidas SJR o JCR del año de publicación del artículo. Si no existe, poner las métricas del último año SJR o JCR disponible. Como no hay una definición específica para las métricas CORE, se muestra todo el conjunto de métricas CORE por cada año existente.

De acuerdo a como se ha mencionado anteriormente, en esta estrella se incluye la tabla nombre_lugar_publicacion**,** que es la misma que el modelo anterior. Sólo se muestra nuevamente para tenerla en cuenta en la estrella de Impacto Lugar de Publicación. Su función es conectar la estrella de Impacto Lugar de Publicación con la estrella de Impacto Artículo.

Para este diseño también se ha aplicado el concepto de "*Bridge Table*" [6] para solucionar el problema de que un lugar de publicación puede estar relacionado con muchos grupos CORE, y un grupo CORE puede estar relacionado con muchos artículos. Esto por la explicación anteriormente comentada de que no hay definición para mostrar los indicadores CORE.

Adicionalmente se aplican los conceptos de "*custom fact table*" y "*base fact table*" [7] para particionar la tabla de Impacto_Lugar_Publicacion en dos:

- Impacto_Revista: muestra las métricas específicas de las revistas (SJR y JCR).
- Impacto_Congreso: muestra las Métricas específicas de los congresos (SJR).

Compartiendo los encabezados de la tabla padre (Impacto_Lugar_Publicacion), lo que se consigue con esto es en primer lugar evitar que existan tantos nulos en la tabla de hechos y evitar errores de cálculo en las métricas. Además es más eficiente en cuanto a espacio de almacenamiento y resolución en las búsquedas.

El resultado del diseño final de la estrella de Impacto Lugar Publicación es:

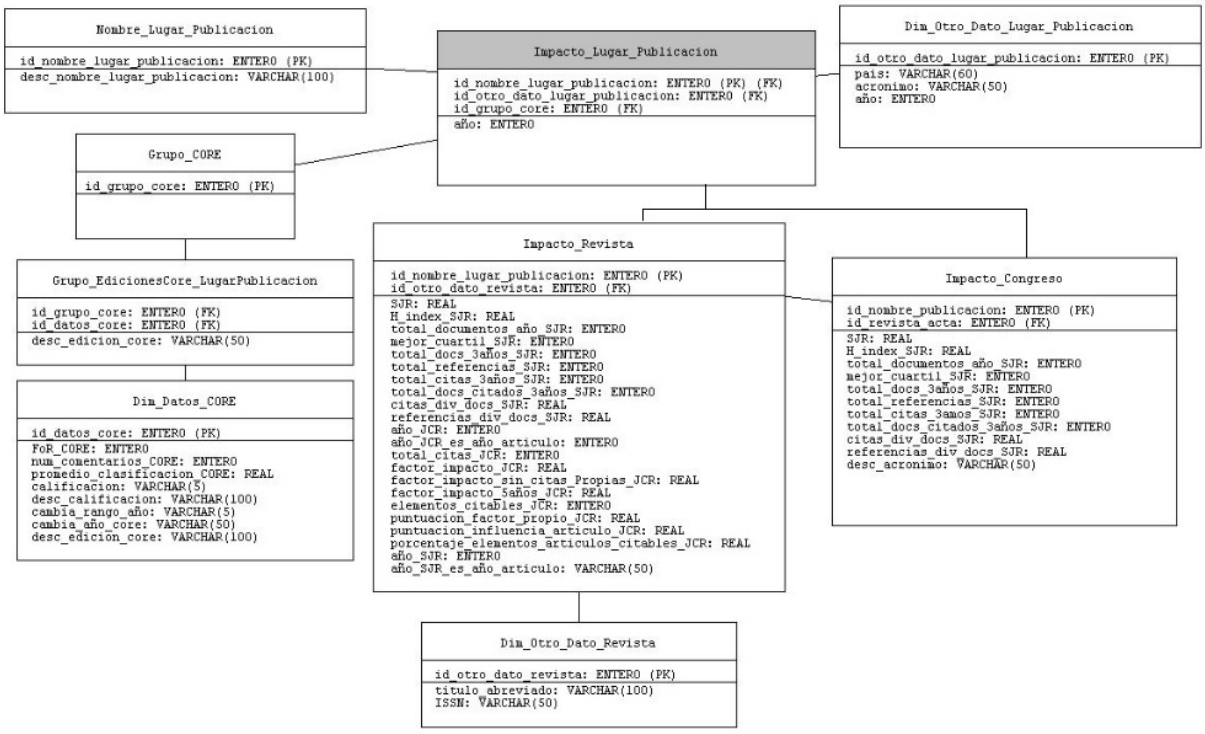


Figura 12: Estrella Impacto Lugar de Publicación

Tal como se puede observar en la Figura 12, el concepto de "Bridge Table" está aplicado en la relación de las tablas Impacto_Lugar_Publicacion – Grupo_Core – Grupo_EdicionesCore_LugarPublicacion – Dim_Datos_Core.

A su vez se puede ver el concepto de "custom fact table" y "base fact table" en la relación de las tablas Impacto_Lugar_Publicacion – Impacto_Revista – Impacto_Congreso.

3.3.7 Directorios de carga de ficheros planos

Inicialmente se ha desarrollado el proyecto teniendo una sola carpeta de fuentes de información, pero era muy difícil manejarla además de que daba una impresión de que el proceso de carga de ficheros planos era desordenado. Así que se ha solucionado creando un árbol de ficheros de carga de información, tal como se explica a continuación.

Una vez se sigan las instrucciones de descarga de los ficheros de información que se pueden consultar en el Anexo "Manuales de descarga de información", y poniendo cada uno de los ficheros en el lugar correspondiente, se tendría una estructura de directorios de carga de información como la siguiente:

- La ruta donde se aloja el directorio principal de carga puede ser:

  "C:\Planos_TFM". No obstante, la ruta donde se alojan estos ficheros es configurable de manera sencilla como se observa en el Anexo "Manual de configuración de rutas del proceso".

  Esto para que al momento de implementar el proyecto no surjan problemas, puesto que toda la solución se configura con esta ruta.

- Dentro de esta ruta encontraremos dos directorios:

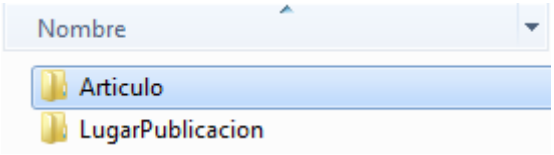


Figura 13: Directorios fuentes de información

- Dentro de la carpeta artículos, se encontrará el directorio por cada carga de información correspondiente a los artículos:

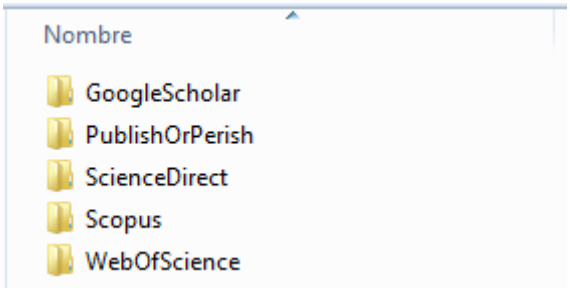


Figura 14: Directorio "Articulo"

- Dentro de cada uno de estos directorios se alojan los ficheros directamente descargados de la fuente de información. Por ejemplo los ficheros de la carpeta ScienceDirect podrían ser:

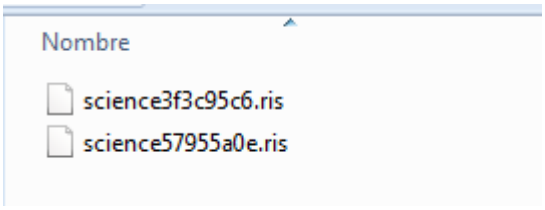


Figura 15: Ficheros de ScienceDirect (.ris)

- Con una estructura similar se encuentran los ficheros de carga de información de los lugares de publicación:

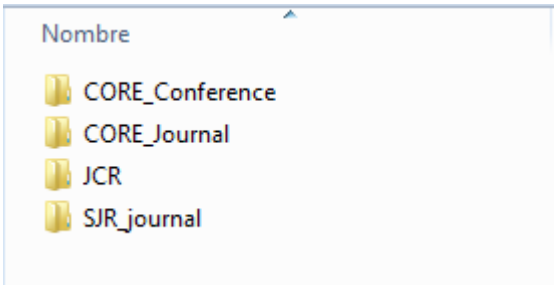


Figura 16: Directorio "LugarPublicacion"

- Cada uno de estos directorios aloja los ficheros descargados directamente de la fuente de información. Por ejemplo los ficheros para el lugar de publicación de JCR podrían ser:

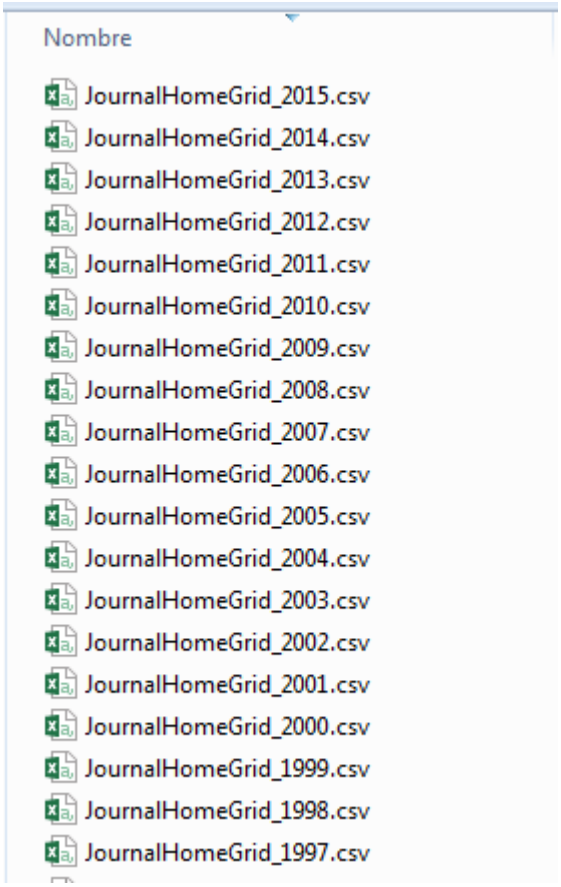


Figura 17: Ficheros JCR Ranking

3.4 Procesos ETL (Extracción, Transformación y Carga)

Los datos de las fuentes de información son cargados a través de procesos ETL (Extracción, Transformación y Carga), inicialmente en tablas de staging, que son las tablas cargadas tal y como vienen los datos en los ficheros descargados de las fuentes.

Hay otra solución ETL donde son cargadas las tablas auxiliares que son las tablas "arregladas" de staging, y finalmente la solución ETL para la carga del almacén de datos, teniendo en cuenta las reglas de negocio planteadas y el emparejamiento (*matching*) entre las fuentes de información.

En el Anexo J "Descripción de procesos ETL para el proyecto (*Staging* y tablas auxiliares)". Se puede consultar como se ha realizado el proceso ETL en el trabajo de fin de más, al mismo tiempo se menciona el funcionamiento de la herramienta y elementos utilizados.

3.4.1 Matching – Emparejamiento

Una de las principales características que hacen novedoso este trabajo de fin de máster es la unificación de diferentes fuentes de información descritas en el punto 3.3.1 "Análisis y recolección de las fuentes de información". Esto significa que se debe unificar toda la información extraída de cada fuente, transformarla y guardarla en el

almacén de datos. Las fuentes de datos relacionadas con información de las publicaciones científicas en la estrella de "Impacto_Articulo" y las fuentes de datos relacionadas con la información de los lugares de publicación en la estrella de "Impacto_Lugar_Publicacion". En resumen se debe realizar un *matching* - emparejamiento entre las fuentes de información.

El *matching* entre las fuentes de información significa que el emparejamiento se debe hacer a nivel de frase como tal y no a nivel de coincidencias de palabras, por ejemplo si la consulta es NY en el contexto de países (el tema), debería hacer match con la consulta "New York". Las consultas y los documentos también pueden hacer *matching* con respecto a la estructura, donde estructura significa estructura lingüística, por ejemplo la consulta "distancia entre el sol y la tierra" coincide con la consulta "cuán lejos está el sol de la tierra". [22]

Un problema que surge al momento de desarrollar el proceso ETL es el emparejamiento entre las diferentes fuentes de información. Para el desarrollo del proyecto ha sido un problema emparejar (hacer *matching*) los nombres de los autores, los nombres de los lugares de publicación (revistas, congresos y conferencias), incluso difieren en algunos casos el nombre de la publicación.

Para superar este problema, se ha utilizado la solución de calidad de datos de *Data Quality Services* (DQS) que permite mantener una calidad de datos en las bases de datos asegurándose de que los datos son apropiados para su uso en el proyecto.

La herramienta de *Integration Services* que aplica el módulo de calidad de datos del DQS se llama *Fuzzy Lookup* (búsqueda borrosa – aproximada). Esta herramienta hace búsquedas de coincidencias exactas y similares de acuerdo a una tabla de referencia.

Para utilizar esta esta transformación es necesario tener una tabla de referencia con los valores "limpios" de acuerdo al concepto que se esté manejando.

Si se está haciendo un emparejamiento de autores, la tabla de referencia es una tabla con la representación canónica (aceptada) de los nombres de los autores. Por ejemplo, "S. Ilarri", "Ilarri, S." y "Ilarri, Sergio" son variantes de "Sergio Ilarri" (esta última forma es la que se puede utilizar como representación canónica).

La transformación por un lado tiene como entrada la tabla de referencia y por otro lado los valores de los registros (a emparejar) de la base de datos. Por ejemplo, los nombres de los autores descargados de las diferentes fuentes de información. Una vez tenga estas dos entradas, hace un emparejamiento calculando dos valores: el primero es la puntuación de similitud y el segundo es la puntuación de confianza.

Puntuación de Similitud

Es una medida matemática que calcula la similitud del texto de referencia y la del texto a emparejar. El umbral de la puntuación de similitud es de 0 a 1, mientras más se acerque al umbral 1, los dos textos emparejados son más similares, si el valor es 1, significa que los dos textos son el mismo (duplicados exactos).

La puntuación de similitud está basada en el algoritmo de q-gram, también conocido como n-gram. [23], en resumen, es una subcadena de longitud *q*. El principio de esta función para calcular la similitud entre dos textos es que cuando dichas cadenas son

muy similares tienen muchos *q-grams* comunes, siendo *q*, generalmente un número pequeño, utilizándose uni-grams (*q*=1), bi-grams (*q*=2) y tri-grams (*q*=3). Se podría agregar *q*=-1 para ocurrencias de caracteres no definidos en el contexto. Por ejemplo, alguna letra del alfabeto griego. Esto hace que dos cadenas puedan tener un puntaje mayor de similitud. Por ejemplo, la cadena "Sergio" contiene los *bi-grams* "Se","er", "rg", "gi" y "io" y los 1-skip-2-g*rams* (los mismos *bi-grams*, pero con un salto) "Sr","eg","ri" y "go", luego calcula la similitud aplicando un algoritmo de comparación O(n) [24].

Puntuación de Confianza

Es una medida de probabilidad de que un texto dado es el mejor *match* de un texto buscado en toda la tabla de referencia, esta puntuación depende de la cantidad de registros devueltos por la búsqueda.

Propiedades y parametrización

Esta herramienta cuenta con una serie de propiedades para su parametrización de acuerdo a las necesidades del usuario, entre las más importantes y utilizadas para este proyecto están:

- Delimitadores: La transformación utiliza los delimitadores por defecto o insertados por el usuario para dividir en *tokens* los valores de las columnas.
- Similitud mínima: Es el umbral de similitud utilizada por el componente que comprende valores entre 0 y 1. Las filas con umbral igual a uno, se consideran coincidencias exactas.
- Tipo de *Join*: Es el valor que especifica si la transformación realiza una coincidencia aproximada o exacta. Por defecto tiene asignado el valor "Aproximada" cuyo valor es 2. Para coincidencias exactas el valor es 1.

Para consultar todas las propiedades de parametrización ver el Anexo "Propiedades Parametrización Búsqueda Aproximada DQS".

Ejemplo para el trabajo fin de máster

Este trabajo para las publicaciones científicas tiene un total de cinco fuentes de información (Google Scholar [9], Publish or Perish [12], Scopus [8], Science Direct [13] y Web of Science [10]). Cada publicación científica tiene uno o más autores. El problema es que, para nombrar un autor, cada fuente de información puede llamarlo de una forma diferente. Por ejemplo, el autor Sergio Ilarri en las diferentes fuentes de información:

Results Messages

	fuente_fichero	autor
1	GoogleScholar	Ilarri, Sergio
2	publishOrPerish	S Ilarri
3	scienceDirect	Ilarri, Sergio
4	Scopus	Ilarri, S.
5	webOfScience	Ilarri, Sergio

Figura 18: Ejemplo de nombre de autor por cada fuente de información

o como se ve en la tabla de auxiliar de autores (sin hacer el emparejamiento):

Results Messages

	id_autor	nombre_Autor
1	13241	Ilarri S
2	13242	Ilarri S.
3	13243	Ilarri, Sergio
4	13244	Ilarri, Sergio

Figura 19: Nombre de autor en la tabla auxiliar de autores

En la tabla de referencia, este autor está almacenado como se ve en la figura siguiente (representación canónica del nombre del autor):

Results Messages

	Autor_OK
1	Sergio Ilarri

Figura 20: Ejemplo representación canónica nombre de autor

Una vez pasan estos registros por la transformación *Fuzzy Lookup* para resolver la calidad de datos se tiene como resultado la tabla de emparejamientos, que a su vez es la tabla de Log:

Fuzzy Lookup Output Data Viewer at Data Flow Task

Detach Copy Data

id_autor	nombre_Autor	Autor_OK	_Similarity	_Confidence	_Similarity_no...
13241	Ilarri S	Sergio Ilarri	0.691369	0.9875	0.691369
13242	Ilarri S.	Sergio Ilarri	0.691369	0.9875	0.691369
13243	Ilarri, Sergio	Sergio Ilarri	0.98	0.5679234	0.98
13244	Ilarri, Sergio	Sergio Ilarri	0.98	0.5679234	0.98

Attached | Total rows: 0, buffers: 0 | Rows displayed = 4

Figura 21: Tabla de emparejamiento - Log

Donde el resultado del emparejamiento de los registros es la columna “Autor_OK” y a su vez, se muestran los valores de similitud y confianza, el tercer valor es el mismo que el de similitud.

Se observa en la Figura 21 que los dos primeros emparejamientos (excepto por le punto final) y los dos últimos son iguales, esto porque cada uno de los repetidos viene de una fuente de información diferente y el proceso igual los tiene en cuenta por separado.

Para el trabajo fin de master, se insertan todas las tablas resultantes de este proceso, para en pasos siguientes actualizar los registros de acuerdo a la tabla de emparejamiento obtenida teniendo en cuenta los valores de similitud y confianza. Por ejemplo, elegir los valores cuya puntuación de similitud sea mayor a 0.60 y puntuación de confianza > 0.5. Estos valores son configurables por el usuario actualizando en el fichero de configuración “Configuracion_DQS_similitud_confianza”. Consultar el Anexo Configuración DQS (*Data Quality Services*) similitud y confianza.

Almacenamiento de log de emparejamientos.

En la base de datos DataQualityServices de la instancia del trabajo de fin de máster, quedan almacenados en tablas todos los registros y puntuaciones de los emparejamientos automáticos que se hagan. Por ejemplo, el de autores mostrado en la Figura 21.

La estructura de todas las tablas de log son iguales a la de la Figura 21. Tiene en total seis columnas; 5 las que utilizamos en el proceso y la sexta que muestra la misma información de la similitud:

1. Id: Es el identificador de la tabla auxiliar de la dimensión.
2. Nombre_: es el nombre del artículo, del lugar de publicación o del autor a emparejar.
3. Nombre_OK: es el nombre sugerido Luego de la ejecución del proceso de emparejamiento.
4. _Similarity: Puntuación de similitud.
5. _Confidence: Puntuación de confianza.

¿Cómo se podría utilizar el fichero de configuración de nombres de autor después del emparejamiento automático?

Imaginar que desde una fuente de información de publicaciones científicas se descarga de un artículo el nombre de autor “S Ilarri” haciendo referencia a un tal “Sergio ilarri” y se descarga de otro artículo el nombre del autor “S Ilarri” que hace referencia a un tal “Sebastián Ilarri”.

Hay dos opciones: la primera es configurar los valores de similitud y confianza relativamente bajos para que no se haga un emparejamiento automático y utilizar directamente el fichero de configuración de nombres de autor para asignar la representación canónica de cada autor manualmente y la segunda es poner los valores de similitud y confianza justos para que este emparejamiento se haga automático y en caso de emparejamientos errados (como este caso), hacer uso del fichero de configuración de nombres.

El proceso de emparejamiento automático podría poner a los dos “S Ilarri” con la misma forma canónica (“Sergio Ilarri”, con el mismo ejemplo anterior).

Para estos casos particulares se puede utilizar el fichero de configuración de autores, se debe mirar (para este ejemplo) la tabla de log “log_Autor_Matching” para consultar el id del autor a modificar y rellenar el fichero de configuración (Excel) como se observa en la Figura 22.

	A	B
1	**id_Autor**	**Nombre_Autor**
2	232	Sebastián Ilarri
3		

Figura 22: Fichero configuración de nombres de autor

La ETL se encarga de cambiar el nombre del autor y relacionarlo con toda la información del mismo. Por ejemplo, sus publicaciones científicas.

3.4.2 Carga del Almacén de Datos

Para la carga de información del almacén de datos (*Data Warehouse*) se han desarrollado principalmente dos scripts de procedimientos almacenados: el primero para la carga de la estrella de “Impacto Articulo” y el segundo para la carga de la estrella de “Impacto Lugar de Publicación”, el segundo dependiendo del primero.

Los scripts para su consulta o análisis están en el Anexo “Script carga estrella Impacto Articulo” y en el Anexo “Script carga estrella Lugar de Publicación”. Lo que se describirá a continuación son los pasos generales que se siguen para la carga de cada estrella.

Carga Modelo Impacto Artículo

El proceso de la carga del Modelo de Impacto Artículo (a partir del proceso ETL de unificación de artículos) realiza los siguientes pasos (hay que tener en cuenta que el proceso se ejecuta registro por registro del flujo de datos):

Paso 1: Insertar un nuevo id_articulo en Impacto_Articulo y capturarlo.

- 1.1: Insertar demás los datos de la tabla Impacto_Articulo de acuerdo al id asignado.

Paso 2: Insertar id_datos_articulo en Dim_Datos_Articulo.

- 2.1: Insertar valores de Dim_Datos_Articulo.
- 2.2: Actualizar en Impacto_Articulo el Id_datos_articulo.

Paso 3: Insertar nuevo id_grupo_autores en Grupo_Autores.

- 3.1 Actualizar Impacto_Articulo con el id_grupo_autores.

Paso 4: Insertar en GrupoAutores_Articulo el valor de id_grupo_autores por cada autor encontrado en los autores del artículo, crear un registro con el id_grupo_autores y el id_autor.

Paso 4.2:Validar si hay un grupo con los mismos autores:

4.2.1 Consultar si existe un grupo con los mismos autores.

4.2.2 Si existe un grupo con los mismos autores, actualizar el Impacto_Articulo con el grupo de autores existente.

4.3.3 Eliminar el nuevo grupo creado en GrupoAutores_Articulo. (Actualizar BD)

Paso 5: Insertar datos de editorial si existen.

Paso 6: Insertar id_nombre_lugar_publicacion y el nombre del Lugar de Publicación, actualizar id_nombre_lugar_publicacion en id_impacto_articulo.

Paso 7: Insertar id_nombre_lugar_publicacion en Impacto_Lugar_publicacion.

Carga Modelo Impacto Lugar Publicación

Teniendo en cuenta la misma anotación de la carga del modelo de Impacto Artículo, y a partir del último paso de la carga del modelo anterior, inicia la carga del segundo modelo, el de Impacto lugar de Publicación:

Tarea: Buscar los datos del Lugar de Publicación de los Artículos. Es necesario además del nombre del lugar de publicación, tener siempre presente el año del artículo, puesto que este año es el que asigna las métricas según la fuente de información.

NOTA: Sólo si el lugar de publicación es SJR puede tener revistas y conferencias a la vez.

TIPO_LP: Tipo de Lugar de publicación (definiciones):

Tipo 1 -> REVISTA (book de SJR va a revista): buscar métricas del año del artículo o último año disponible actual o última.

Tipo 2 -> CONFERENCIA: de SJR buscar métricas del año del artículo o último año disponible actual o última.

Tipo 3 -> CORE_REVISTA: Inserta toda la información de CORE Revista.

Tipo 4 -> CORE_CONFERENCIA: Inserta toda la información de CORE Conferencia.

Tipo 5 -> REVISTA (de JCR): busca métricas del año del artículo o último año disponible actual o última.

Paso 8: Con los datos del lugar de publicación, inserta id_otro_dato_lugar_publicacion. Inserta los demás datos del Lugar de Publicación.

Paso 9:

Si TIPO_LP = 1

- Con el año del artículo, buscar si hay un lugar de publicación (LP) del mismo año o el último año disponible. Inserta en la tabla si es el mismo año o si no lo es.

- Inserta id_nombre_lugar_publicacion en Impacto_Revista.
- Inserta datos Impacto_Revista.
- Inserta id_otro_datos_revista en Dim_otro_Dato_Revista.
- Inserta la información restante.

Si TIPO_LP = 2

- Con el año del artículo, buscar si hay un LP del mismo año o el último año disponible
- Inserta id_nombre_lugar_publicacion en Impacto_Congreso.
- Inserta datos Impacto_Congreso.

Si TIPO_LP = 3 o 4

- Inserta id_grupo_core en Impacto_Lugar_publicacion, a su vez, almacenar este id en una variable.
- Inserta id_grupo_core en Grupo_Core, a su vez, almacenar este id en una variable.
- Inserta id_grupo_core en Impacto_Lugar_publicacion.
- Busca todos los CORES de ese lugar de publicación y por cada uno, inserta en Grupo_EdicionesCore_LugarPublicacion y Dim_Datos_Core.

Si TIPO_LP = 5

- Con el año del artículo, busca si hay un LP del mismo año o el último año disponible.
- Inserta id_nombre_lugar_publicacion en Impacto_Revista.
- Inserta datos Impacto_Revista.
- Inserta id_otro_datos_revista en Dim_otro_Dato_revista.

  Inserta demás información de los datos de revista.

3.5 Explotación de la Información

La explotación de información se realiza a través de una herramienta de la suite de Microsoft llamada Power BI. Es un servicio para el análisis de información que proporciona visualizaciones variadas de los datos más críticos del caso de estudio y permite crear informes interactivos para tener toda la información controlada; es allí donde se muestran paneles de información con las métricas de los autores, artículos y lugares de publicación.

No obstante, los paneles diseñados y desarrollados en el proyecto puede ser solo una base para todas las posibilidades que tiene el usuario de ver la información; si bien la herramienta es fácil de utilizar, hay bastante información y tutoriales en el link https://powerbi.microsoft.com/es-es/learning/ donde toda la información es presentada en tres secciones:

Figura 23: Presentación módulo "Learnig Power BI Microsoft"

Es importante mencionar que se ha revisado una herramienta también desarrollada por la universidad de Zaragoza llamada Kampal: "Kampal es una herramienta que construye y estudia las redes formadas por los miembros de la Universidad de Zaragoza a través de su actividad investigadora: artículos y proyectos. El énfasis está puesto en las relaciones entre personas, en la dinámica de la interacción, más que en las propiedades individuales." [9]. Una visualización de ejemplo de Kampal, es la interacción entre las macro áreas de la universidad de Zaragoza (Figura 44).

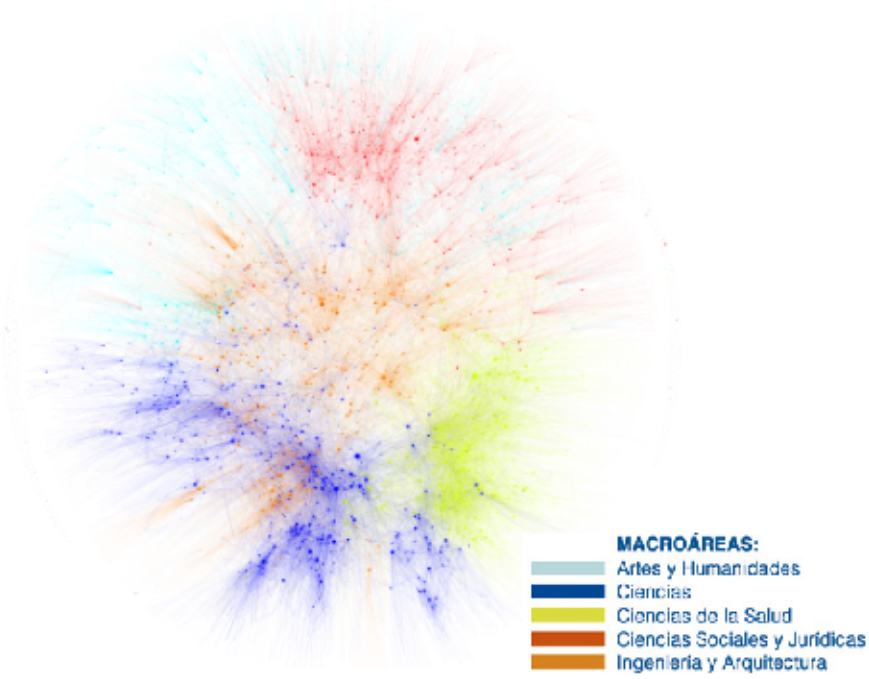


Figura 24: Colaboraciones en la Universidad de Zaragoza; coloreado por todas las macroáreas [9]

Observando las visualizaciones de Kampal e investigando en la biblioteca de visualizaciones de la tienda de Microsoft, se ha encontrado en *Power BI* una visualización que permite representar de manera visual las coautorías de las publicaciones científicas. Con la inspiración dada por las visualizaciones de kampal se ha logrado incorporar en el proyecto una visualización de este estilo:

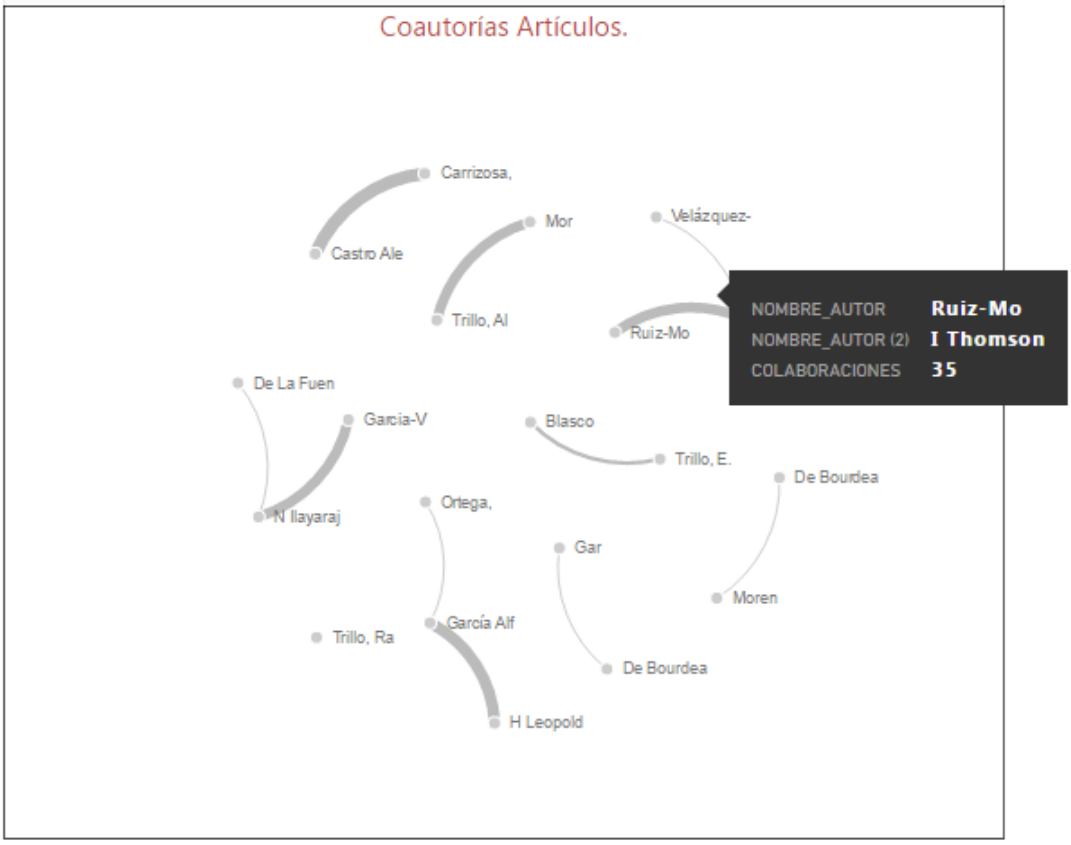


Figura 25: Visualización Coautorías con el desarrollo del TFM

Se visualiza en la imagen algunas de las coautorías de los artículos cargados en las fuentes de información de este proyecto. El grosor de la línea de conexión entre autores es directamente proporcional a la cantidad de artículos en los que tengan coautoría.

Las presentaciones visuales desarrolladas para el trabajo de fin de máster como se ha comentado, están hechas en *Power BI* así que este trabajo de fin de master tiene todas las fortalezas y capacidades que ofrece esta herramienta.

Una de las características más importantes de *POWER BI* y que no se puede observar en el documento, son los filtros dinámicos, es decir, estos filtros que permiten cambiar la visualización y la información a medida que se explora el panel visual. Por ejemplo, cuando estamos mostrando la información de autores, en principio se muestra la información de todos los autores que estén en el panel, pero en caso de querer ver la información solo de un autor, se hace click sobre la información del mismo y el panel cambia la visualización mostrando la información solo de este autor.

Para el desarrollo se estas visualizaciones se han considerado algunas métricas del total de métricas obtenidas por publicación científica o lugar de publicación. Las métricas que no se visualizan están en el almacén de datos y disponibles para los usuarios puedan armar visualizaciones de interés con estas.

Las descripciones de todas las métricas se encuentran en el Anexo" Definición de las métricas". Tener en cuenta que la información mostrada en este documento es de carácter meramente orientativo.

Es importante hacer un pre-filtro de los autores y/o publicaciones de interés puesto que, si no se hace esto, es posible que la herramienta muestre mucha información y puede ser difícil de analizar.

Los informes creados para el desarrollo del proyecto son:

3.5.1 Información del autor por artículo

En la figura 46 se observa la visualización de los autores por artículos. Especificamente se puede visualizar el total de artículos publicados (de los autores seleccionados), la cantidad de citas por cada fuente de información y el H- index del autor.

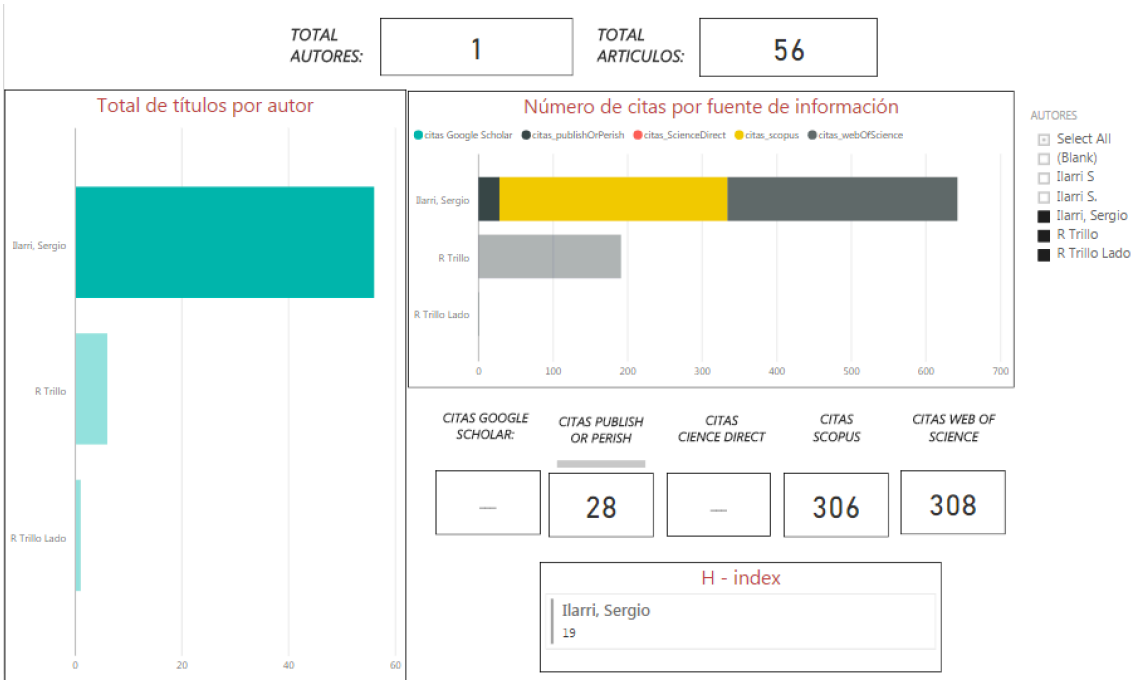


Figura 26: Visualización informacuón del autor por artículo Power BI

Se puede visulizar un autor o más, la herramienta directamente muestra la comparativa entre los autores seleccionados.

3.5.2 Información métricas del autor con respecto a los lugares de publicación

En este panel de visualización se observan las métricas que tienen que ver con los lugares de publicación. En la parte superior, un resumen de la cantidad de autores mostrados y el total de publicaciones científicas.

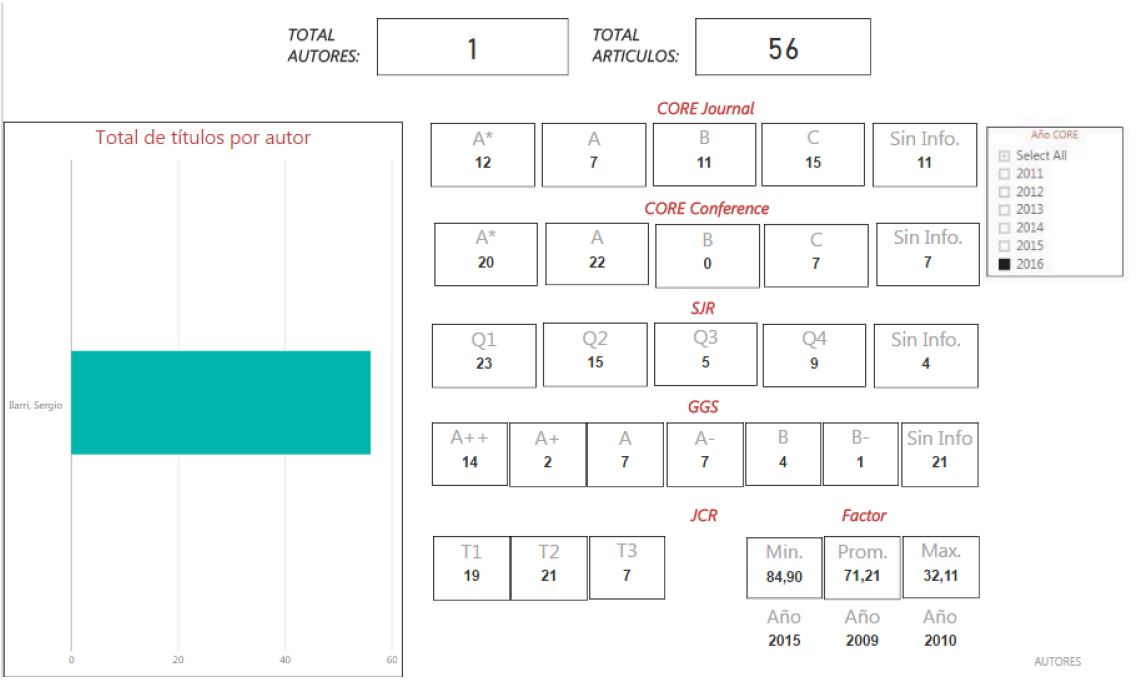


Figura 27: VIsualización de métricas del autor con respecto al lugar de publicación

En la parte derecha un panel con los autores consultados, al poner el cursor sobre las barras, la herramienta muestra el nombre del autor y la cantidad de publicaciones científicas.

En la parte central del panel están las métricas de las publicaciones del autor con respecto a los lugares de publicación. Se visualiza la cantidad de artículos del autor que está en cada una de las medidas del lugar de publicación. Por ejemplo, el lugar de publicación GGS [14] tiene entre otras, las calificaciones A++, B, B-, etc. (estas calificaciones son del lugar de publicación) y se visualiza cuantos artículos están en cada calificación del lugar de publicación. Lo mejor sería que los autores tuvieran publicaciones en los lugares de publicación con mejor *ranking*.

Para el caso de los lugares de publicación CORE, se debe seleccionar el año que se quiere visualizar puesto que para estos lugares de publicación no hay definida ninguna regla para mostrar los rankings (como ya se ha en el documento).

3.5.3 Detalle de las publicaciones por autor

Es una representación dónde se visualiza algún detalle de las publucaciones de los autores. Igual como en todos los páneles de vsualización, es posible ver la información de un autor o más al mismo tiempo:

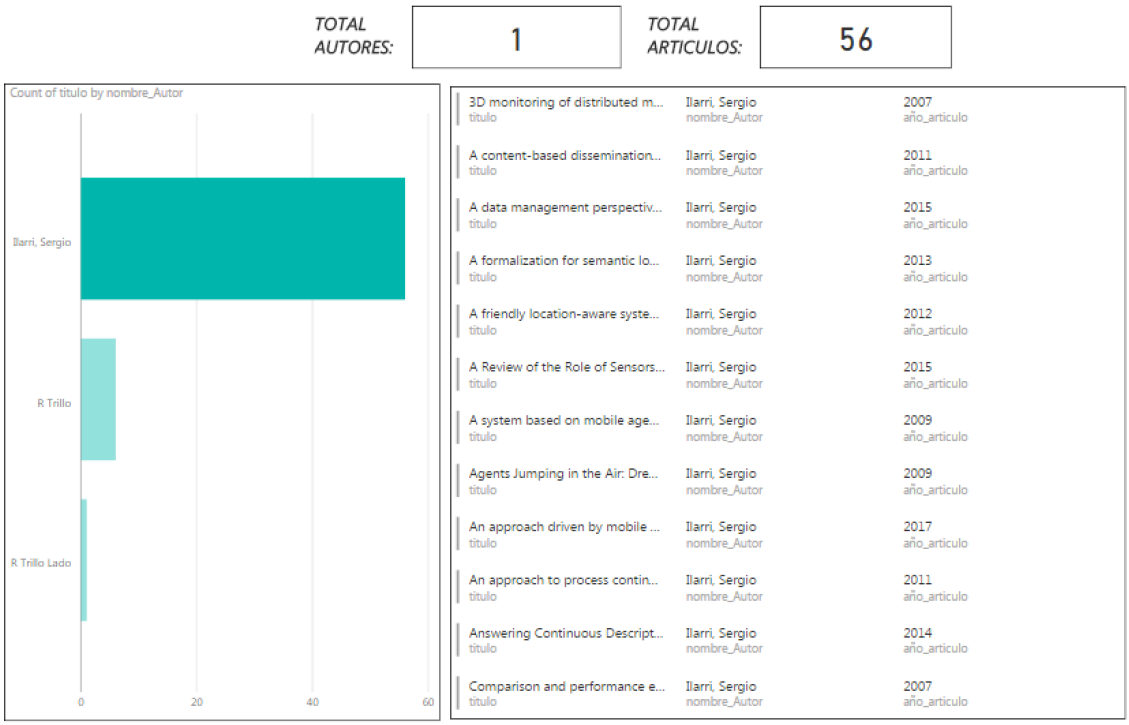


Figura 28: Detalle de las publicaciones por autor

En este panel se ve el detalle de las publicciones científicas de cada autor; la cantidad de publicaciones en la parte superior, el titulo de cada artículo, el autor y el año del artículo.

3.5.4 Diagrama de coautorías

Este diagrama ya mencionado en este capítulo muestra las coautorías.

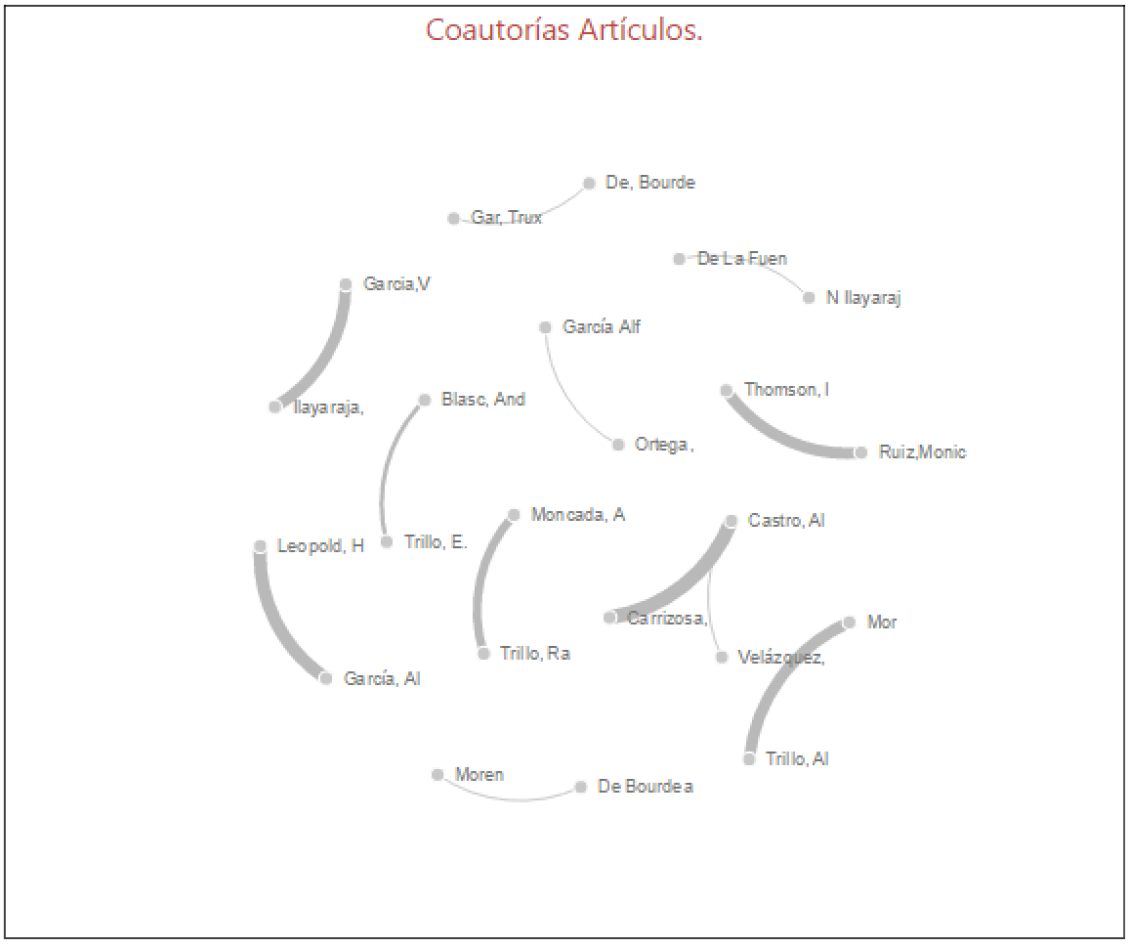


Figura 29: Coautorías de publicaciones científicas

Igual como se ha dicho anteriormente, se visualiza en la Figura 49 algunas de las coautorías de los artículos cargados en las fuentes de información de este proyecto. El grosor de la línea de conexión entre autores es directamente proporcional a la cantidad de artículos en los que tengan coautoría.

3.5.5 Ejemplo de visualización de grano grueso

La herramienta también permite hacer visualizaciones de grano grueso (más globales), para hacer algún tipo de resumen, alguna estadística en especial o lo que el usuario considere ver a rasgos generales.

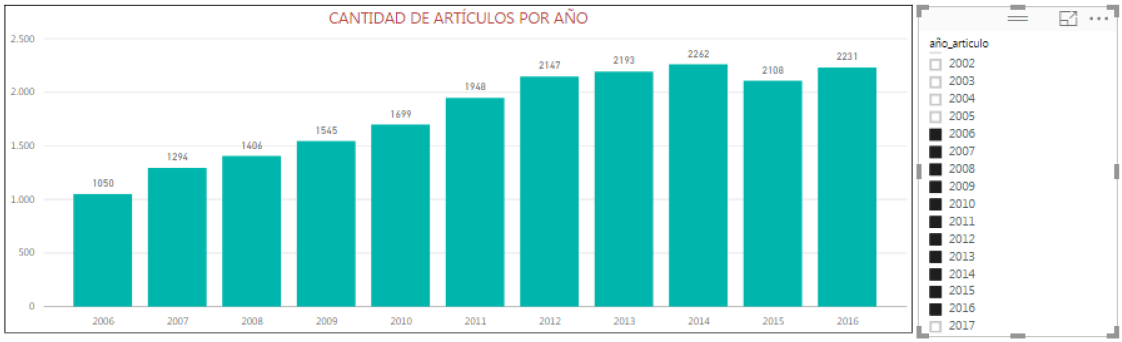


Figura 30: Publicaciones científicas por año

En la figura 50 muestra una visualización de información general: la cantidad de publicaciones científicas por año (también referenciada al inicio del documento), estas cifras son mostradas de acuerdo a las fuentes de información cargadas. Como esta visualización, la herramienta permite hacer visualizaciones de grano grueso. Por ejemplo, cantidad de autores por año.

IV. Gestión del proyecto

En este capítulo se describe el todo el proceso de programación del trabajo fin de master: la metodología de desarrollo, los principios de la metodología SCRUM, la planificación y el calendario de actividades del mismo.

4.1 Programación del proyecto

Una de las etapas más complicadas para el gestor de proyecto es la programación del mismo. Establecer los tiempos y recursos totales para el desarrollo de un proyecto de forma coherente es una ardua tarea porque los proyectos son diferentes, al igual que las metodologías y resultados esperados. Generalmente las estimaciones son optimistas aun cuando se intente tener en cuenta todas las variables que afecten el desarrollo del mismo.

La programación del proyecto implica separar todo el trabajo de un proyecto en actividades complementarias y considerar el tiempo requerido para completar dichas actividades." [10]

4.2 Metodología

Para el desarrollo del trabajo de fin se máster se ha tomado como base la metodología SCRUM: "Scrum es una metodología ágil de desarrollo de proyectos que toma su nombre y principios de los estudios realizados sobre nuevas prácticas de producción por Hirotaka Takeuchi e Ikujijo Nonaka a mediados de los 80" [14]. Los roles principales en Scrum son el *ScrumMaster*, que mantiene los procesos y trabaja de forma similar al director de proyecto, el *ProductOwner*, que representa a los *stakeholders* (clientes externos o internos), y el *Team* que incluye a los desarrolladores.

Dado que el equipo de trabajo para el desarrollo de este proyecto son dos personas: el estudiante del máster (desarrollador del proyecto) y el profesor director del TFM. Se hace una adaptación de la metodología donde el estudiante es el equipo de desarrollo y el profesor es el *ScrumMaster* con tareas de director de proyecto y a su vez representa al equipo de *stakeholders* puesto que es él se considera como el cliente.

4.2.1 Generales de la metodología Scrum

- Los equipos de trabajo pequeños están organizados para "maximizar la comunicación, minimizar los gastos generales y maximizar el hecho de compartir conocimiento tácito e informal".
- El proceso debe adaptarse a los cambios técnicos y de negocios "para asegurar que se produzca el mejor producto posible".
- El proceso produce incrementos frecuentes de software "los cuales se pueden inspeccionar, ajustar, probar, documentar y construir".

- El trabajo de desarrollo y la gente que lo realiza están divididos en “particiones o paquetes de bajo acoplamiento”.
- Conforme se construye el producto se realizan pruebas y documentación constantes.

4.3 Planificación y seguimiento

Para el desarrollo de este trabajo de fin de máster, se ha invertido un total de 672 horas de trabajo en un total de 168 días. Teniendo en cuenta que la inversión diaria en promedio fue de 4 horas.

La fecha de inicio del proyecto, teniendo en cuenta el desarrollo de la propuesta ha sido el día 24 de febrero de 2017 y finalizado el día 28 de agosto de 2017. Durante este periodo se han tomado 10 días de receso que inició el día 17 de junio de 2017 y ha finalizado el día 27 de junio de 2017.

En la figura se puede ver el cronograma de actividades generales del proyecto.

Figura 31: Cronograma de actividades

Se ha considerado incluir un gráfico que represente el esfuerzo aproximado por tarea realizada:

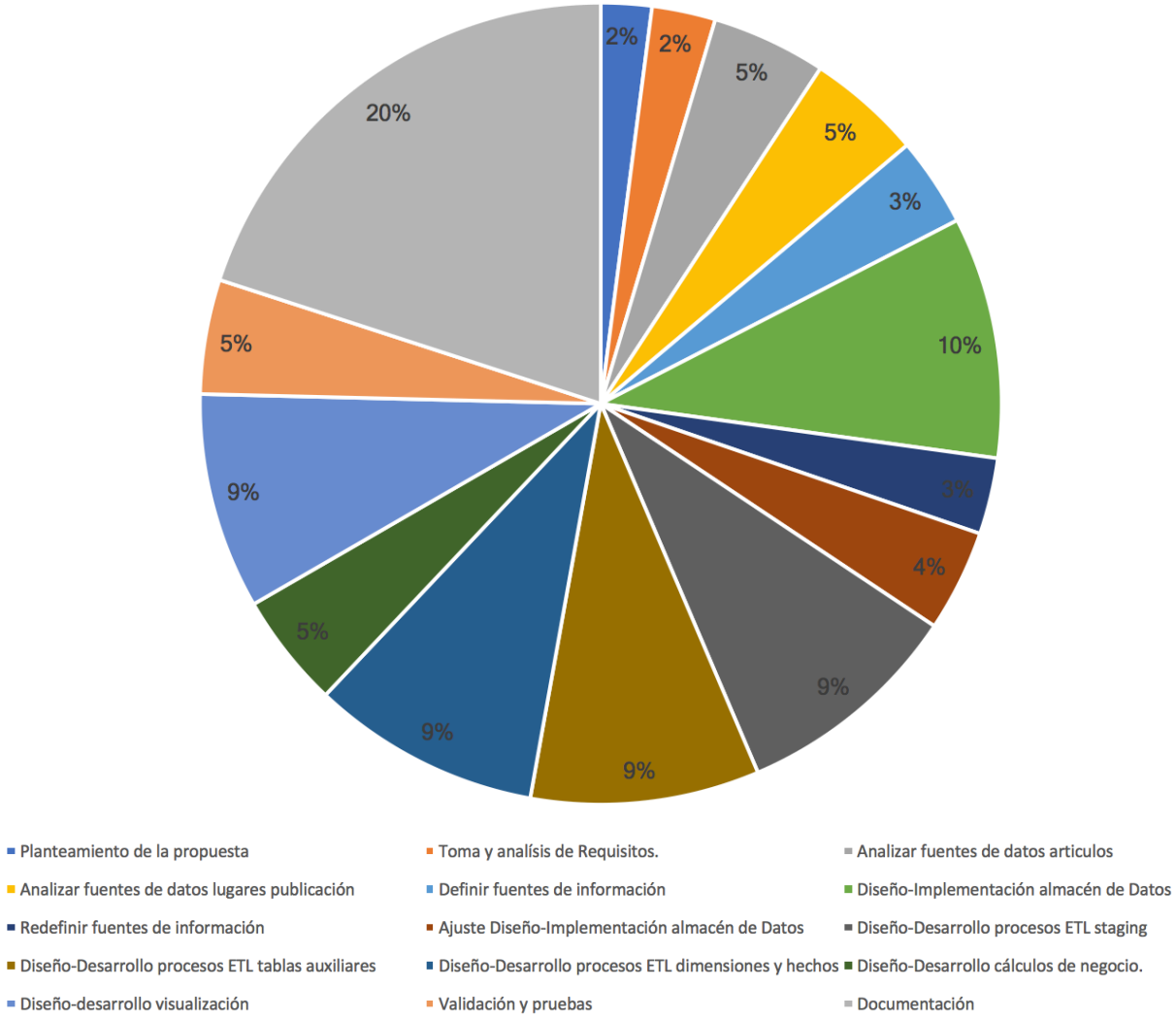


Figura 32: Diagrama de esfuerzo por tarea

Aunque no ha sido considerado, la documentación del proyecto ha sido una de las tareas que más esfuerzo ha tenido, y por esta misma, se ha tenido que cambiar la fecha de depósito planeada, además de los ajustes realizados a la herramienta como tal.

Esto que ha ocurrido, se tomará como aprendizaje profesional a la hora de gestionar un proyecto. Es muy importante tener en cuenta el tiempo y el esfuerzo que se invierte al momento de hacer la documentación de un proyecto. En este caso el impacto no ha sido grave, pero en el ámbito laboral puede generar algún impacto grave el no tener en cuenta estos puntos.

V. Conclusiones y Trabajo Futuro

5.1 Conclusiones

- El sistema desarrollado e implementado cumple con el objetivo propuesto: se puede hacer un análisis sencillo de las métricas de las publicaciones científicas. El almacén de datos es un modelo en estrella que cualquier usuario de bases de datos podría entender, se han aplicado conceptos científicos como se ha descrito en la memoria y cumple con las características planteadas en el proyecto. Es importante mencionar que se han podido consolidar todas las fuentes en el mismo almacén de datos.
- En el transcurso del proyecto, de acuerdo al análisis de las fuentes de información, algunas fuentes de todas las consideradas inicialmente fueron descartadas porque no cumplían con los estándares de información requerida por el proyecto. Por ejemplo, que la fuente de información al momento de descargarse, mostrara información de citas, autores, lugar de publicación, etc. De igual manera el diseño del almacén de datos y el proceso ETL experimentaron cambios con respecto a lo previsto.
- El sistema es capaz de almacenar y gestionar todo el inventario de fuentes que se encuentra en el directorio de carga de ficheros. Adicionalmente permite hacer cambios y configuraciones con los ficheros de configuración, los cuales son desarrollados para que el usuario final pueda gestionarlos.
- Las fuentes de información de prueba y la cual está cargada en el proyecto, se encuentra actualizada a julio de 2017.
- Se ha logrado hacer una adaptación de la metodología de desarrollo ágil SCRUM de acuerdo a las necesidades del proyecto y los dos integrantes del mismo: director del proyecto y alumno.
- El diseño del almacén de datos de acuerdo a la metodología de Kimball, es un modelo estrella relacional que facilita el manejo de la persistencia y el cumplimiento de las reglas de negocio planteadas en el proyecto. Se han implementado dos estrellas llamadas “Impacto_Articulo” e “Impacto_lugar_publicación”.

5.2 Trabajo Futuro

- Implementación automática de fuentes de información: si bien la implementación de una fuente de información puede llegar a ser sencilla gracias al modelo del almacén de datos y el desarrollo del proceso ETL, no deja de ser un trabajo de desarrollo manual. Para trabajo futuro se podría contemplar incorporar algún software o herramienta que consiga hacer una carga automática de una nueva fuente de información.

- Matching-Emparejamiento: se han implementado en el proyecto técnicas de emparejamiento y en caso de que sea claro que no se ha hecho correctamente un emparejamiento, la herramienta permite hacer emparejamientos manuales a través de los ficheros de configuración. No obstante, un trabajo futuro podría ser el desarrollo o incorporación de un módulo especializado de emparejamiento de datos.
- Sería interesante incluir en el proyecto algún módulo de minería de texto. Aunque este tema actualmente es una tecnología emergente y seguramente mejor desarrollada, sería útil para obtener información importante de nuestro modelo de base de datos. Por ejemplo, con un análisis de minería de textos, sería posible analizar de forma automática qué tema está de moda según las publicaciones científicas incluidas en el almacén de datos, otro ejemplo podría ser, hacer un análisis automático de las necesidades de un tópico de investigación, para asimismo poder plantear nuevas investigaciones y poder solventar dichas necesidades.
- Teniendo en cuenta la evolución y constante cambio de cualquier herramienta de software, es posible plantear opciones de mejora y optimización del sistema de acuerdo a las mejoras del software utilizado.

VI. Bibliografía

[1] K. North y R. Rivas, Gestión del Conocimiento. Una Guia Práctica Hacia la Empresa Inteligente., Libros en Red, 2008.

[2] KNOWMAP, Una Metodología y Herramienta de Soporte a la Gestión del Conocimiento, Socintec Corporación, 2004.

[3] Universidad de Zaragoza, «Servicio de Gestión de Datos,» [En línea]. Available: http://www.unizar.es/datuz. [Último acceso: 16 09 2017].

[4] Universidad de Zaragoza, «KAMPAL,» 05 06 2017. [En línea]. Available: http://kampal.unizar.es/.

[5] A.-W. H. «Publish or Perish,» Harzing, 1999. [En línea]. Available: http://www.harzing.com/resources/publish-or-perish. [Último acceso: 16 09 2017].

[6] Elsevier, «Scopus,» [En línea]. Available: https://www.scopus.com/. [Último acceso: 16 09 2017].

[7] Google, «Google Scholar,» Google, [En línea]. Available: https://scholar.google.es/ . [Último acceso: 16 09 2017].

[8] Thomson Reuters, «Web of Science,» FECYT, [En línea]. Available: http://wos.fecyt.es. [Último acceso: 16 09 2017].

[9] «Science Direct,» Elsevier, [En línea]. Available: http://www.sciencedirect.com/. [Último acceso: 16 09 2017].

[10] «CORE Conference Portal,» Computing Research & Education, 16 09 2017. [En línea]. Available: http://portal.core.edu.au/conf-ranks/.

[11] «CORE Journal Portal,» Computing Research & Education, [En línea]. Available: http://portal.core.edu.au/jnl-ranks/. [Último acceso: 16 09 2017].

[12] «JCR Ranking,» [En línea]. Available: http://jcr.fecyt.es/ . [Último acceso: 16 09 2017].

[13] «Scimago Journal & Country Rank,» Scimago Lab, [En línea]. Available: http://www.scimagojr.com/journalrank.php. [Último acceso: 16 09 2017].

[14] GII, GRIN, SCIE, «The GII-GRIN-SCIE (GGS) Conference Rating,» 16 09 2017. [En línea]. Available: http://valutazione.unibas.it/gii-grin-scie-rating/conferenceRating.jsf.

[15] C. A. Benavides y C. Quintana, Gestión del Conocimiento y Calidad Total., Madrid: Ediciones Diaz de Santos, S.A., 2003.

[16] J. Alegre Vidal, La Gestión del Conocimiento como Motor de la Innovación.:Lecciones de la Industria de Alta Tecnología para la Empresa., Universitat Jaume-I, 2004.

[17] C. Imhoff, N. Galemmo y J. G. Geiger, Mastering Data Warehouse Design - Relational and Dimensional Techniques, Indianapolis,: Wiley Publishing Inc, 2003.

[18] R. Kimball y M. Ross, The Data Warehouse Toolkit - The Definitive Guide to Dimensional Modeling, Indianapolis,: Wiley, 2013.

[19] R. Kimball, M. Ross, W. Thornthwaite, J. Mundy y B. Becker, The Data Warehouse Lifecycle Toolkit, 2nd ed., New York: Wiley, 2008.

[20] Thomson Reuters, «Researcher ID,» [En línea]. Available: http://www.researcherid.com/. [Último acceso: 16 09 2017].

[21] T. Reuters, “RIS” Format Documentation Adding a "Direct Export" Button to Your Web Page or Web Application, ResearchSoft, 2008.

[22] R. Kimball, M. Ross, T. Warren, M. Joy y B. Bob, Practical Techniques for Building Data warehouse and Business Intelligence Systems., Indianaplis, IN 46256: Wiley Publishing, Inc., 2008.

[23] R. Kimball, L. Reeves, M. Ross y W. Thornthwaite, Expert Methods for Designing, Developing and Deploying Data Warehouses., NY, Chichester, Weinheim, Brisbane, Singapore and Toronto.: John Wiley & Sons, Inc., 1998.

[24] L. Hang y X. J. , Semantic Matching in Search, vol. 7, Foundations and Trends in Information Retrieval, 2013.

[25] J. D. Cohen, Recursive hashing functions for n-grams, New York: ACM, 1997.

[26] E. Ukkoen, Approximate string-matching with q-grams and maximal matches, Helsinki: Theoretical Computer Science, 1992.

[27] I. Sommerville, Ingeniería del Software, Pearson Education, 2006.

[28] J. Palacio, Flexibilidad con SCRUM, Safe Creative, 2008.

[29] I. Jacobson, G. Booch y J. Rumbaugh, El Proceso Unificado de Desarrollo de Software, Pearson Education, 2000.

[30] K. E. Kendall y J. E. Kendall, Análisis y Diseño de Sistemas, Pearson Education, 2005.

Apéndice A. Definición de las métricas.

En este anexo se encuentra la definición de las métricas de las fuentes de información.

Indicadores CORE Journal – Conferences:

Descripción de los indicadores de CORE.

Ranking	Calificación	Características
A*	Excepcional	Visibles y conocidos tanto en su comunidad, como entre científicos fuera de esa comunidad. Si una conferencia es A* casi todos los investigadores son calificación A* . Las revisiones de los trabajos son realizadas por líderes internacionales que han publicado área del trabajo presentado, y proporcionar retroalimentación detallada y extendida. El tiempo entre la fecha límite de presentación y la revisión se mide en múltiples meses, y es similar a lo que se logra en revistas excelentes Ser invitado a ser un Presidente de PC de una Conferencia A * es un profesional altamente significativo gran reconocimiento y probablemente tendrán h-index de +25 y varias publicaciones en esa conferencia.
A	Excelente	Lideres de investigación rutinarios (envían trabajos a conferencias, pero a veces no asisten si su trabajo no ha sido aceptado). Ser invitado a ser un Presidente de PC de una Conferencia A, es un gran hito profesional.

B	Bueno - Muy Bueno	Los presidentes de B, suelen ser investigadores significativos en mitad de su carrera que han establecido solidos antecedentes. Las revisiones de los trabajos B están realizadas por personas que son conocedoras del tema, pero no son tan completas o detalladas como para ser A o A*.
C	Escuchado - Satisfactorio	Trabajos que exhiben el mínimo requerido para obtener una calificación CORE. Las revisiones suelen ser breves, y solo una fracción de los comentarios, serán comentarios técnicos detallados. En las conferencias: Se influirán autores dónde un arbitro o incluso dos lo han rechazado. Ser moderador de una conferencia C, representa un nivel modesto de reconocimiento.
Australasia	Conferencias locales	Solo para conferencias. Son conferencia nacionales relacionadas con solo un país. Eliminadas de la lista de CORE a partir del 2013. se pueden encontrar en el portal como parte de las listas anteriores.
Sin calificación.		se han proporcionado datos de calidad insuficientes para obtener una calificación CORE.

Indicadores GGS

Descripción de los indicadores GGS.

Clase	Calificación	Medida	Descripción
Clase !	A++, A+	34+39 = 73 Conferencias	Excelentes, conferencias de primera categoría.

Clase II	A, A-	85+84 = 169 Conferencias.	Eventos muy buenos.
Clase III	B, B-	273+145=418 Conferencias.	Eventos de buena calidad.
-	Trabajo en progreso.	1573 Conferencias.	Trabajo en progreso.

Las definiciones de los indicadores de JCR [12] y SJR [13] se han extraído textualmente de las fuentes citadas.

Indicadores JCR

Descripción de los indicadores de JCR.

Total Cites

The total number of times that a journal has been cited by all journals included in the database in the JCR year.

Citations to journals listed in JCR are compiled annually from the JCR years combined database, regardless of which JCR edition lists the journal and regardless of what kind of article was cited or when the cited article was published. Each unique article-to-article link is counted as a citation.

Citations from a journal to an article previously published in the same journal are compiled in the total cites. However, some journals listed in JCR may be cited-only journals, in which case self-cites are not included.

Journal Impact Factor Percentile

The Journal Impact Factor Percentile transforms the rank in category by Journal Impact Factor into a percentile value, allowing more meaningful cross-category comparison. It is calculated by using the following formula:

$$Journal\ Impact\ Factor\ Percentile = \frac{(N - R + 0.5)}{N}$$

Where:

- *N is the number of journals in the category*
- *R is the Descending Rank*

Journal Impact Factor

The Journal Impact Factor is defined as all citations to the journal in the current JCR year to items published in the previous two years, divided by the total number of scholarly items (these comprise articles, reviews, and proceedings papers) published in the journal in the previous two years.

Though not a strict mathematical average, the Journal Impact Factor provides a functional approximation of the mean citation rate per citable item. A Journal Impact Factor of 1.0 means that, on average, the articles published one or two years ago have been cited one time. A Journal Impact Factor of 2.5 means that, on average, the articles published one or two years ago have been cited two and a half times. The citing works may be articles published in the same journal. However, most citing works are from different journals, proceedings, or books indexed in Web of Science Core Collection.

Journal Impact Factor - Subject Category

The JIF - Subject Category is a quartile comparison metric that takes the Journal Impact Factor without Self Cites and applies a journal to a quartile based on that value.

5-Year Journal Impact Factor

The 5-year journal Impact Factor, available from 2007 onward, is the average number of times articles from the journal published in the past five years have been cited in the JCR year. It is caclulated by dividing the number of citations in the JCR year by the total number of articles published in the five previous years.

Citable Items

Citable items are those items that comprise the figure in the denominator of the Journal Impact Factor calculation. These items are those identified in the Web of Science as an article, review or proceedings paper and are considered the substantive articles that contribute to the body of scholarship in a particular research field and those most likely to be cited by other articles. Other forms of journal content, such as editorial materials, letters, and meetings abstracts, are not considered as citable items.

Aggregate Cited Half-Life

Aggregate Cited half-life is the median age, in years, of items in any journal in the category that were cited during the JCR year.

Citing Half-Life Data

The Citing Half-Life is the median age of the citations produced by a journal during the JCR year. A citation's age is equal to the publication year of the citing item (i.e., JCR year) minus the publication year of the cited item. By definition, half of a journal's outbound citations are to items published before the Citing Half-Life, and half are to items published after the Citing Half-Life. In the example histogram, a journal produced 10,500 citations during the JCR year (the JCR year is the period marked "0-1"; the year prior is the period marked "1-2"; and so on). The Citing Half-Life is 4.6, meaning the median age of the citations is 4.6 years old. Half of the citations are to items that are newer than 4.6 years old (orange zone), and half are to items that are older (blue zone).

The maximum Citing Half-Life that will be displayed in the Key Indicators table is 10 years. Any value greater than this will be displayed as ">10.0".

Eigenfactor® Score

The Eigenfactor Score calculation is based on the number of times articles from the journal published in the past five years have been cited in the JCR year, but it also considers which journals have contributed these citations so that highly cited journals will influence the network more than lesser cited journals. References from one article in a journal to another article from the same journal are removed, so that Eigenfactor Scores are not influenced by journal self-citation.

Article Influence Score

The Article Influence Score determines the average influence of a journal's articles over the first five years after publication. It is calculated by multiplying the Eigenfactor Score by 0.01 and dividing by the number of articles in the journal, normalized as a fraction of all articles in all publications. This measure is roughly analogous to the 5-Year Journal Impact Factor in that it is a ratio of a journal's citation influence to the size of the journal's article contribution over a period of five years.

The equation is as follows: $\frac{0.01 * EigenFactor\ Score}{X}$ *where* X = *5-year Journal Article Count divided by the 5-year Article Count from All Journals.*

The mean Article Influence Score for each article is 1.00. A score greater than 1.00 indicates that each article in the journal has above-average influence. A score less than 1.00 indicates that each article in the journal has below-average influence.

% Articles in Citable Items

The % of Articles in Citable Items emphasizes a journal's original research by calculating the percentage of articles that count toward the total Citable Items. For example, in 2013, Nature has 829 articles and 28 reviews for a total Citable Items of 857.

96% of the Citable Items are original research.

Indicadores SJR

Descripción de los indicadores SJR.

SJR (SCImago Journal Rank) indicator

It expresses the average number of weighted citations received in the selected year by the documents published in the selected journal in the three previous years, --i.e. weighted citations received in year X to documents published in the journal in years X-1, X-2 and X-3.

H Index

The h index expresses the journal's number of articles (h) that have received at least h citations. It quantifies both journal scientific productivity and scientific impact and it is also applicable to scientists, countries, etc.

Total Documents

Output of the selected period. All types of documents are considered, including citable and non-citable documents.

Total Documents (3years)

Published documents in the three previous years (selected year documents are excluded), i.e.when the year X is selected, then X-1, X-2 and X-3 published documents are retrieved. All types of documents are considered, including citable and non citable documents.

Citable Documents (3 years)

Number of citable documents published by a journal in the three previous years (selected year documents are excluded). Exclusively articles, reviews and conference papers are considered.

Total Cites (3years)

Number of citations received in the seleted year by a journal to the documents published in the three previous years, --i.e. citations received in year X to documents published in years X-1, X-2 and X-3. All types of documents are considered.

Cites per Document (2 years)

Average citations per document in a 2 year period. It is computed considering the number of citations received by a journal in the current year to the documents published in the two previous years, --i.e. citations received in year X to documents published in years X-1 and X-2.

Cites per Document (3 years)

Average citations per document in a 3 year period. It is computed considering the number of citations received by a journal in the current year to the documents published in the three previous years, --i.e. citations received in year X to documents published in years X-1, X-2 and X-3.

Cites per Document (4 years)

Average citations per document in a 4 year period. It is computed considering the number of citations received by a journal in the current year to the documents published in the four previous years, --i.e. citations received in year X to documents published in years X-1, X-2, X-3 and X-4.

Self Cites

Number of journal's self-citations in the seleted year to its own documents published in the three previous years, --i.e. self-citations in year X to documents published in years X-1, X-2 and X-3. All types of documents are considered.

Cited Documents

Number of documents cited at least once in the three previous years, --i.e. years X-1, X-2 and X-3

Uncited Documents

Number of uncited documents in the three previous years, --i.e. years X-1, X-2 and X-3

Total References

It includes all the bibliographical references in a journal in the selected period.

References per Document

Average number of references per document in the selected year.

% International Collaboration

Document ratio whose affiliation includes more than one country address.

Apéndice B. Manuales de descarga de información.

Se encuentran todos los manuales de usuario para la descarga de información de las fuentes del proyecto.

Publicaciones científicas:

Scopus

Ingresar al link https://www.scopus.com y buscar el autor, Documento, etc.:

Para esta descarga se va a hacer una búsqueda por *Affiliation name*: "Universidad de Zaragoza":

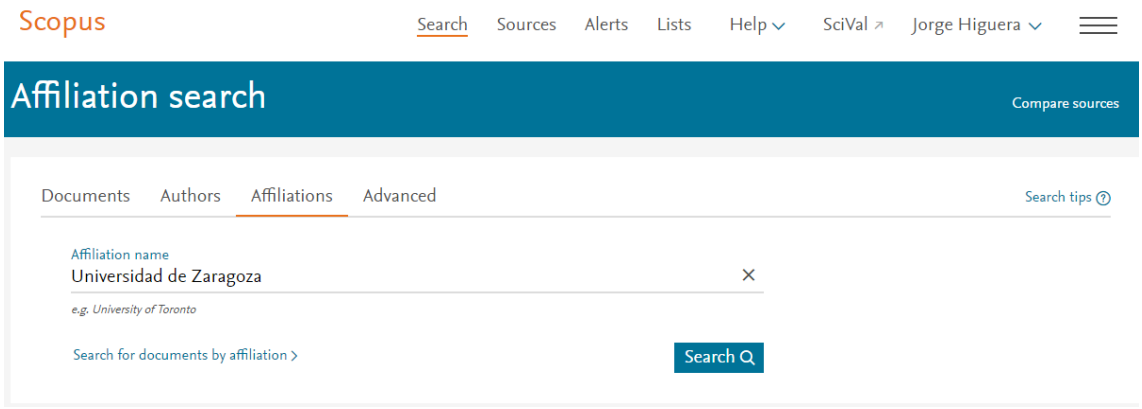


Figura 33:Inicio de Scopus

Al buscar, seleccionar el resultado deseado:

Figura 34:Buscar Scopus

Hacer click en el total de documentos "29.339":

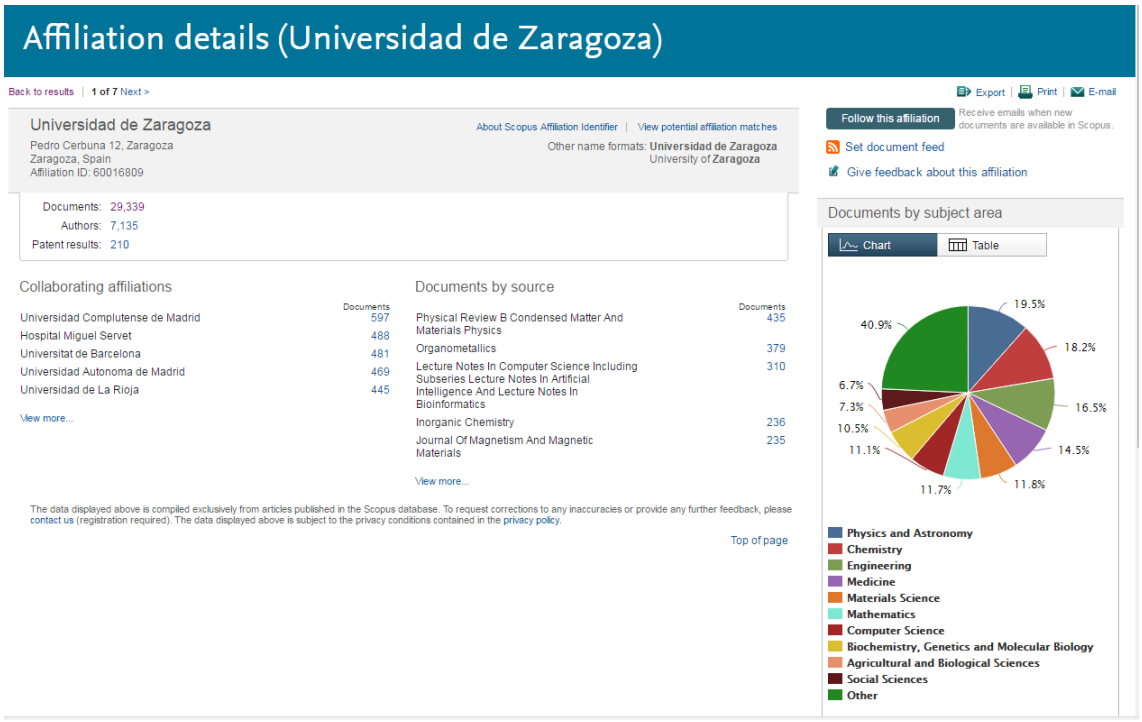


Figura 35: Total documentos Scopus

Seleccionar todos los documentos y desplegar las opciones de exportación:

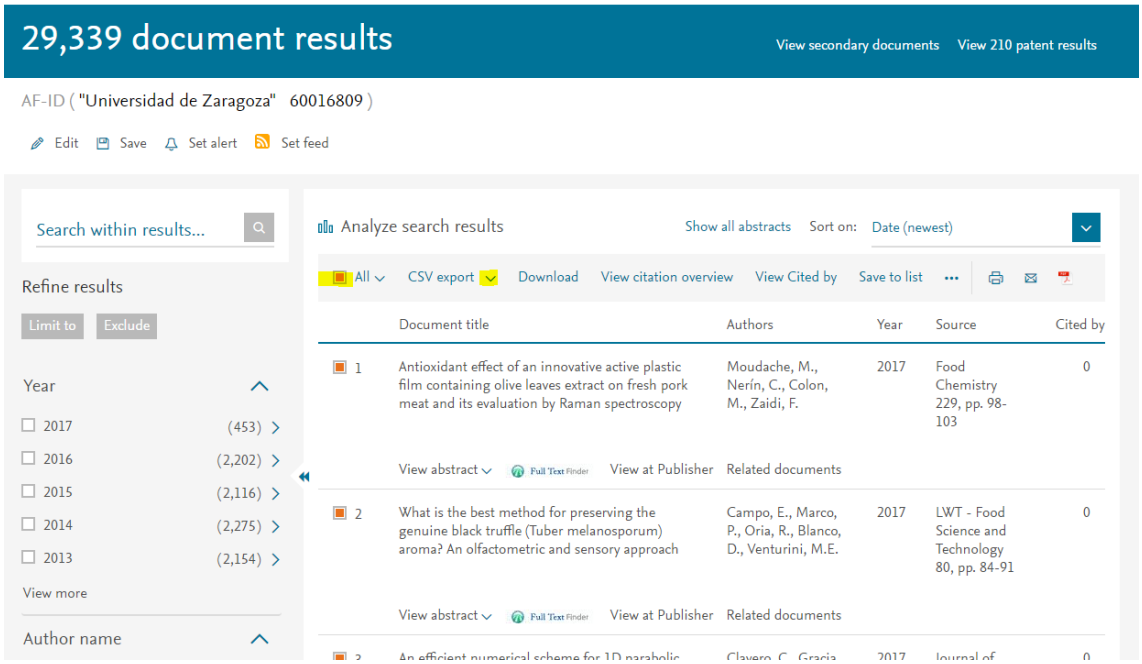


Figura 36: Seleccionar descarga Scopus

Seleccionar todas las opciones de "*Citation information*", "*Bibliographical information*" y "*Abstract and Keywords*" y hacer click en exportar:

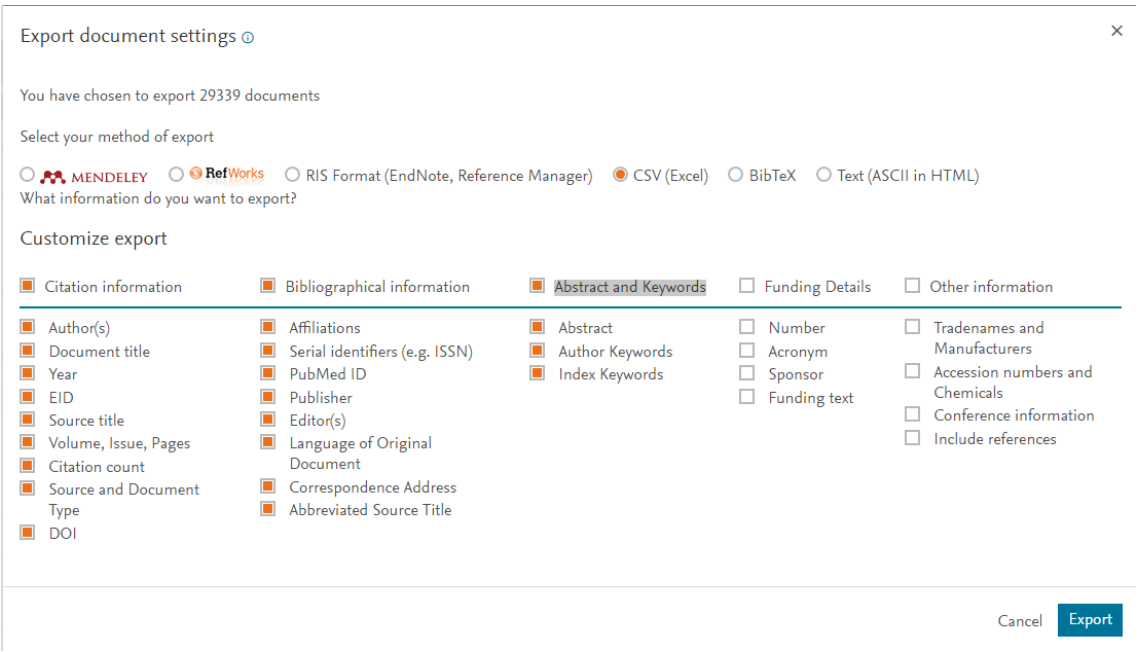


Figura 37: Seleccionar contenido de descarga Scopus

El fichero se exportará, si el contenido supera los 20.000 registros, llegará un email con el link de descarga:

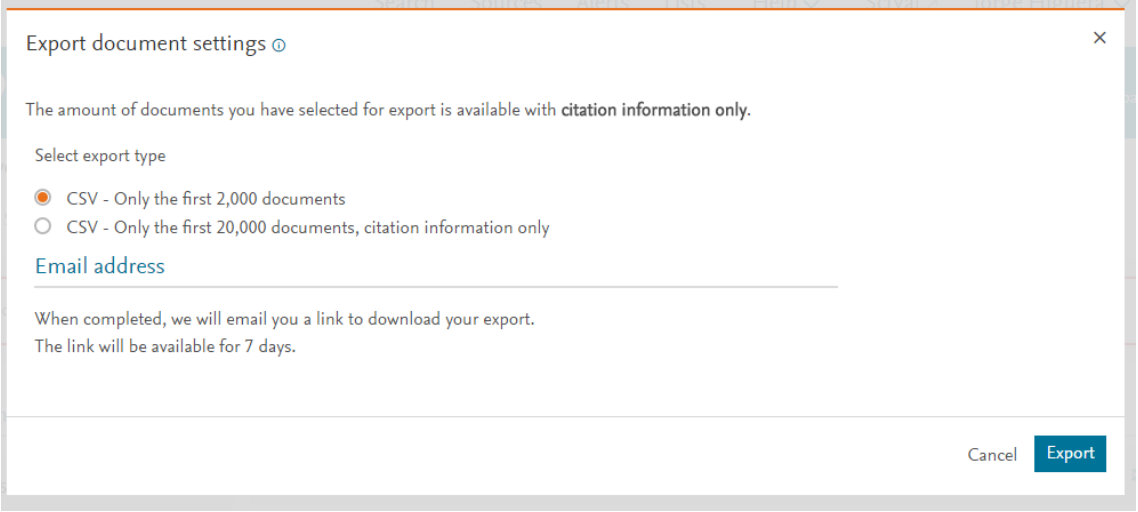


Figura 38: Seleccionar formato de descarga Scopus

Google Scholar

Ingresar al link https://scholar.google.es/ y buscar el autor, artículo, *paper*, etc:

Figura 39: Inicio Google Scholar

Seleccionar el autor de un artículo escrito por dicho autor:

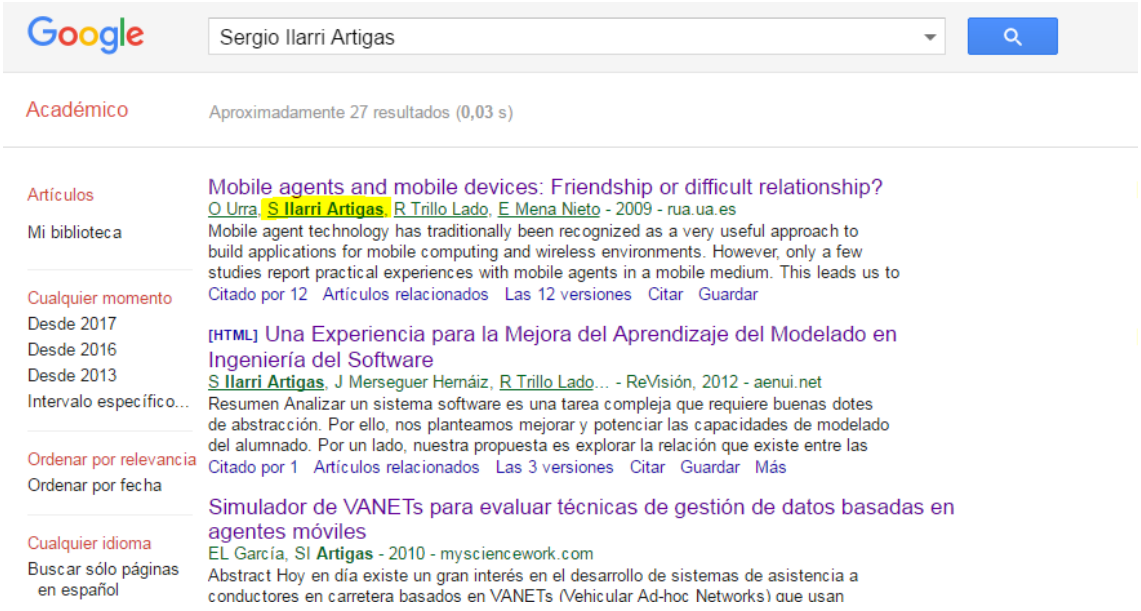


Figura 40: Buscar Google Scholar

En el listado de artículos del autor, se muestra la información del nombre del artículo, las citas y el año; seleccionar el artículo de interés:

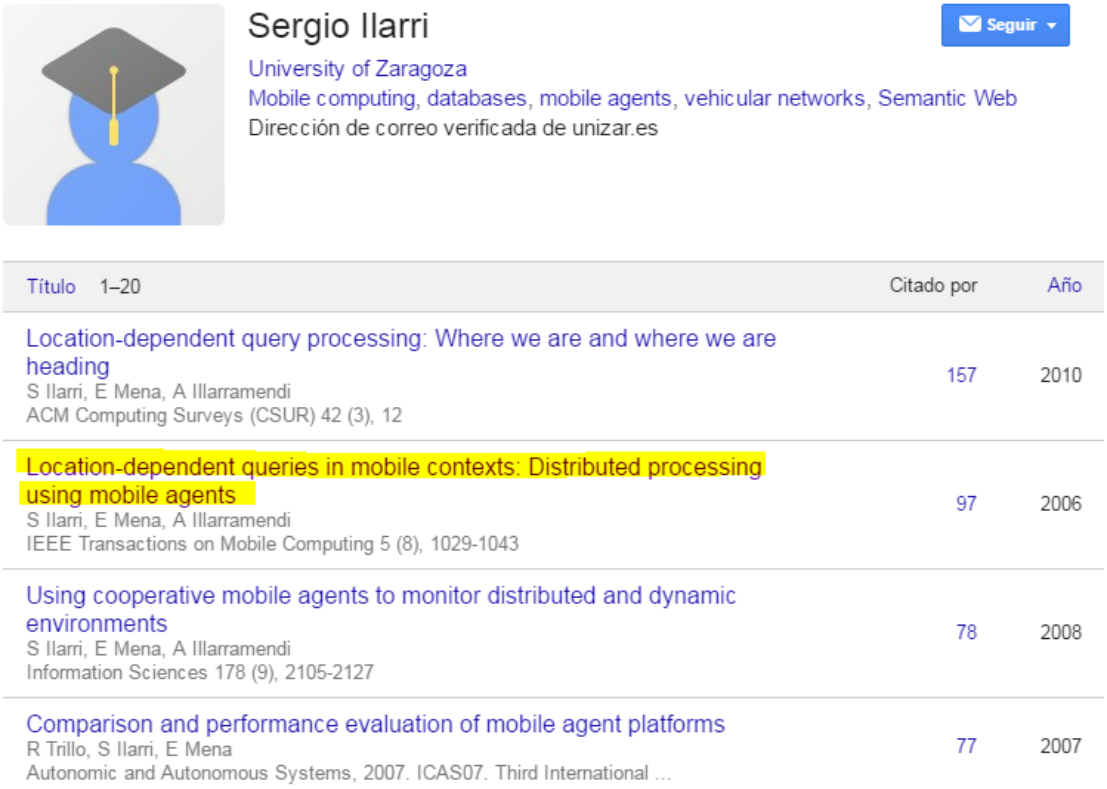


Figura 41: Seleccionar descarga Google Scholar

Una vez dentro de la información del artículo seleccionamos la opción de "Exportar" y la opción de "RefMan":

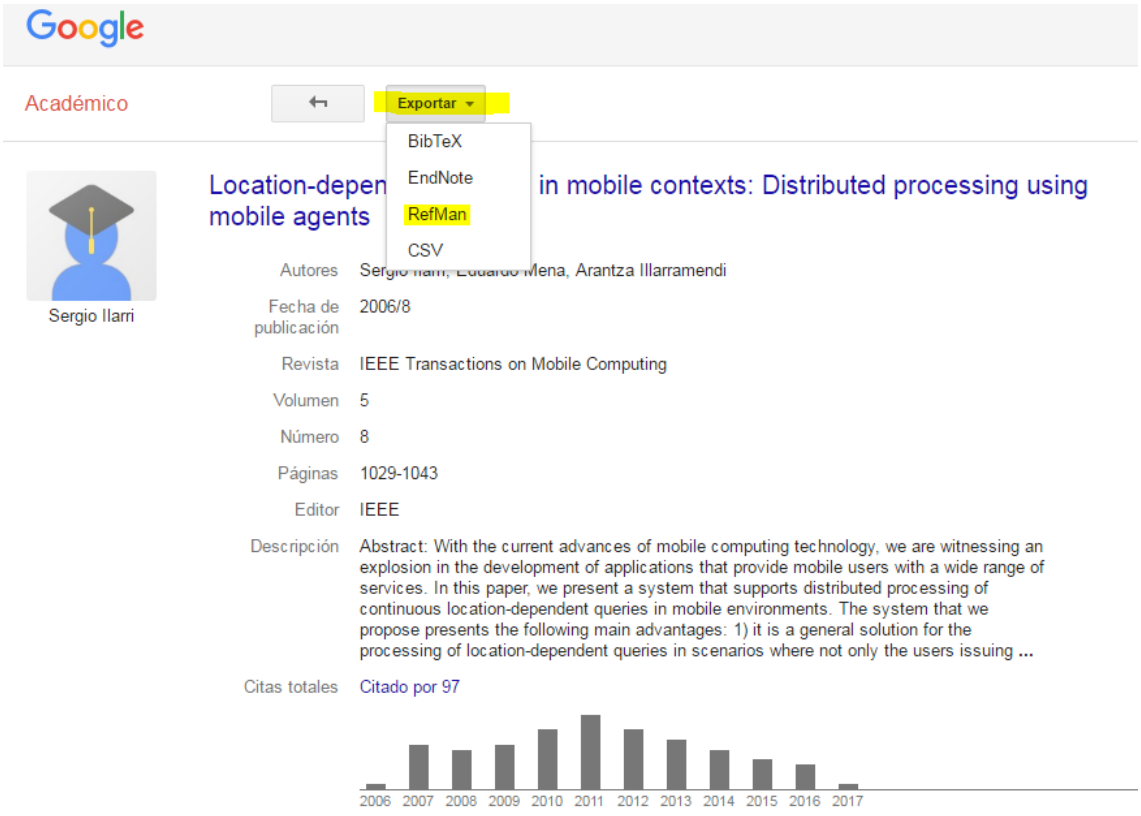


Figura 42: Seleccionar formato de descarga Google Scholar

Web of Science

Ingresar al link http://wos.fecyt.es y seleccionamos nuestra búsqueda, para el ejemplo se ha seleccionado autor:

Figura 43: Inicio de Web of Science

Una vez tenemos la lista de artículos, tener en cuenta el número de resultados, hacer click en "Crear informe de citas":

Figura 44: Listado artículos Web of Science

La web crea un informe, ir al final del mismo y hacer click en "Guardar en archivo de Excel":

Figura 45: Guardar descarga de Web of Science

Seleccionar la cantidad total de registros, para este ejemplo 58 (el número de artículos) y hacer click en enviar:

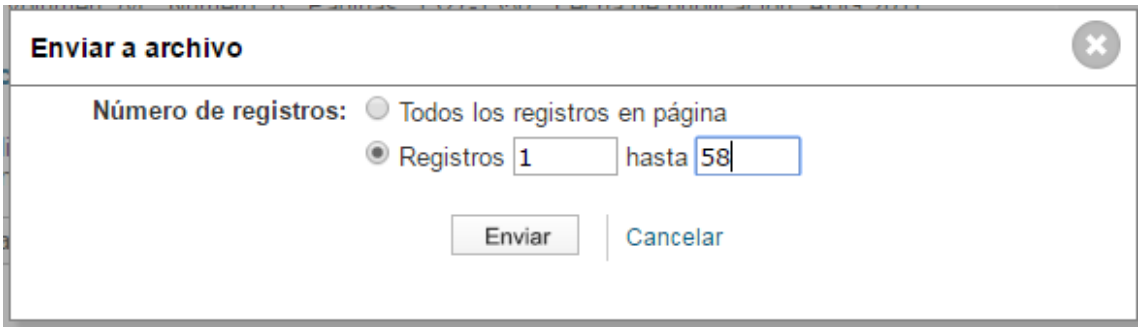


Figura 46: Seleccionar la cantidad de registros Web of Science

El archivo ha sido descargado.

Publish or Perish

Science Direct

Ingresar al link http://www.sciencedirect.com/ y buscar por palabras clave, autor, revista, volumen, etc:

Para esta descarga se va a hacer una búsqueda por autor Ilarri:

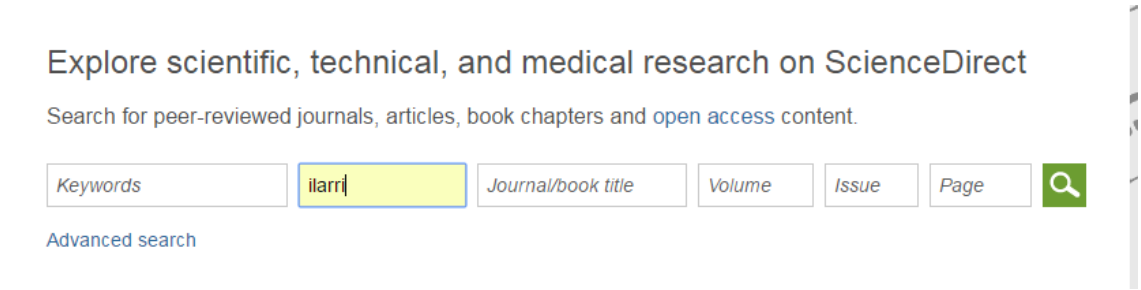


Figura 47: Inicio de Science Direct

Al buscar, arroja un listado de los artículos que coinciden, sin elegir ninguno, hacer click en "*Export*" -> la opción de "RIS (for EndNote, Referencia Manager, ProCite)", en contenido seleccionar "*Citations and Abstracts*" y exportar:

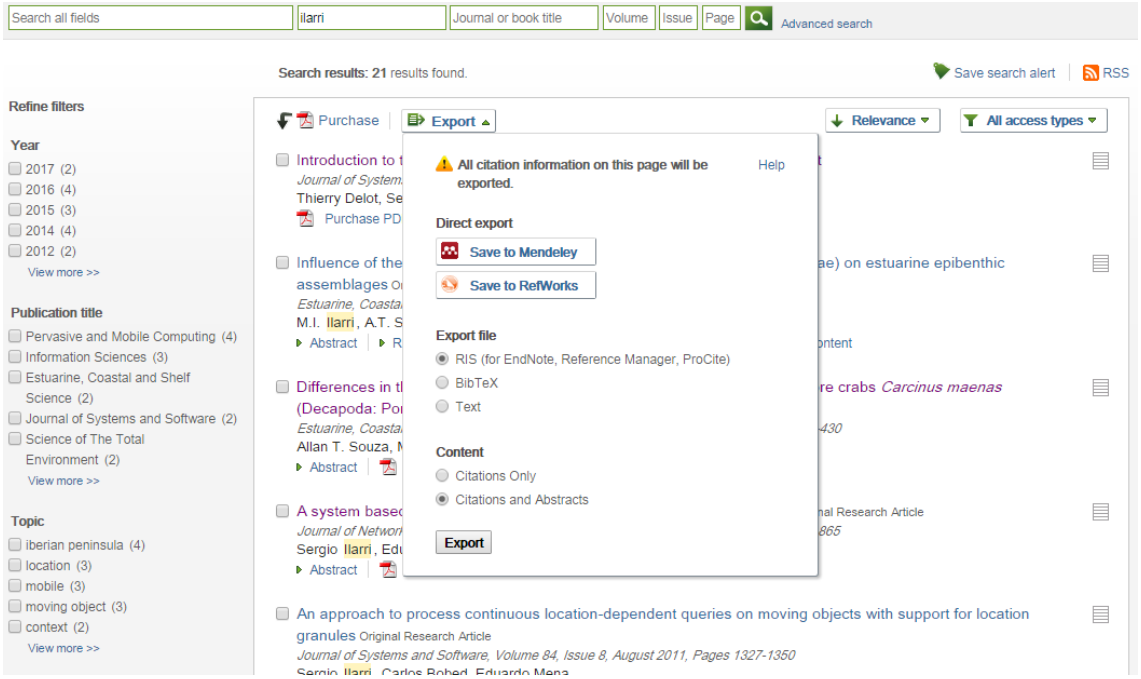


Figura 48: Exportar archivo de Science Direct

Ya se ha exportado el archivo, ponerlo en la ruta de archivos planos de carga.

Lugares de publicación:

CORE Conference Ranking

Ingresar al link http://portal.core.edu.au/conf-ranks/ dejar la primera casilla en blanco y en *Search by* y en *Source*, elegir "*All*" para consultar todos los registros. Hacer Click en *Search*:

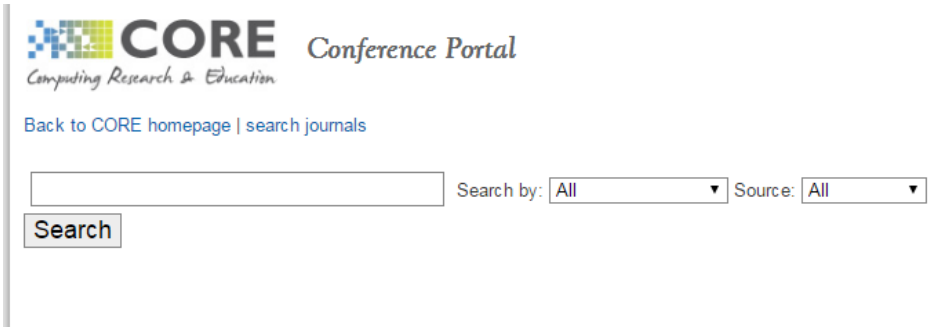


Figura 49: Inicio Portal CORE Conference

A continuación, nos aparece una lista con todos los registros:

Al final toca descargar uno año por año puesto que solo muestra el ultimo ranking de cada artículo. Si ha cambiado de rango en dos años, solo muestra el último año.

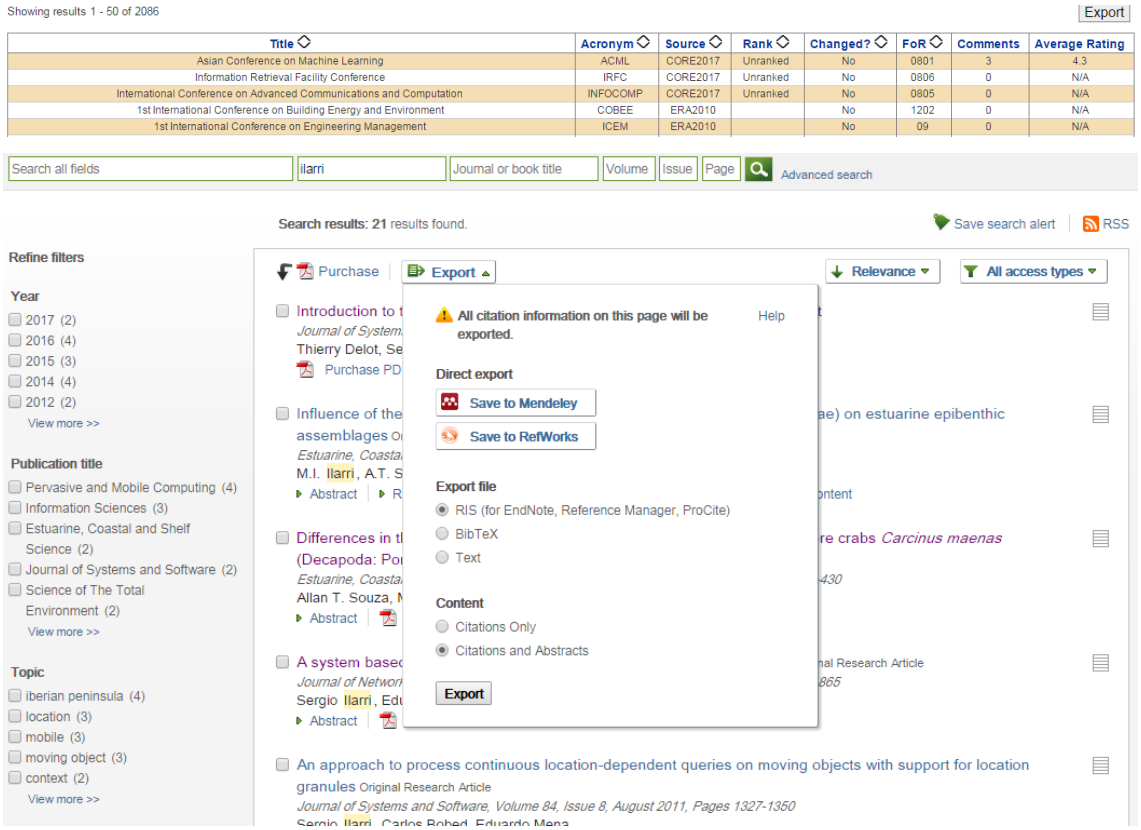


Figura 50: Descarga de archivo portal CORE Conference

Ya se ha exportado el archivo, ponerlo en la ruta de archivos planos de carga.

CORE Journal Ranking

Ingresar al link http://portal.core.edu.au/jnl-ranks/ dejar la primera casilla en blanco, en *Search by* elegir "*All*" y en *Source* elegir "ERA2010" que al momento es la única opción. Hacer Click en *Search*:

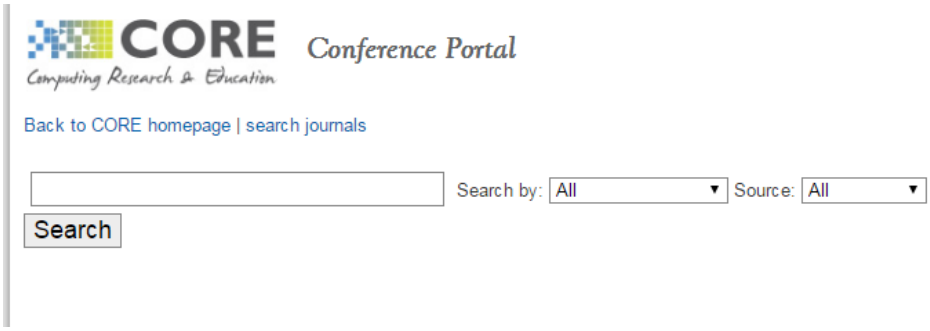
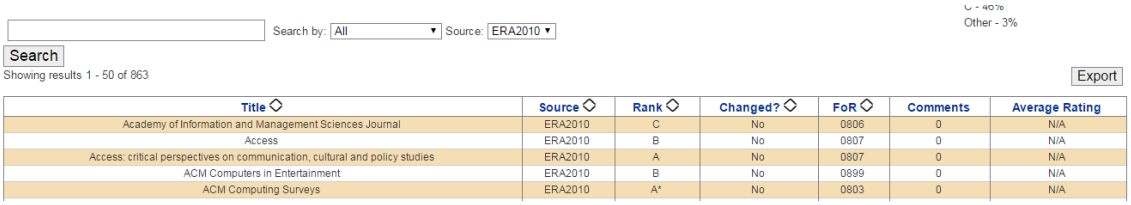


Figura 51: Inicio portal CORE Journal

A continuación, nos aparece una lista con todos los registros:

Al final toca descargar uno año por año puesto que solo muestra el ultimo ranking de cada artículo. Si ha cambiado de rango en dos años, solo muestra el último año.



Title	Source	Rank	Changed?	FoR	Comments	Average Rating
Academy of Information and Management Sciences Journal	ERA2010	C	No	0806	0	N/A
Access	ERA2010	B	No	0807	0	N/A
Access: critical perspectives on communication, cultural and policy studies	ERA2010	A	No	0807	0	N/A
ACM Computers in Entertainment	ERA2010	B	No	0899	0	N/A
ACM Computing Surveys	ERA2010	A*	No	0803	0	N/A

Figura 52: Descarga de archivo portal CORE Journal

Hacer Click en *Export* y el archivo se ha exportado en formato csv.

Poner el archivo en la ruta de fuentes del proyecto.

JCR Ranking

SJR Journal Ranking

Ingresar al link http://www.scimagojr.com/journalrank.php

En la caja de las áreas seleccionar "*Computer science*", (luego repetir los pasos con *Engineering*, las dos áreas que serán utilizadas en el proyecto, el usuario podrá descargar las áreas que desee.

Dejar "*All Subcategories*", "*All regions / countries*", "*All types*".

Y para los años ir seleccionando uno a uno los años de interés:

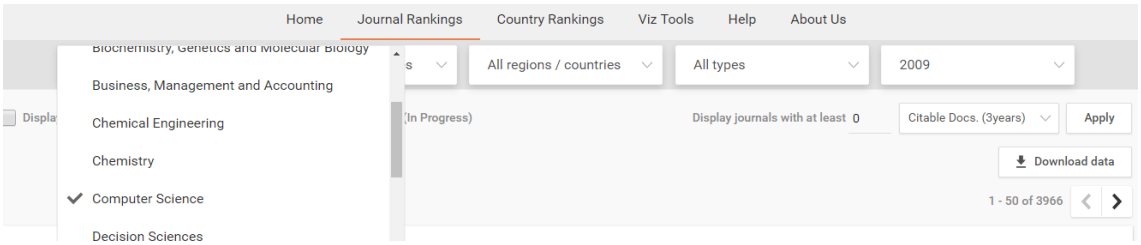


Figura 53: Seleccionar búsqueda SJR

Una vez seleccionado los filtros, hacer click en "*Download data*":

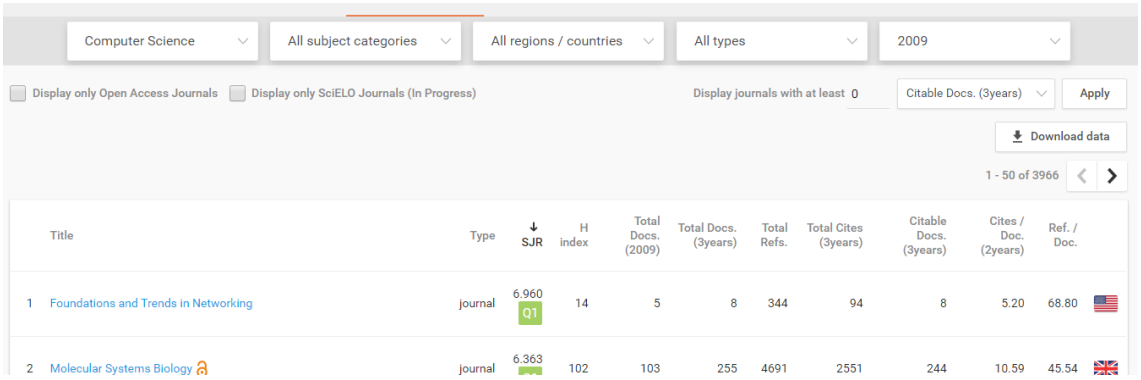


Figura 54: Descargar datos SJR

Se descarga en la carpeta de descargas un fichero llamado "scimagojr"; este nombre debe ser modificado de la siguiente manera:

4 primeras letras del área seleccionada + "_" + "scimagojr" + "_" + AAAA (año seleccionado)

Al final para el área *Computer Science* quedará cada nombre del archivo de la siguiente manera:

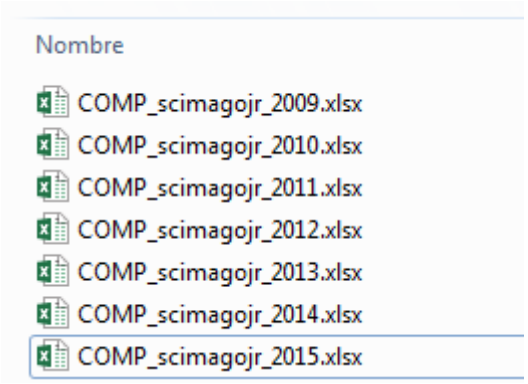


Figura 55: Ejemplo ficheros SJR

Una vez los nombres estén de acuerdo al formato, enviar los documentos a la carpeta de fuentes del proyecto.

GGS Ranking

Ingresar al Link y hacer click en "*Search*":

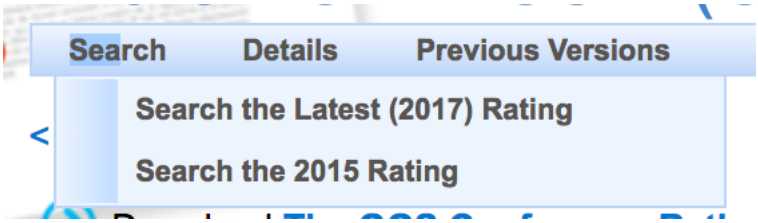


Figura 56: Buscar GGS Ranking

Estos indicadores son relativamente nuevos, así que solo hay disponibles para el año 2015 y 2017 (último), hacer click en el año a descargar.

Directamente aparece la opción de descargar todo el fichero en formato Excel:

Figura 57: Descargar GGS Ranking

Al hacer click, directamente descarga el archivo.

Ponerlo en la ruta de planos TFM.

Apéndice C. Manual configuración Instancia base de datos.

Configuración (por consola) de la instancia MyInstance de la base de datos.

Después de la instalación de SQL Server.

Abrir la consola de Windows (CMD) y ejecutar el comando [sqllocaldb create "MyInstance"], pulsar Enter:

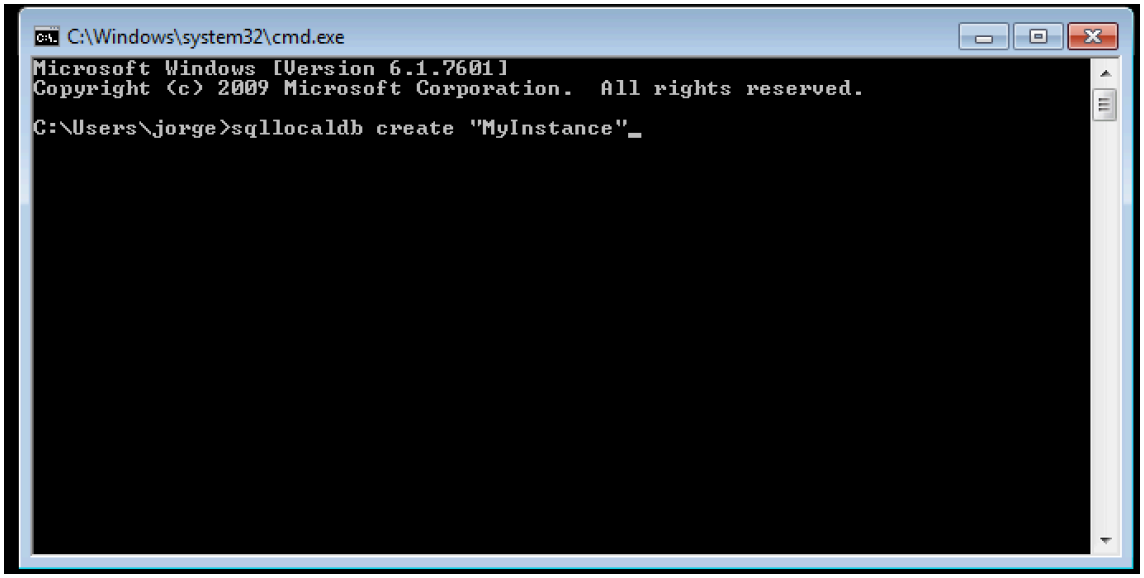


Figura 58: Consola

Una vez creado, podemos verificar la creación ejecutando en la consola el comando [sqllocaldb info]; podemos ver la información de la nueva instancia creada:

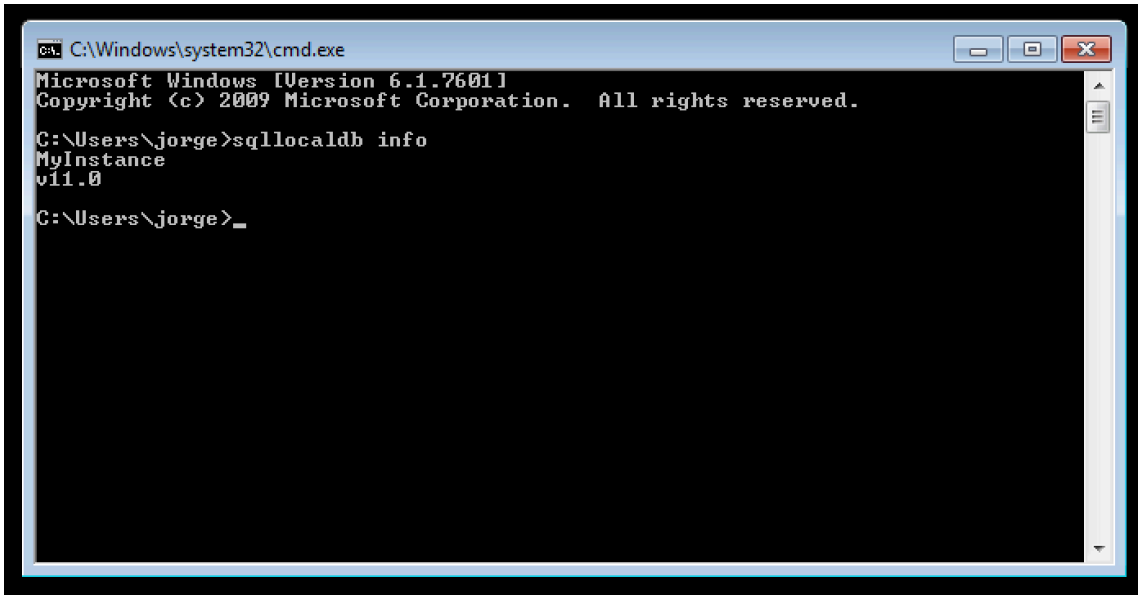


Figura 59: Verificar nueva instancia por consola

Una vez creada y validada la instancia, iniciar el SQL Management introduciendo los datos de acceso:

Figura 60: inicio de acceso base de datos

Ingresar, y verificar conexión.

Apéndice D. Manual de Restauración de las bases de datos.

En este manual se restaura la base de datos DATAWAREHOUSE, pero son tres bases de datos a restaurar para poner en marcha el proyecto: DataQualityService, DATAWAREHOUSE, db_Config y db_STAGING. Todas se restauran de la misma manera a partir de su *backup* (archivo .bak).

1. Abrir el SQL Server Management Studio:

Figura 61: Abrir SQL Management Studio

2. Hacer click derecho en "*Databases*" -> "*Restore Database*":

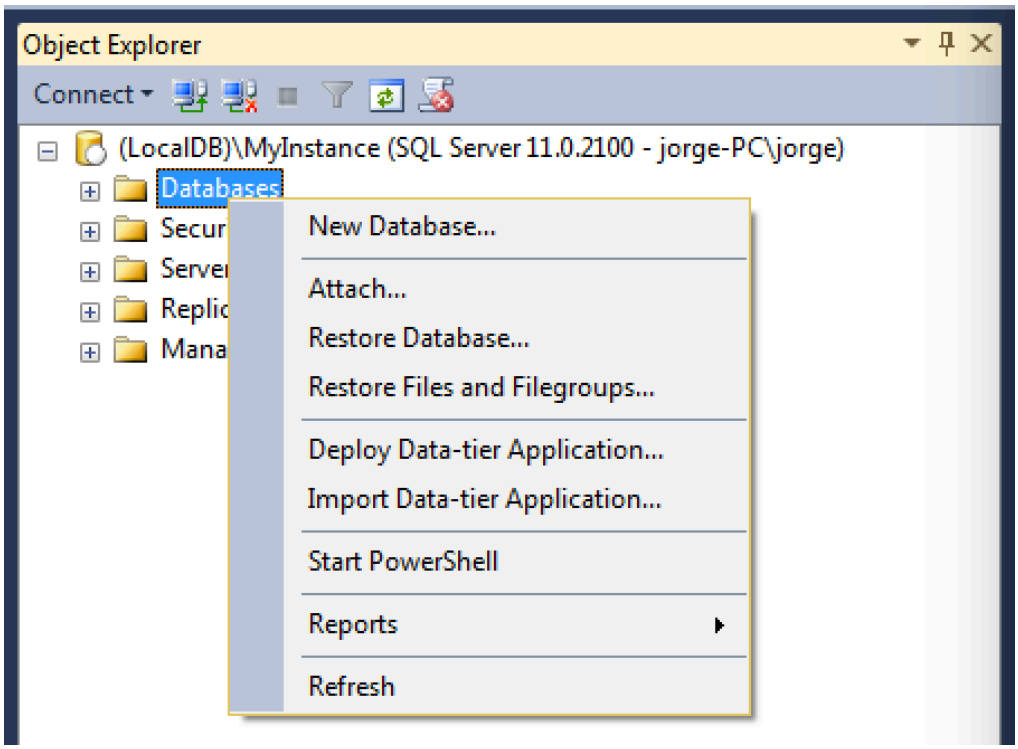


Figura 62: Restaurar base de datos

3. Seleccionar la opción de "*Device*" y buscar el directorio donde esté el *BackUp* de la base de datos "DATAKAREHOUSE.bak". hacer click en OK y luego, en OK de nuevo:

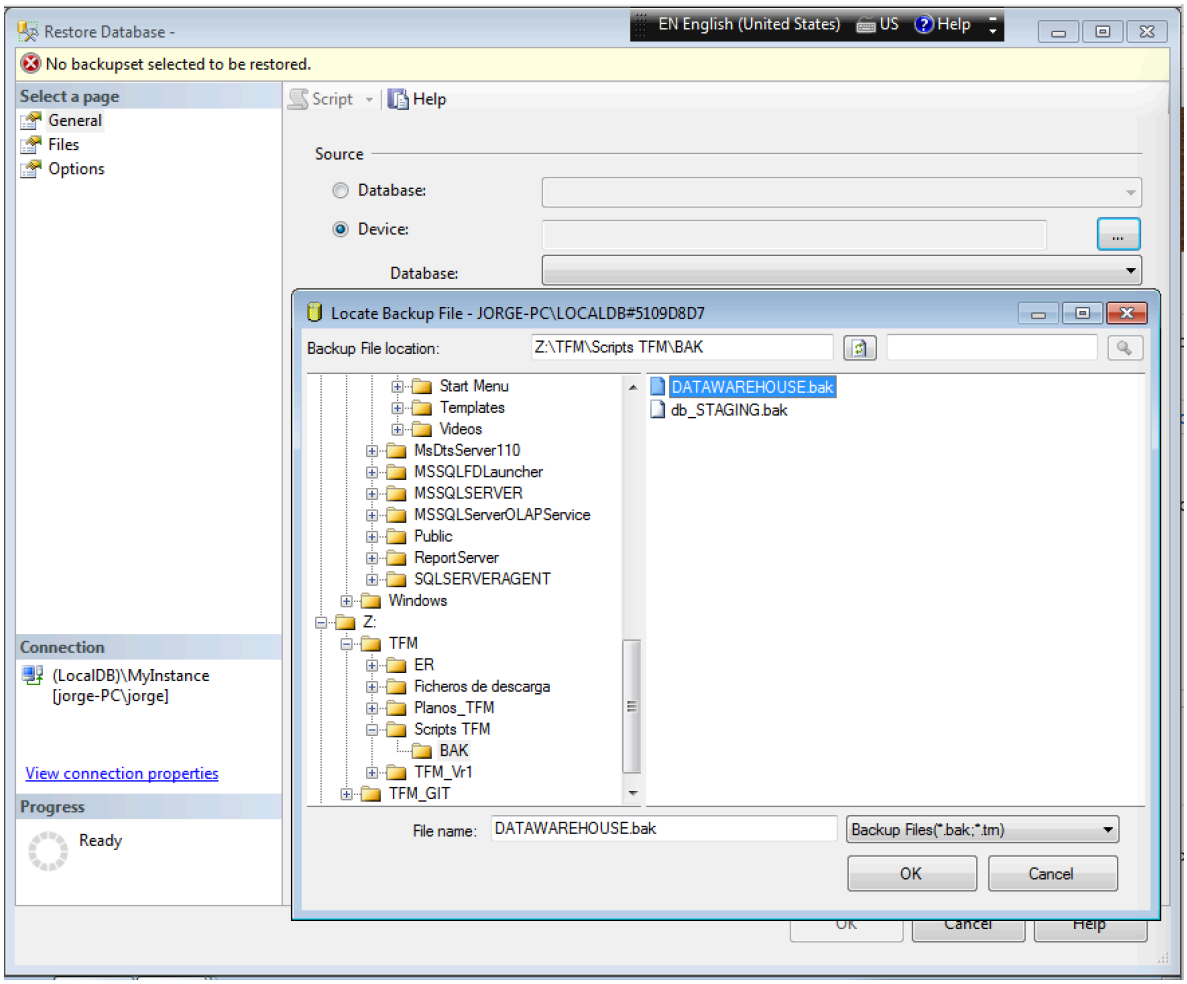


Figura 63: Restaurar base de datos II

4. El programa muestra un aviso de restablecimiento correcto, hacer click en OK:

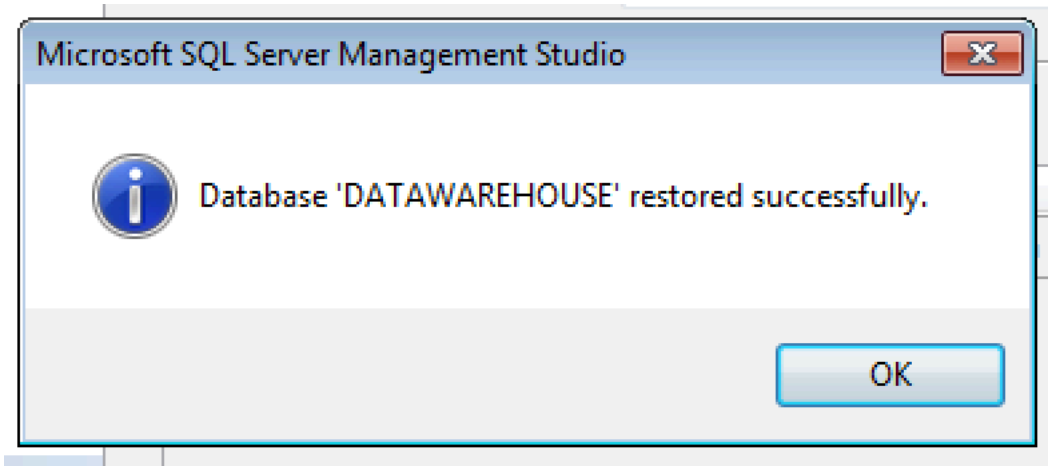


Figura 64: Restauración correcta

5. Se visualiza en la parte izquierda en las "*Databases*" la nueva base de datos incorporada:

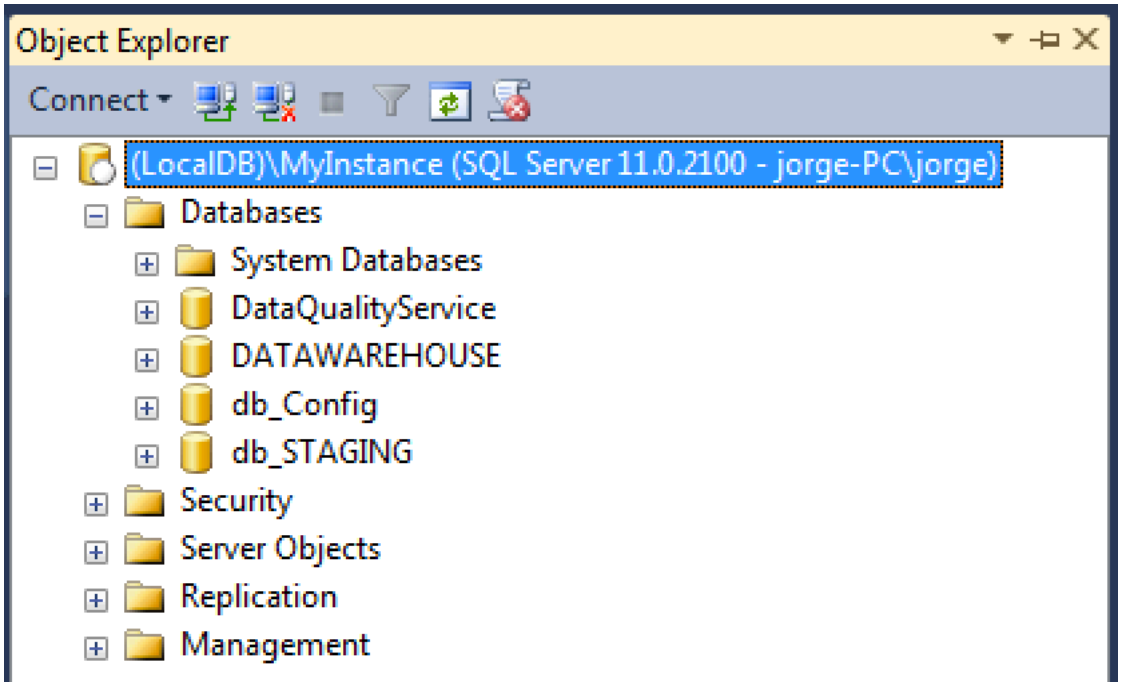


Figura 65: Listado de bases de datos

Apéndice E. Manual de configuración de rutas del proceso.

- Configurar la ruta de los ficheros planos:
  Una vez se ha ingresado a la instancia de Base de datos, ejecutar la consulta

  UPDATE db_Config.dbo.tbl_config

  SET desc_configuracion = 'Ruta_ Directorio'

  where id_configuracion = 1

  Siendo 'Ruta_ Directorio' la ruta donde se encuentra el directorio de los ficheros planos. Por ejemplo 'C:\Planos_TFM\' importante que tenga la barra del final:

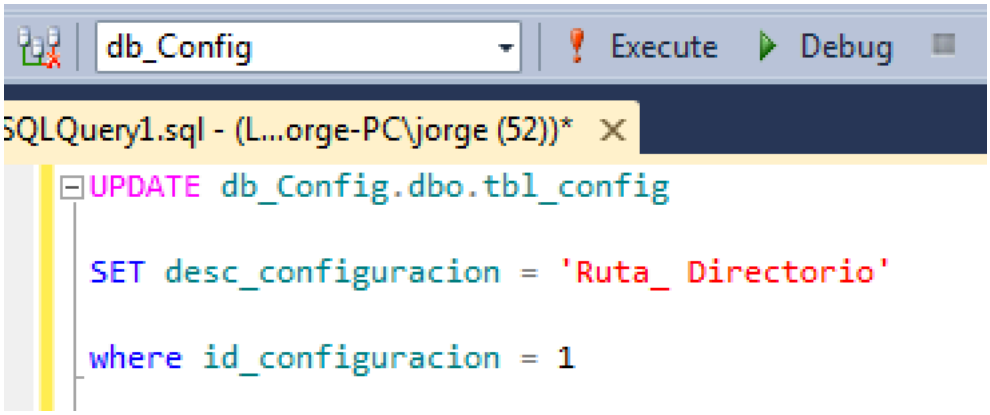


Figura 66: Ejecución de comando en la base de datos

Apéndice F. Script carga estrella Impacto Artículo.

```
 --antes se debe ejecutar el cursor de la creación de las columnas de autor.
declare @titulo as varchar(max)
declare @lugarPublicacion   as varchar(max)
declare @DOI    as varchar(max)
declare @ISSN   as varchar(max)
declare @Issue  as varchar(max)
declare @t2 as varchar(max)
declare @conferencia as varchar(max)
declare @autoríaCorpotariva as varchar(max)
declare @editores   as varchar(max)
declare @año    as varchar(max)
declare @Volumen    as varchar(max)
declare @paginaInicio   as varchar(max)
declare @paginaFinal as varchar(max)
declare @paginas    as varchar(max)
declare @publicador as varchar(max)
declare @autores    as varchar(max)
declare @citas_Publish_2    as varchar(max)
declare @citas_ScienceDirect    as varchar(max)
declare @citas_webOf_science_2  as varchar(max)
declare @citas_googleScholar_2  as varchar(max)
declare @citas_Scopus_2 as varchar(max)
declare @nuevoArticulo as int
declare @nuevoDatoArticulo as int
declare @nuevoGrupoAutores int
declare @nuevoLugarPublicacion int
declare @nuevoGrupoCore int
declare @id_autor int
declare @nombreAutor varchar (100)
declare @flag as varchar (10)
declare @flag_autor as varchar (10)
declare @cadena_insert_autor as varchar(255)
--Asignación de valores a los flag
  set @flag = 1
  set @flag_autor = 1
--Dejar las tablas vacías.
truncate table DATAWAREHOUSE.dbo.Impacto_Articulo
truncate table DATAWAREHOUSE.dbo.Dim Datos Articulo
```

```
truncate table DATAWAREHOUSE.dbo.Grupo_Autores
truncate table DATAWAREHOUSE.dbo.GrupoAutores_Autor
truncate table DATAWAREHOUSE.dbo.Nombre_Lugar_Publicacion
truncate table DATAWAREHOUSE.dbo.Impacto_Lugar_publicacion
--Declarar el cursor
DECLARE cursor_DWH CURSOR FOR
--consultar uno a uno los articulos.
    select  titulo, isnull(lugarPublicacion, 'Sin info') lugarPublicacion , max(DOI)
    DOI, max (ISSN) ISSN, max(Issue) Issue, max(t2) t2, max(conferencia) conferencia,
    max(autoríaCorpotariva) autoríaCorpotariva, max(editores) editores, max(año) año,
    max(Volumen) Volumen, max(paginaInicio) paginaInicio,   max(paginaFinal) paginaFinal,
    max(paginas) paginas,max(publicador) publicador,max(autores) autores,
    max(citas_Publish_2) citas_Publish_2, max(citas_ScienceDirect) citas_ScienceDirect,
    max(citas_webOf_science_2) citas_webOf_science_2,
    max(citas_googleScholar_2) citas_googleScholar_2,max(citas_Scopus_2) citas_Scopus_2
    from db_STAGING.dbo.[A_Aux_unificacion_articulos]
    --Where lugarPublicacion = 'Asia Information Retrieval Symposium' --> pruebas
    where lugarPublicacion is not null
    group by titulo,lugarPublicacion
        --se abre el cursor
        OPEN cursor_DWH;
        --Asignación del resultado de las consultas en las variables
        FETCH NEXT FROM cursor_DWH INTO @titulo ,@lugarPublicacion  ,@DOI   ,@ISSN  ,@Issue
        ,@t2    ,@conferencia   ,@autoríaCorpotariva    ,@editores  ,@año   ,@Volumen
        ,@paginaInicio  ,@paginaFinal ,@paginas ,@publicador,@autores   ,@citas_Publish_2
        ,@citas_ScienceDirect,@citas_webOf_science_2,@citas_googleScholar_2 ,@citas_Scopus_2
        --operación a realizar
         WHILE @@FETCH_STATUS = 0
         BEGIN
            --print @lugarPublicacion
            --print @titulo

            --AQUÍ HACE LA OPERACIÓN POR CADA REGISTRO, EN ESTE CASO LO METO EN UNA TABLA
            --1 insertar nuevo  id_articulo en Impacto_Articulo y capturarlo
            --insertar demás datos
            set @nuevoArticulo = (select isnull(max(id_articulo),0) + 1
            from  DATAWAREHOUSE.dbo.Impacto_Articulo)
            insert into DATAWAREHOUSE.dbo.Impacto_Articulo (id_articulo, citas_publishOrPerish,

           citas_scopus,citas_googleScholar,citas_webOfScience,citas_ScienceDirect) values
           (@nuevoArticulo, @citas_Publish_2   ,@citas_Scopus_2,@citas_googleScholar_2,
           @citas_webOf_science_2,@citas_ScienceDirect)

           --2 insertar id_datos_articulo en Dim_Datos_Articulo
           --insertar valores de Dim_Datos_Articulo
           set @nuevoDatoArticulo = (select isnull(max(id_datos_articulo),0) + 1
           from  DATAWAREHOUSE.dbo.Dim_Datos_Articulo)
           insert into DATAWAREHOUSE.dbo.Dim_Datos_Articulo (id_datos_articulo,titulo,año_articulo,
           volumen,Pagina_inicio,pagina_fin,DOI,edicion, paginas, ISSN) values
           (@nuevoArticulo,@titulo, @año,@Volumen,@paginaInicio    ,@paginaFinal,
           @DOI,@Issue, @paginas, @ISSN)
           --actualizar en Impacto_Articulo el Id_datos_articulo
           update DATAWAREHOUSE.dbo.Impacto_Articulo
           set id_datos_articulo = @nuevoDatoArticulo
           where id_articulo = @nuevoArticulo

           --3 insertar nuevo id_grupo_autores en Grupo_Autores
           set @nuevoGrupoAutores = (select isnull(max(id_grupo_autores),0) + 1
           from  DATAWAREHOUSE.dbo.Grupo_Autores)
           insert into DATAWAREHOUSE.dbo.Grupo_Autores (id_grupo_autores) values
           (@nuevoGrupoAutores)
           --actualizar Impacto_Articulo con el id_grupo_autores
           update DATAWAREHOUSE.dbo.Impacto_Articulo
           set id_grupo_autores = @nuevoGrupoAutores
           where id_articulo = @nuevoArticulo

           --4 insertar en GrupoAutores_Articulo el valor de id_grupo_autores
                        --por cada autor encontrado en los autores del articulo,
                        --crear un registro con el id_grupo_autores y el id_autor
                        DECLARE cursor_Busqueda_Autores_DWH CURSOR FOR

                                select id_autor,nombre_autor
                                from DATAWAREHOUSE.dbo.Dim_Autor

                        OPEN cursor_Busqueda_Autores_DWH;

                        --Asignación del resultado de las consultas en las variables
```

```
                        FETCH NEXT FROM cursor_Busqueda_Autores_DWH INTO @id_autor, @nombreAutor
                        WHILE @@FETCH_STATUS = 0
                        BEGIN
                        --por cada autor encontrado en los autores del articulo,
                        --crear un registro con el id_grupo_autores y el id_autor
                        --SELECT CHARINDEX('ilarri', @document);

                        if (SELECT CHARINDEX(@nombreAutor, @autores COLLATE Latin1_General_CS_AS)) > 0
                        begin
                            insert into DATAWAREHOUSE.dbo.GrupoAutores_Autor (id_grupo_autores,id_autor ) values
                            (@nuevoGrupoAutores, @id_autor)
                        end

                        FETCH NEXT FROM cursor_Busqueda_Autores_DWH INTO @id_autor, @nombreAutor
                        END
                        CLOSE cursor_Busqueda_Autores_DWH;
                        DEALLOCATE cursor_Busqueda_Autores_DWH;

            --6 insertar id_nombre_lugar_publicacion y el nombre del LP
            set @nuevoLugarPublicacion = (select isnull(max(id_nombre_lugar_publicacion),0) + 1 f
            rom  DATAWAREHOUSE.dbo.Nombre_Lugar_Publicacion)
            insert into DATAWAREHOUSE.dbo.Nombre_Lugar_Publicacion (id_nombre_lugar_publicacion,
            desc_nombre_lugar_publicacion) values
            (@nuevoLugarPublicacion, @lugarPublicacion)
            --Actualizar id_nombre_lugar_publicacion en id_impacto_articulo
            update DATAWAREHOUSE.dbo.Impacto_Articulo
            set id_nombre_lugar_publicacion = @nuevoLugarPublicacion
            where id_articulo = @nuevoArticulo

            --7 Insertar id_nombre_lugar_publicacion en Impacto_Lugar_publicacion y id_grupo_core
            --así este grupo_core no sea utilizado
            set @nuevoGrupoCore = (select isnull(max(id_grupo_core),0) + 1
            from  DATAWAREHOUSE.dbo.Impacto_Lugar_Publicacion)
            insert into DATAWAREHOUSE.dbo.Impacto_Lugar_publicacion (id_nombre_lugar_publicacion, id_grupo_core) values
            (@nuevoLugarPublicacion, @nuevoGrupoCore)


/* --Validaciones

select * from DATAWAREHOUSE.dbo.Dim_Datos_Articulo
select * from DATAWAREHOUSE.dbo.Grupo_Autores
select * from DATAWAREHOUSE.dbo.GrupoAutores_Autor
select * from DATAWAREHOUSE.dbo.Nombre_Lugar_Publicacion
select * from DATAWAREHOUSE.dbo.Impacto_Lugar_publicacion
*/


             FETCH NEXT FROM cursor_DWH INTO @titulo    ,@lugarPublicacion  ,@DOI   ,@ISSN  ,@Issue ,@t2
             ,@conferencia  ,@autoríaCorpotariva    ,@editores  ,@año   ,@Volumen   ,@paginaInicio  ,@paginaFinal
            ,@paginas   ,@publicador    ,@autores   ,@citas_Publish_2   ,@citas_ScienceDirect,@citas_webOf_science_2
            ,@citas_googleScholar_2 ,@citas_Scopus_2
            --fin del cursor
        END
 --cerrar cursor
 CLOSE cursor_DWH;
 DEALLOCATE cursor_DWH;
```

Apéndice G. Script carga estrella Lugar de Publicación.

```
 --Antes se debe ejecutar el cursor de la creación de las columnas de autor.
 --declaración de variables
declare @lugar_publicacion as varchar(max)
declare @SJR_Total_Refs  as int
declare @JCR_total_citas     as int
declare @JCR_citing_half_life_2  as varchar(max)
declare @JCR_factor_propio_normalizado_2 as varchar(max)
declare @CORE_Confcomentarios_2  as varchar(max)
declare @SJR     as varchar(max)
declare @SJR_Ref_div_Doc     as varchar(max)
declare @JCR_elementos_citables_2    as varchar(max)
declare @JCR_promedio_percentil_factor_impacto_2 as varchar(max)
declare @CORE_Jour_FoR2_2    as varchar(max)
declare @nuevoArticulo as int
declare @SJR_Total_Docs_3years  as varchar(max)
declare @SJR_area_conocimiento  as varchar(max)
declare @año as int
declare @SJR_Total_Cites_3years  as varchar(max)
declare @JCR_factor_impacto  as varchar(max)
declare @JCR_puntuacion_factor_propio_2  as varchar(max)
declare @CORE_Jour_rango_2  as varchar(max)
declare @CORE_Confpromedio_raiting_2     as varchar(max)
declare @SJR_best_quartile   as varchar(max)
declare @SJR_Country     as varchar(max)
declare @JCR_cited_half_life_2  as varchar(max)
declare @CORE_Jour_FoR3_2    as varchar(max)
declare @SJR_ISSN    as varchar(max)
declare @SJR_Cites_div_Doc_2years    as varchar(max)
declare @SJR_Total_Docs_año as varchar(max)
declare @SJR_tipo as varchar(max)
declare @SJR_Citable_Docs_3years     as varchar(max)
declare @JCR_factor_impacto_sin_citas_propias as varchar(max)
declare @JCR_puntuacion_influencia_articulo_2 as varchar(max)
declare @CORE_Jour_cambia_año_2  as varchar(max)
declare @tipo_LP as varchar(max)
declare @SJR_H_index     as varchar(max)
declare @SJR_Total_Docs_2009     as varchar(max)
declare @JCR_factor_impacto_5anhos_2     as varchar(max)
```

```
declare @CORE_Jour_ISSN_2 as varchar(max)
declare @JCR_porcentaje_elementos_articulos_citables_2 as varchar(max)
declare @CORE_Jour_FoR1_2    as varchar(max)
--el ID del  nombre del lugar de publicacion
declare @id_nombre_lugar_publicacion as int
declare @desc_nombre_lugar_publicacion as varchar (max)
declare @año_articulo   as varchar (max)
declare @GrupoCore as int
declare @nuevoIdDatosCore as int
--Dejar las tablas vacías.
truncate table DATAWAREHOUSE.dbo.Impacto_Revista
truncate table DATAWAREHOUSE.dbo.Impacto_congreso
truncate table Grupo_EdicionesCore_LugarPublicacion
truncate table Dim_Datos_CORE
--declarar el cursor
declare cursor_DWH_LP CURSOR FOR

    --una vez la estrella de impacto articulo esté llena, se prosigue a llenar la estrella de impacto lugar de
    --publicacion
    --se necesita saber:
    --el nombre del lugar de publicación
    --el año del articulo
    select convert(int,NLP.id_nombre_lugar_publicacion) id_nombre_lugar_publicacion, desc_nombre_lugar_publicacion,
    año_articulo
    from DATAWAREHOUSE.dbo.nombre_lugar_publicacion NLP
    inner join DATAWAREHOUSE.dbo.Impacto_Articulo IA on NLP.id_nombre_lugar_publicacion = IA.id_nombre_lugar_publicacion
    inner join DATAWAREHOUSE.dbo.Dim_Datos_Articulo DDA on IA.id_datos_articulo = DDA.id_datos_articulo
    --where desc_nombre_lugar_publicacion = 'Asia Information Retrieval Symposium'
    group by NLP.id_nombre_lugar_publicacion, desc_nombre_lugar_publicacion, año_articulo

        OPEN cursor_DWH_LP;
        --Asignación del resultado de las consultas en las variables
        FETCH NEXT FROM cursor_DWH_LP INTO @id_nombre_lugar_publicacion ,@desc_nombre_lugar_publicacion ,@año_articulo
        WHILE @@FETCH_STATUS = 0
        BEGIN

        --necesitamos otro cursor que recorra uno a uno los resultados y haga lo que tenga que hacer de acuerdo al
        --tipo_LP
```

```
DECLARE cursor_Busqueda_tipo_LP CURSOR FOR

        select lugar_publicacion,convert(int,año) año,max(SJR_tipo) SJR_tipo,   max(SJR_ISSN) SJR_ISSN,
        max(SJR) SJR,   max(SJR_best_quartile) SJR_best_quartile, max(SJR_H_index) SJR_H_index,
        max(SJR_Total_Docs_3years) SJR_Total_Docs_3years,   max(SJR_Total_Refs) SJR_Total_Refs,
        max(SJR_Total_Cites_3years) SJR_Total_Cites_3years, max(SJR_Citable_Docs_3years)
        SJR_Citable_Docs_3years,max([SJR_Cites_/_Doc#_2years]) [SJR_Cites_/_Doc#_2years],
        max([SJR_Ref_/_Doc]) [SJR_Ref_div_Doc], max(SJR_Country) SJR_Country,   max(SJR_Total_Docs_2009)
        SJR_Total_Docs_2009,max(SJR_area_conocimiento) SJR_area_conocimiento,
        max(JCR_total_citas) JCR_total_citas,
        max(JCR_factor_impacto) JCR_factor_impacto, max(JCR_factor_impacto_sin_citas_propias)
        JCR_factor_impacto_sin_citas_propias,max(JCR_factor_impacto_5anhos_2) JCR_factor_impacto_5anhos_2,
        max(JCR_elementos_citables_2) JCR_elementos_citables_2,     max(JCR_cited_half_life_2)
        JCR_cited_half_life_2,max(JCR_citing_half_life_2) JCR_citing_half_life_2,
        max(JCR_puntuacion_factor_propio_2) JCR_puntuacion_factor_propio_2,
        max(JCR_puntuacion_influencia_ariculo_2) JCR_puntuacion_influencia_ariculo_2,
        max(JCR_porcentaje_elementos_articulos_citables_2) JCR_porcentaje_elementos_articulos_citables_2,
        max(JCR_promedio_percentil_factor_impacto_2) JCR_promedio_percentil_factor_impacto_2,
        max(JCR_factor_propio_normalizado_2) JCR_factor_propio_normalizado_2,   max(CORE_Jour_rango_2)
        CORE_Jour_rango_2,  max(CORE_Jour_cambia_año_2) CORE_Jour_cambia_año_2,
        max(CORE_Jour_FoR1_2) CORE_Jour_FoR1_2, max(CORE_Jour_FoR2_2) CORE_Jour_FoR2_2,
        max(CORE_Jour_FoR3_2) CORE_Jour_FoR3_2, max(CORE_Jour_ISSN_2) CORE_Jour_ISSN_2,
        max(CORE_Confcomentarios_2) CORE_Confcomentarios_2, max(CORE_Confpromedio_raiting_2)
        CORE_Confpromedio_raiting_2,max(tipo_LP) tipo_LP
        from db_STAGING.dbo.LP_Aux_unificacion_lugar_publicacion
        where  lugar_publicacion = @desc_nombre_lugar_publicacion --'Asia Information Retrieval Symposium'
        --and tipo_LP in (3,4)
        group by lugar_publicacion, año

OPEN cursor_Busqueda_tipo_LP;
--Asignación del resultado de las consultas en las variables
FETCH NEXT FROM cursor_Busqueda_tipo_LP INTO @lugar_publicacion,    @año,   @SJR_tipo,  @SJR_ISSN   ,@SJR
,@SJR_best_quartile ,@SJR_H_index   ,@SJR_Total_Docs_3years ,@SJR_Total_Refs    ,@SJR_Total_Cites_3years
,@SJR_Citable_Docs_3years   ,@SJR_Cites_div_Doc_2years  ,@SJR_Ref_div_Doc   ,@SJR_Country
,@SJR_Total_Docs_año
,@SJR_area_conocimiento ,@JCR_total_citas   ,@JCR_factor_impacto    ,@JCR_factor_impacto_sin_citas_propias
,@JCR_factor_impacto_5anhos_2   ,@JCR_elementos_citables_2  ,@JCR_cited_half_life_2 ,@JCR_citing_half_life_2
,@JCR_puntuacion_factor_propio_2    ,@JCR_puntuacion_influencia_ariculo_2
,@JCR_porcentaje_elementos_articulos_citables_2 ,@JCR_promedio_percentil_factor_impacto_2
,@JCR_factor_propio_normalizado_2   ,@CORE_Jour_rango_2,    @CORE_Jour_cambia_año_2 ,@CORE_Jour_FoR1_2
,@CORE_Jour_FoR2_2  ,@CORE_Jour_FoR3_2  ,@CORE_Jour_ISSN_2  ,@CORE_Confcomentarios_2
,@CORE_Confpromedio_raiting_2   ,@tipo_LP

WHILE @@FETCH_STATUS = 0
BEGIN
--si TIPO_LP = 1 SJR
--  con el año del articulo, buscamos si hay un LP  del mismo año o el último año disponible
--  insertar id_nombre_lugar_publicacion y datos en Impacto_Revista
--  insertar id_otro_datos_revista en Dim_otro_Dato_revista
--  insertar demás información
--select * from DATAWAREHOUSE.dbo.Impacto_congreso
if @tipo_LP = 1 and @año = @año_articulo
--print @año
--print @año_articulo + "añoArt"
begin
    if (select id_nombre_lugar_publicacion from DATAWAREHOUSE.dbo.Impacto_Revista
        where  id_nombre_lugar_publicacion =  @id_nombre_lugar_publicacion) is null
            begin
            insert into DATAWAREHOUSE.dbo.Impacto_Revista (
            id_nombre_lugar_publicacion,
            año_SJR,
            [año_SJR_es_año_articulo],[H_index_SJR],
            [total_documentos_año_SJR],
            [mejor_cuartil_SJR] ,[total_docs_3años_SJR],[total_referencias_SJR],[total_citas_3años_SJR],
            [total_docs_citados_3años_SJR],[citas_div_docs_SJR],
            [referencias_div_docs_SJR]) values
            (@id_nombre_lugar_publicacion, @año, 'SI',@SJR_H_index,@SJR_Total_Docs_año, @SJR_best_quartile,
            @SJR_Total_Docs_3years,@SJR_Total_Refs,@SJR_Total_Cites_3years,@SJR_Citable_Docs_3years
            ,@SJR_Cites_div_Doc_2years,@SJR_Ref_div_Doc)
            end
        else
            begin
            update DATAWAREHOUSE.dbo.Impacto_Revista
            set año_SJR = @año,
            [año_SJR_es_año_articulo] = 'SI',
            [H_index_SJR] = @SJR_H_index ,
```

```
        [total_documentos_año_SJR] = @SJR_Total_Docs_año,
        [mejor_cuartil_SJR] = @SJR_best_quartile,
        [total_docs_3años_SJR] = @SJR_Total_Docs_3years,
        [total_referencias_SJR] =@SJR_Total_Refs ,
        [total_citas_3años_SJR] = @SJR_Total_Cites_3years,
        [total_docs_citados_3años_SJR] = @SJR_Citable_Docs_3years,
        [citas_div_docs_SJR] = @SJR_Cites_div_Doc_2years,
        [referencias_div_docs_SJR] = @SJR_Ref_div_Doc
        where id_nombre_lugar_publicacion = @id_nombre_lugar_publicacion
        end

end
--si tipo_LP es = 1 y el año menor al año del articulo y el año menor al maximo año del
--id_nombre_lugar_publicacion
if @tipo_LP = 1 and @año < @año_articulo and @año > (select isnull(max(año_SJR),1)
                                    from  DATAWAREHOUSE.dbo.Impacto_Revista
                                    where id_nombre_lugar_publicacion =  @id_nombre_lugar_publicacion)
    begin
        if (select id_nombre_lugar_publicacion from DATAWAREHOUSE.dbo.Impacto_Revista
        where  id_nombre_lugar_publicacion =  @id_nombre_lugar_publicacion) is null
        begin
        insert into DATAWAREHOUSE.dbo.Impacto_Revista (
        id_nombre_lugar_publicacion,
        año_SJR,
        [año_SJR_es_año_articulo],[H_index_SJR],
        [total_documentos_año_SJR],
        [mejor_cuartil_SJR] ,[total_docs_3años_SJR],[total_referencias_SJR],[total_citas_3años_SJR],
        [total_docs_citados_3años_SJR],[citas_div_docs_SJR],[referencias_div_docs_SJR]) values
        (@id_nombre_lugar_publicacion, @año, 'NO',@SJR_H_index,@SJR_Total_Docs_año, @SJR_best_quartile,
        @SJR_Total_Docs_3years,@SJR_Total_Refs,@SJR_Total_Cites_3years,@SJR_Citable_Docs_3years
        ,@SJR_Cites_div_Doc_2years,@SJR_Ref_div_Doc)
    end
else
    begin
        update DATAWAREHOUSE.dbo.Impacto_Revista
        set año_SJR = @año,
        [año_SJR_es_año_articulo] = 'NO',
        [H_index_SJR] = @SJR_H_index ,

         [total_documentos_año_SJR] = @SJR_Total_Docs_año,
         [mejor_cuartil_SJR] = @SJR_best_quartile,
         [total_docs_3años_SJR] = @SJR_Total_Docs_3years,
         [total_referencias_SJR] =@SJR_Total_Refs ,
         [total_citas_3años_SJR] = @SJR_Total_Cites_3years,
         [total_docs_citados_3años_SJR] = @SJR_Citable_Docs_3years,
         [citas_div_docs_SJR] = @SJR_Cites_div_Doc_2years,
         [referencias_div_docs_SJR] = @SJR_Ref_div_Doc
         where id_nombre_lugar_publicacion = @id_nombre_lugar_publicacion
     end

end
--si TIPO_LP = 2
--  con el año del articulo, buscamos si hay un LP  del mismo año o el último año disponible
--  insertar id_nombre_lugar_publicacion en Impacto_Congreso
--  insertar datos Impacto_Congreso
if @tipo_LP = 2 and @año = @año_articulo
--print @año
--print @año_articulo + "añoArt"
begin
    if (select id_nombre_lugar_publicacion from DATAWAREHOUSE.dbo.Impacto_Congreso
        where  id_nombre_lugar_publicacion =  @id_nombre_lugar_publicacion) is null
        begin
        insert into DATAWAREHOUSE.dbo.Impacto_Congreso (
        id_nombre_lugar_publicacion,
        año_SJR,
        [año_SJR_es_año_articulo],[H_index_SJR],
        [total_documentos_año_SJR],
        [mejor_cuartil_SJR] ,[total_docs_3años_SJR],[total_referencias_SJR],[total_citas_3años_SJR],
        [total_docs_citados_3años_SJR],[citas_div_docs_SJR],[referencias_div_docs_SJR]) values
        (@id_nombre_lugar_publicacion, @año, 'SI',@SJR_H_index,@SJR_Total_Docs_año,
        @SJR_best_quartile,@SJR_Total_Docs_3years,@SJR_Total_Refs,@SJR_Total_Cites_3years,
        @SJR_Citable_Docs_3years,@SJR_Cites_div_Doc_2years,@SJR_Ref_div_Doc)
        end
    else
        begin
        update DATAWAREHOUSE.dbo.Impacto_Congreso
        set año_SJR = @año,
```

```
        [año_SJR_es_año_articulo] = 'SI',
        [H_index_SJR] = @SJR_H_index ,
        [total_documentos_año_SJR] = @SJR_Total_Docs_año,
        [mejor_cuartil_SJR] = @SJR_best_quartile,
        [total_docs_3años_SJR] = @SJR_Total_Docs_3years,
        [total_referencias_SJR] =@SJR_Total_Refs ,
        [total_citas_3años_SJR] = @SJR_Total_Cites_3years,
        [total_docs_citados_3años_SJR] = @SJR_Citable_Docs_3years,
        [citas_div_docs_SJR] = @SJR_Cites_div_Doc_2years,
        [referencias_div_docs_SJR] = @SJR_Ref_div_Doc
        where id_nombre_lugar_publicacion = @id_nombre_lugar_publicacion
        end


end
--si tipo_LP es = 2 y el año menor al año del articulo y el año menor al maximo año del
--id_nombre_lugar_publicacion
if @tipo_LP = 2 and @año < @año_articulo and @año > (select isnull(max(año_SJR),1)
                        from  DATAWAREHOUSE.dbo.Impacto_Revista
                        where id_nombre_lugar_publicacion =  @id_nombre_lugar_publicacion)
    begin
        if (select id_nombre_lugar_publicacion from DATAWAREHOUSE.dbo.Impacto_Congreso
                        where  id_nombre_lugar_publicacion =  @id_nombre_lugar_publicacion) is null
        begin
        insert into DATAWAREHOUSE.dbo.Impacto_Congreso (
        id_nombre_lugar_publicacion,
        año_SJR,
        [año_SJR_es_año_articulo],[H_index_SJR],
        [total_documentos_año_SJR],
        [mejor_cuartil_SJR] ,[total_docs_3años_SJR],[total_referencias_SJR],[total_citas_3años_SJR],
        [total_docs_citados_3años_SJR],[citas_div_docs_SJR],[referencias_div_docs_SJR]) values
        (@id_nombre_lugar_publicacion, @año, 'NO',@SJR_H_index,@SJR_Total_Docs_año,
        @SJR_best_quartile,@SJR_Total_Docs_3years,@SJR_Total_Refs,@SJR_Total_Cites_3years,
        @SJR_Citable_Docs_3years,@SJR_Cites_div_Doc_2years,@SJR_Ref_div_Doc)
    end
else
    begin
        update DATAWAREHOUSE.dbo.Impacto_Congreso

         set año_SJR = @año,
         [año_SJR_es_año_articulo] = 'NO',
         [H_index_SJR] = @SJR_H_index ,
         [total_documentos_año_SJR] = @SJR_Total_Docs_año,
         [mejor_cuartil_SJR] = @SJR_best_quartile,
         [total_docs_3años_SJR] = @SJR_Total_Docs_3years,
         [total_referencias_SJR] =@SJR_Total_Refs ,
         [total_citas_3años_SJR] = @SJR_Total_Cites_3years,
         [total_docs_citados_3años_SJR] = @SJR_Citable_Docs_3years,
         [citas_div_docs_SJR] = @SJR_Cites_div_Doc_2years,
         [referencias_div_docs_SJR] = @SJR_Ref_div_Doc
         where id_nombre_lugar_publicacion = @id_nombre_lugar_publicacion
     end
 end
      ----------------------------------------
 if @tipo_LP = 5 and @año = @año_articulo
 --print @año
 --print @año_articulo + "añoArt"
 begin
     if (select id_nombre_lugar_publicacion from DATAWAREHOUSE.dbo.Impacto_Revista
                         where  id_nombre_lugar_publicacion =  @id_nombre_lugar_publicacion) is null
         begin
         insert into DATAWAREHOUSE.dbo.Impacto_Revista (
         id_nombre_lugar_publicacion,[año_JCR],[año_JCR_es_año_articulo],[total_citas_JCR],
         [factor_impacto_JCR],[factor_impacto_sin_citas_Propias_JCR],[factor_impacto_5años_JCR],
         [elementos_citables_JCR],[puntuacion_factor_propio_JCR],[puntuacion_influencia_articulo_JCR],
         [porcentaje_elementos_articulos_citables_JCR],[factor_propio_normalizado_JCR],
         [cited_half_life [varchar_JCR],[citing_half_life_JCR]  ) values
         (@id_nombre_lugar_publicacion, @año, 'SI',@JCR_total_citas,@JCR_factor_impacto,
         @JCR_factor_impacto_sin_citas_propias,@JCR_factor_impacto_5anhos_2,
         @JCR_elementos_citables_2,@JCR_puntuacion_factor_propio_2,@JCR_puntuacion_influencia_ariculo_2,
         @JCR_porcentaje_elementos_articulos_citables_2,
         @JCR_factor_propio_normalizado_2,@JCR_cited_half_life_2,@JCR_citing_half_life_2)
         end
     else
         begin
         update DATAWAREHOUSE.dbo.Impacto_Revista
         set [año_JCR] = @año,
```

```
        [año_JCR_es_año_articulo] = 'SI',
        [total_citas_JCR] = @SJR_H_index ,
        [factor_impacto_JCR] = @SJR_Total_Docs_año,
        [factor_impacto_sin_citas_Propias_JCR] = @SJR_best_quartile,
        [factor_impacto_5años_JCR] = @SJR_Total_Docs_3years,
        [elementos_citables_JCR] =@SJR_Total_Refs ,
        [factor_propio_normalizado_JCR] = @SJR_Total_Cites_3years,
        [cited_half_life [varchar_JCR] = @SJR_Citable_Docs_3years,
        [citing_half_life_JCR] = @SJR_Cites_div_Doc_2years
        where id_nombre_lugar_publicacion = @id_nombre_lugar_publicacion
        end

end
--si tipo_LP es = 1 y el año menor al año del articulo y el año menor al maximo año del
--id_nombre_lugar_publicacion
if @tipo_LP = 5 and @año < @año_articulo and @año > (select isnull(max(año_SJR),1)
                        from  DATAWAREHOUSE.dbo.Impacto_Revista
                        where id_nombre_lugar_publicacion =  @id_nombre_lugar_publicacion)
    begin
        if (select id_nombre_lugar_publicacion from DATAWAREHOUSE.dbo.Impacto_Revista
                        where  id_nombre_lugar_publicacion =  @id_nombre_lugar_publicacion) is null
        begin
        insert into DATAWAREHOUSE.dbo.Impacto_Revista (
        id_nombre_lugar_publicacion,[año_JCR],[año_JCR_es_año_articulo],[total_citas_JCR],
        [factor_impacto_JCR],[factor_impacto_sin_citas_Propias_JCR],[factor_impacto_5años_JCR],
        [elementos_citables_JCR],[puntuacion_factor_propio_JCR],[puntuacion_influencia_articulo_JCR],
        [porcentaje_elementos_articulos_citables_JCR],[factor_propio_normalizado_JCR],
        [cited_half_life [varchar_JCR],[citing_half_life_JCR]  ) values
        (@id_nombre_lugar_publicacion, @año, 'NO',@JCR_total_citas,@JCR_factor_impacto,
        @JCR_factor_impacto_sin_citas_propias,@JCR_factor_impacto_5anhos_2,
        @JCR_elementos_citables_2,@JCR_puntuacion_factor_propio_2,@JCR_puntuacion_influencia_ariculo_2,
        @JCR_porcentaje_elementos_articulos_citables_2, @JCR_factor_propio_normalizado_2,
        @JCR_cited_half_life_2,@JCR_citing_half_life_2)
    end
else
    begin
        update DATAWAREHOUSE.dbo.Impacto_Revista
        set [año_JCR] = @año,

        [año_JCR_es_año_articulo] = 'NO',
        [total_citas_JCR] = @SJR_H_index ,
        [factor_impacto_JCR] = @SJR_Total_Docs_año,
        [factor_impacto_sin_citas_Propias_JCR] = @SJR_best_quartile,
        [factor_impacto_5años_JCR] = @SJR_Total_Docs_3years,
        [elementos_citables_JCR] =@SJR_Total_Refs ,
        [factor_propio_normalizado_JCR] = @SJR_Total_Cites_3years,
        [cited_half_life [varchar_JCR] = @SJR_Citable_Docs_3years,
        [citing_half_life_JCR] = @SJR_Cites_div_Doc_2years
        where id_nombre_lugar_publicacion = @id_nombre_lugar_publicacion
    end

end
--  si TIPO_LP = 3 -> journal
if  @tipo_LP = 3
begin
--  actualizar id_grupo_core en Impacto_Lugar_publicacion  (capturarlo)
    set @GrupoCore  = (select id_grupo_core from  DATAWAREHOUSE.dbo.Impacto_Lugar_Publicacion
                        where id_nombre_lugar_publicacion =  @id_nombre_lugar_publicacion )
--  insertar id_grupo_core en  Grupo_Core (@GrupoCore)
--si no existe el grupo core
if (select id_grupo_core from Grupo_CORE where id_grupo_core =  @GrupoCore) is null
begin
    insert into DATAWAREHOUSE.dbo.Grupo_CORE(id_grupo_core)values
    (@GrupoCore)
end
--si ya existe el grupo
else
begin
    update DATAWAREHOUSE.dbo.Grupo_CORE
    set id_grupo_core = @GrupoCore
    where id_grupo_core = @GrupoCore
end
    --   insertar dim_datos_core
    set @nuevoIdDatosCore = (select isnull(max(id_datos_core),0) + 1
                            from  DATAWAREHOUSE.dbo.Dim_Datos_CORE)
    insert into DATAWAREHOUSE.dbo.Dim_Datos_CORE (
    id_datos_core,
```

```
    FoR_CORE,
    calificacion_core,
    cambia_año_core,
    desc_edicion_core) values
    (@nuevoIdDatosCore,
    @CORE_Jour_FoR1_2,
    @CORE_Jour_rango_2,
    @CORE_Jour_cambia_año_2,
    concat('core_journal', @año))

    -- crear registro   en Grupo_EdicionesCore_LugarPublicacion
insert into DATAWAREHOUSE.dbo.Grupo_EdicionesCore_LugarPublicacion (id_grupo_core, id_datos_core) values
(@GrupoCore,    @nuevoIdDatosCore)
end
if  @tipo_LP = 4 --> confer
begin
--  actualizar id_grupo_core en Impacto_Lugar_publicacion  (capturarlo)
    set @GrupoCore  = (select id_grupo_core from  DATAWAREHOUSE.dbo.Impacto_Lugar_Publicacion
                       where id_nombre_lugar_publicacion =  @id_nombre_lugar_publicacion )
--  insertar id_grupo_core en  Grupo_Core (@GrupoCore)
--si no existe el grupo core
if (select id_grupo_core from Grupo_CORE where id_grupo_core =  @GrupoCore) is null
begin
    insert into DATAWAREHOUSE.dbo.Grupo_CORE(id_grupo_core)values
    (@GrupoCore)
end
--si ya existe el grupo
else
begin
    update DATAWAREHOUSE.dbo.Grupo_CORE
    set id_grupo_core = @GrupoCore
    where id_grupo_core = @GrupoCore
end
    --   insertar dim_datos_core
    set @nuevoIdDatosCore = (select isnull(max(id_datos_core),0) + 1
                             from  DATAWAREHOUSE.dbo.Dim_Datos_CORE)
    insert into DATAWAREHOUSE.dbo.Dim_Datos_CORE (
    id_datos_core,

    FoR_CORE,
    calificacion_core,
    cambia_año_core,
    desc_edicion_core) values
    (@nuevoIdDatosCore,
    @CORE_Jour_FoR1_2,
    @CORE_Jour_rango_2,
    @CORE_Jour_cambia_año_2,
    concat('core_conference', @año))

    -- crear registro   en Grupo_EdicionesCore_LugarPublicacion
insert into DATAWAREHOUSE.dbo.Grupo_EdicionesCore_LugarPublicacion (id_grupo_core, id_datos_core) values
(@GrupoCore,    @nuevoIdDatosCore)
end
    --select * from Grupo_EdicionesCore_LugarPublicacion
    --select * from Dim_Datos_CORE
    --select * from grupo_core

FETCH NEXT FROM cursor_Busqueda_tipo_LP INTO @lugar_publicacion,    @año,   @SJR_tipo,  @SJR_ISSN
,@SJR   ,@SJR_best_quartile ,@SJR_H_index   ,@SJR_Total_Docs_3years ,@SJR_Total_Refs
,@SJR_Total_Cites_3years    ,@SJR_Citable_Docs_3years   ,@SJR_Cites_div_Doc_2years  ,@SJR_Ref_div_Doc
,@SJR_Country   ,@SJR_Total_Docs_2009   ,@SJR_area_conocimiento ,@JCR_total_citas,@JCR_factor_impacto
,@JCR_factor_impacto_sin_citas_propias  ,@JCR_factor_impacto_5anhos_2   ,@JCR_elementos_citables_2
,@JCR_cited_half_life_2 ,@JCR_citing_half_life_2    ,@JCR_puntuacion_factor_propio_2
,@JCR_puntuacion_influencia_ariculo_2   ,@JCR_porcentaje_elementos_articulos_citables_2
,@JCR_promedio_percentil_factor_impacto_2,@JCR_factor_propio_normalizado_2  ,@CORE_Jour_rango_2,
@CORE_Jour_cambia_año_2 ,@CORE_Jour_FoR1_2  ,@CORE_Jour_FoR2_2  ,@CORE_Jour_FoR3_2  ,@CORE_Jour_ISSN_2
,@CORE_Confcomentarios_2    ,@CORE_Confpromedio_raiting_2   ,@tipo_LP
END
CLOSE cursor_Busqueda_tipo_LP;
DEALLOCATE cursor_Busqueda_tipo_LP;
FETCH NEXT FROM cursor_DWH_LP INTO @id_nombre_lugar_publicacion ,@desc_nombre_lugar_publicacion ,
@año_articulo

END
CLOSE cursor_DWH_LP;
DEALLOCATE cursor_DWH_LP;
```

Apéndice H. Configuración DQS (*Data Quality Services*) similitud y confianza.

Abrir el directorio “Planos_TFM” -> “Configuración” y de esta carpeta abrimos el fichero de Excel “Configuracion_DQS”:

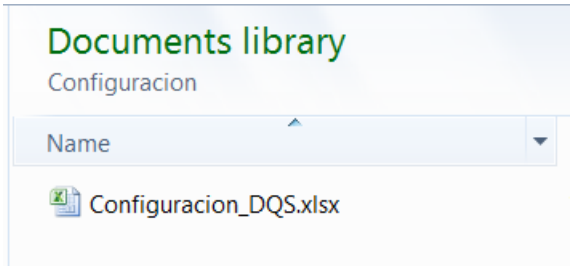


Figura 67: Directorio de configuración

Abrir el fichero en la hoja “Similitud_confianza”:

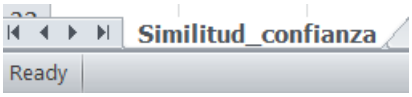


Figura 68: Hoja libro Excel "Similitud_confianza"

Finalmente, actualizar los campos de similitud y confianza.

Tener en cuenta que los valores son entre 0 y 1 puesto que son valores porcentuales, donde 1 significa que es una coincidencia exacta.

Si se desea hacer pruebas para calibrar estas medidas, se puede conseguir apoyo a las tablas de log de DQS y así ver las coincidencias de todas las tablas con sus respectivos valores de similitud y confianza:

	A	B
1	**Similitud**	**Confianza**
2	0,5	0,5

Figura 69: Ingresar valores de similitud y confianza

Ejecutar el proceso “Configuracion_Similitud_confianza”.

Apéndice I. Propiedades Parametrización Búsqueda Aproximada DQS.

Toda la información tomada de *https://docs.microsoft.com/es-es/sql/integration-services/data-flow/transformations/transformation-custom-properties#flookup.*

Las propiedades de la transformación de búsqueda aproximada son:

- Propiedades personalizadas de la transformación (todas menos "ReferenceMetadataXML" son de lectura y escritura):

Propiedad	Tipo de datos	Descripción
CopyReferenceTable	Boolean	Especifica si se debería crear una copia de la tabla de referencia para la construcción del índice de búsqueda aproximada y las búsquedas subsiguientes. El valor predeterminado de esta propiedad es **True**.
Delimiters	String	Delimitadores que la transformación utilizará para dividir en tokens los valores de las columnas. Los delimitadores predeterminados incluyen los siguientes caracteres: espacio (), coma (,), semicolon(;) de punto (.), guión (-) de dos puntos (:), doble comillas rectas ("), marca de comillas rectas ('), marca" y "comercial (&), la barra diagonal (/), barra diagonal inversa (\), arroba (@), signo de exclamación (!), signo de interrogación (?), paréntesis de apertura ((), cerrar paréntesis ()), menor que (<), mayor que (>), abrir corchete ([]), corchete de cierre (]), llave ({}), Cerrar llave (}), canalización (|) de apertura de cierre. almohadilla (#), asterisco (*), símbolo de intercalación (^) y porcentaje (%).*
DropExistingMatchIndex	Boolean	Valor que especifica si el índice de coincidencia especificado en MatchIndexName se elimina cuando MatchIndexOptions no se establece en ReuseExistingIndex. El valor predeterminado de esta propiedad es **True**.
Exhaustive	Boolean	Valor que especifica si cada registro de entrada se compara con el resto. El valor de **True** está destinado sobre todo a fines de depuración. El valor predeterminado de esta propiedad es **False**. Nota: Esta propiedad no está disponible en el Editor de transformación Búsqueda aproximada, pero se puede establecer con el Editor avanzado.

MatchIndexName	String	Nombre del índice de coincidencia. El índice de coincidencia es la tabla en la que la transformación crea y guarda el índice que utiliza. Si se reutiliza el índice de coincidencia, MatchIndexName especifica el índice que se reutilizará. MatchIndexName debe ser un nombre de identificador de SQL Server válido. Por ejemplo, si el nombre contiene espacios, debe escribirse entre corchetes.
MatchIndexOptions	Integer (enumeración)	Valor que especifica cómo administra la transformación el índice de coincidencia. Esta propiedad admite cualquiera de los siguientes valores: ReuseExistingIndex (0), GenerateNewIndex (1), GenerateAndPersistNewIndex (2) y GenerateAndMaintainNewIndex (3)
MaxMemoryUsage	Integer	Tamaño máximo permitido de la caché para la tabla de búsqueda. El valor predeterminado de esta propiedad es **0**, lo que significa que el tamaño de la memoria caché no tiene límite. Puede especificar el valor de esta propiedad con una expresión de propiedad. Nota: Esta propiedad no está disponible en el Editor de transformación Búsqueda aproximada, pero se puede establecer con el Editor avanzado.
MaxOutputMatchesPerInput	Integer	Número máximo de coincidencias que la transformación puede devolver para cada fila de entrada. El valor predeterminado de esta propiedad es **1**. Nota: Los valores mayores que 100 solo se pueden especificar con el Editor avanzado.
MinSimilarity	Integer	El umbral de similitud que la transformación usa en el nivel de componente, especificado como un valor entre 0 y 1. Solo las filas mayores que el umbral se consideran coincidencias.
ReferenceMetadataXML	String	Solamente se identifica con fines informativos. No compatible. La compatibilidad con versiones posteriores no está garantizada.
ReferenceTableName	String	Nombre de la tabla de búsqueda. El nombre debe ser un nombre de identificador de SQL Server válido. Por ejemplo, si el nombre contiene espacios, debe escribirse entre corchetes.
WarmCaches	Boolean	Cuando es true, la búsqueda carga parcialmente el índice y la tabla de referencia en la memoria antes de comenzar la ejecución. Esto puede mejorar el rendimiento.

- Propiedades personalizadas de las columnas de entrada, todas las propiedades son de lectura y escritura:

Propiedad	Tipo de datos	Descripción
FuzzyComparisonFlags	Integer	Valor que especifica cómo compara la transformación los datos de cadena de una columna.
FuzzyComparisonFlagsEx	Integer (enumeración)	Valor que especifica qué marcas de comparación extendida usa la transformación. Los valores pueden incluir MapExpandLigatures, MapFoldCZone, MapFoldDigits, MapPrecomposedy NoMapping.NoMapping no se puede usar con otras marcas.
JoinToReferenceColumn	String	Valor que especifica el nombre de la columna en la tabla de referencia con la que se combina la columna.
JoinType	Integer	Valor que especifica si la transformación realiza una coincidencia aproximada o exacta. El valor predeterminado de esta propiedad es Aproximada. El valor entero para el tipo de combinación exacta es 1 y para el tipo de combinación aproximado es 2.
MinSimilarity	Doble	Umbral de similitud que la transformación usa en el nivel de columna, especificado como un valor entre 0 y 1. Solo las filas mayores que el umbral se consideran coincidencias.

- Propiedades personalizadas de las columnas de salida, todas las propiedades son de lectura y escritura:

Propiedad	Tipo de datos	Descripción
ColumnType	Integer (enumeración)	Valor que identifica el tipo de columna de resultados para las columnas que la transformación agrega a la salida. Esta propiedad admite cualquiera de los siguientes valores: Undefined (0), Similarity (1), Confidence (2) y ColumnSimilarity (3)
CopyFromReferenceColumn	String	Valor que especifica el nombre de la columna en la tabla de referencia que proporciona el valor en una columna de resultados.
SourceInputColumnLineageId	Integer	Valor que identifica la columna de entrada que proporciona los valores de esta columna de resultados.

Apéndice J. Descripción de procesos ETL para el proyecto(*Staging* y tablas auxiliares).

Como se ha mencionado en la memoria del trabajo de fin de máster, la función de las ETL de *Staging* es cargar los datos de las fuentes de información en tablas intermedias tal y como viene de las fuentes de información:

Solución ETL de Tablas Auxiliares

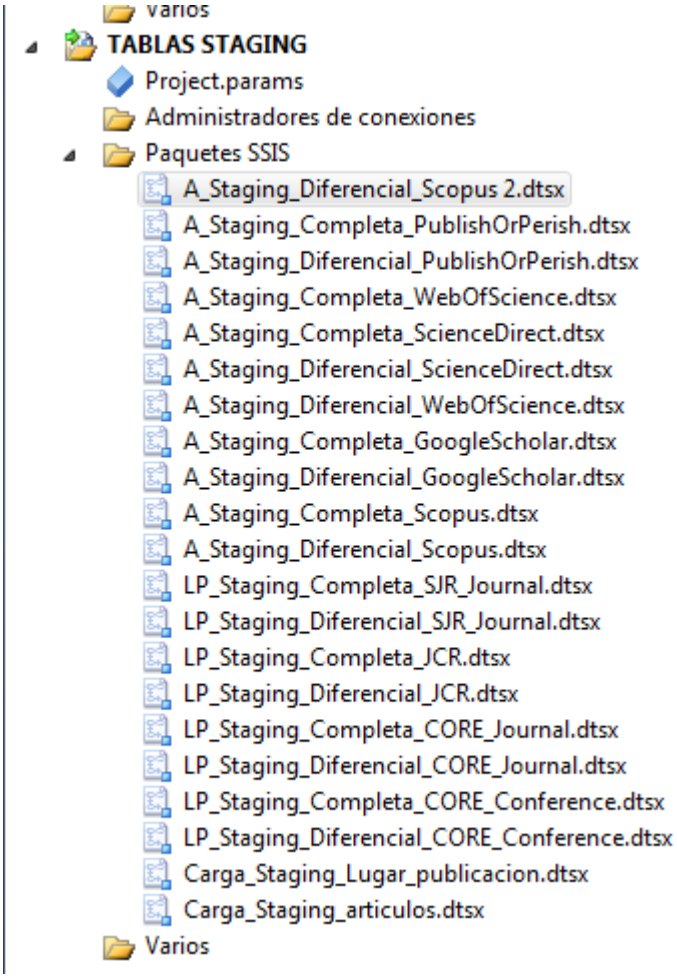


Figura 70: Procesos ETL, carga Staging

En la figura anterior se puede apreciar cada una de las ETL de los ficheros del sistema. Como se puede observar hay dos por cada fuente una llamada “Diferencial” y otra llamada “Completa”.

Las ETL con el nombre “diferencial” se encarga se cargar un fichero de la ruta. Y las ETL con el nombre “completa” se encarga de ejecutar la ETL “diferencial” por cada fichero que se encuentre en un directorio. Por ejemplo, si el directorio “JCR” tiene 10 ficheros

para cargar, la ETL "completa" ejecuta diez veces la ETL "diferencial"; una vez por cada fichero, de esta forma se asegura que carguen todos los ficheros.

Para cada fuente de información se ha desarrollado una ETL diferente puesto que todos los ficheros son diferentes, pero la lógica de carga de cada uno de ellos es la misma.

A continuación se explicará en detalle una carga y para el resto de cargas sólo las mencionaremos, con el fin de no sobrecargar el documento con información similar.

- Carga de fuente de datos CORE JCR.

Todas las cargas están monitorizadas por un log de inicio y un log de finalización, sea esta finalización correcta o error. Lo que se hace es insertar en la tabla de log en la base de datos un estado de inicio con el nombre de la ETL, la fecha-hora del inicio de la ejecución y su finalización (correcta o error), esto para dejar un registro de ejecución por si algún tipo de administrador desea ver las cargas o para hacer seguimiento de algún tipo de error.

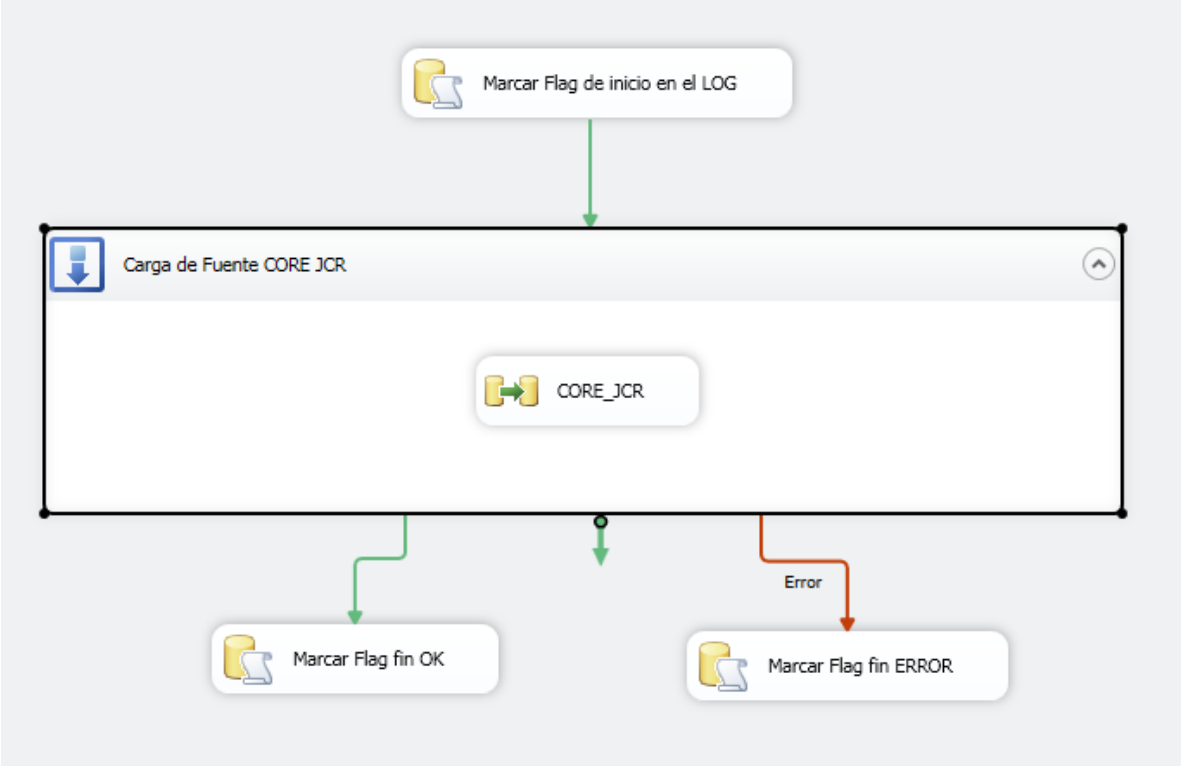


Figura 71: Ejemplo carga CORE JCR

De acuerdo a la explicación anterior, se muestra en la Figura 24 una caja de inicio "Marcar *Flag* de inicio en el LOG", que marca el estado de inicio de ejecución de la ETL. A continuación procesa la "Carga de Fuente CORE JCR", que como su nombre indica, carga la fuente de datos CORE JCR. Al hacer doble clic en la caja del flujo de datos "CORE_JCR" podemos ver lo que se está ejecutando dentro:

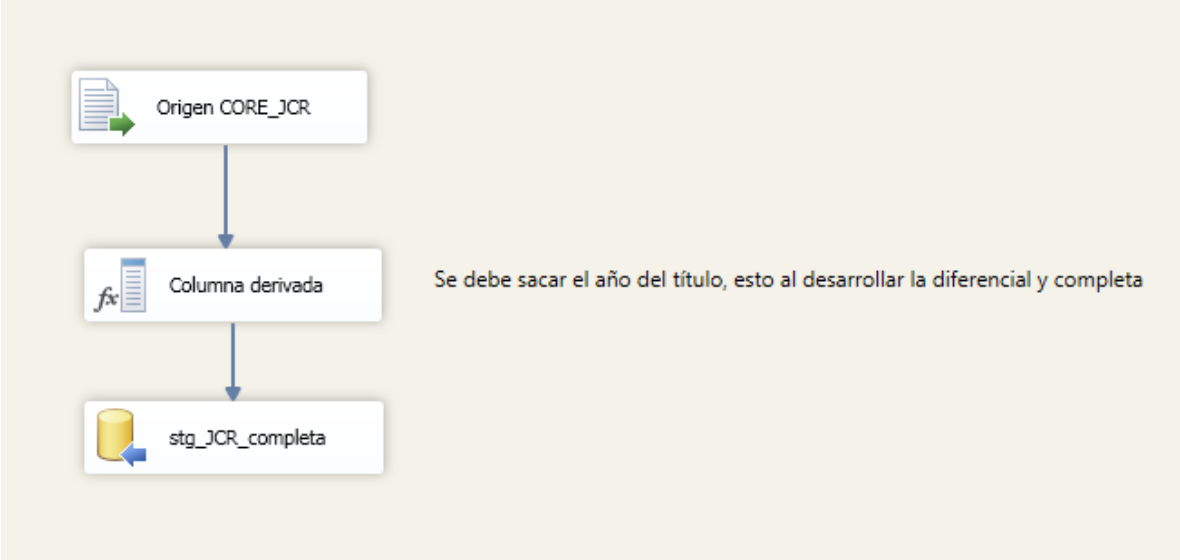


Figura 72: Flujo de datos carga CORE JCR

Se pueden observar tres cajas dentro de la caja del flujo de datos:

- Una caja de Origen de Datos con el nombre “Origen CORE_JRC”

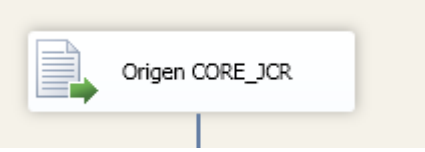


Figura 73: Caja de origen de datos

Esta caja consiste en extraer la información del archivo plano que está descrito en la cadena de conexión de la misma:.

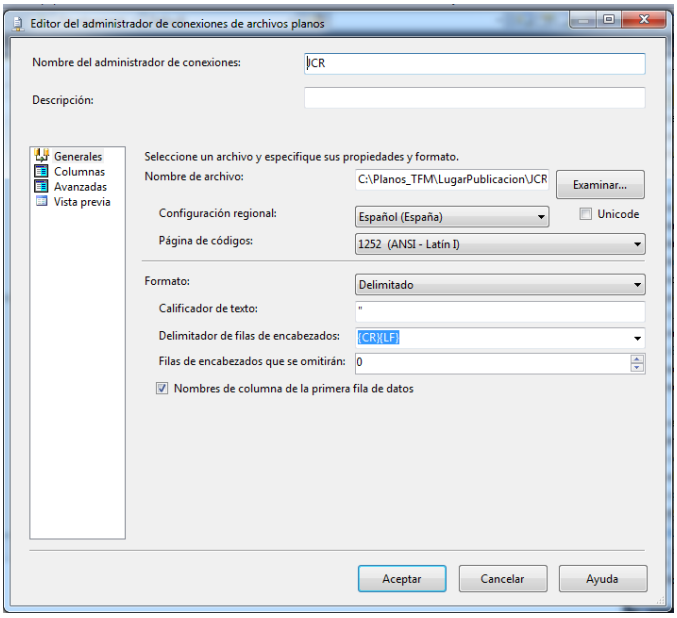
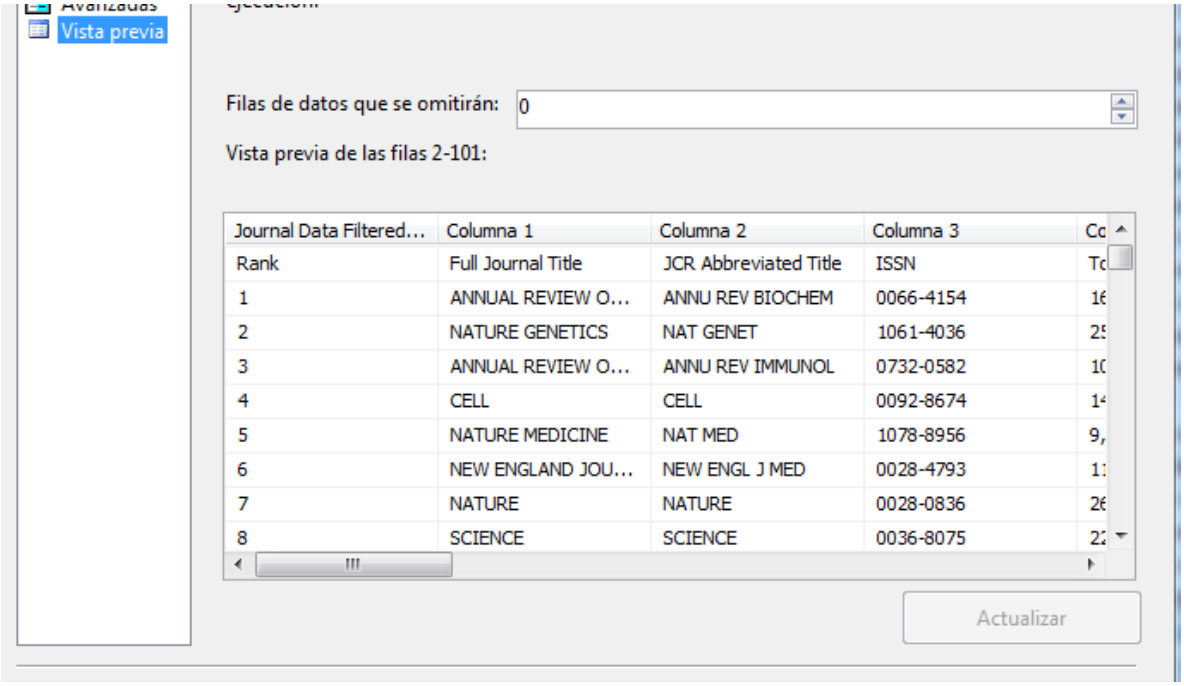


Figura 74: Configuración cadena de conexión

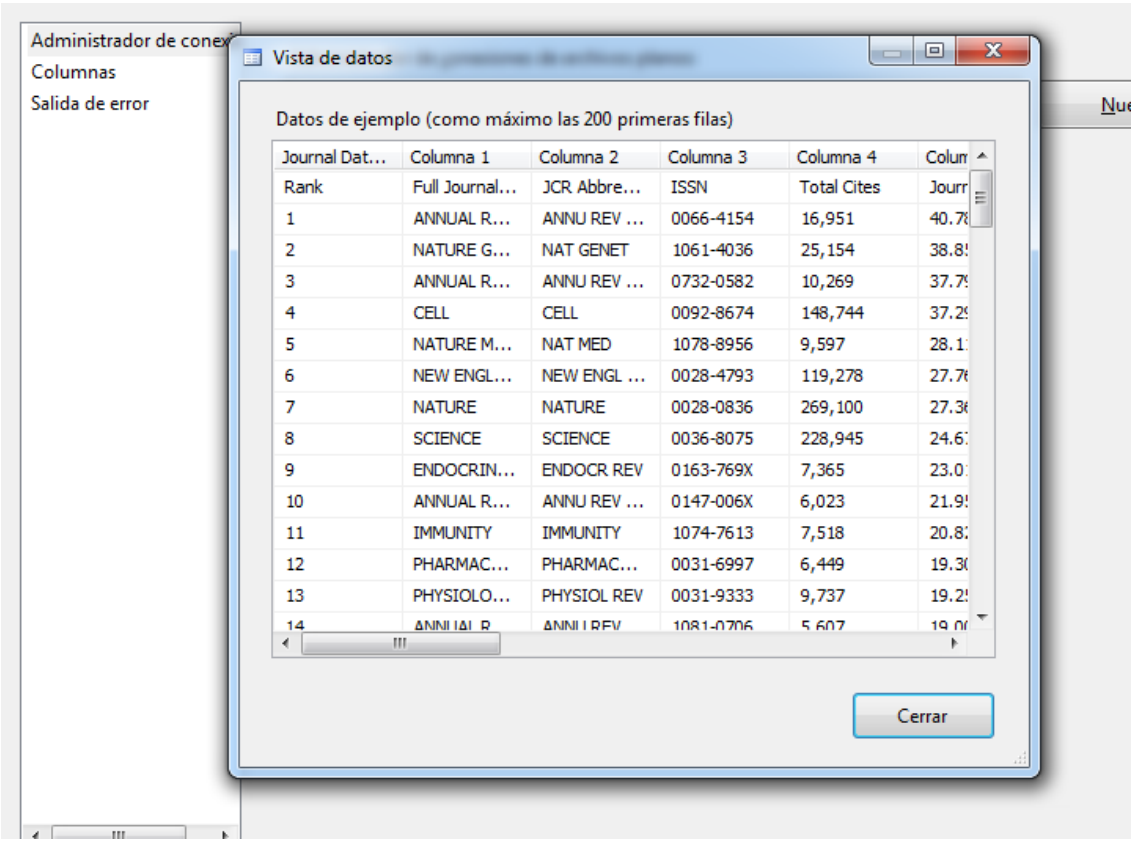
Lo que hace la conexión es ir donde está físicamente el fichero de datos y leer su contenido. Al ir a la opción de "Vista previa" podemos ver el contenido del fichero:



Journal Data Filtered...	Columna 1	Columna 2	Columna 3	Co
Rank	Full Journal Title	JCR Abbreviated Title	ISSN	To
1	ANNUAL REVIEW O...	ANNU REV BIOCHEM	0066-4154	16
2	NATURE GENETICS	NAT GENET	1061-4036	25
3	ANNUAL REVIEW O...	ANNU REV IMMUNOL	0732-0582	10
4	CELL	CELL	0092-8674	14
5	NATURE MEDICINE	NAT MED	1078-8956	9,
6	NEW ENGLAND JOU...	NEW ENGL J MED	0028-4793	11
7	NATURE	NATURE	0028-0836	26
8	SCIENCE	SCIENCE	0036-8075	22

Figura 75: Vista previa fichero CORE JCR

Desde la caja de origen de datos también se puede ver el contenido que se está leyendo haciendo doble clic en la misma y luego clic en "Vista Previa":



Journal Dat...	Columna 1	Columna 2	Columna 3	Columna 4	Colum
Rank	Full Journal...	JCR Abbre...	ISSN	Total Cites	Jourr
1	ANNUAL R...	ANNU REV ...	0066-4154	16,951	40.7
2	NATURE G...	NAT GENET	1061-4036	25,154	38.8
3	ANNUAL R...	ANNU REV ...	0732-0582	10,269	37.7
4	CELL	CELL	0092-8674	148,744	37.2
5	NATURE M...	NAT MED	1078-8956	9,597	28.1
6	NEW ENGL...	NEW ENGL ...	0028-4793	119,278	27.7
7	NATURE	NATURE	0028-0836	269,100	27.3
8	SCIENCE	SCIENCE	0036-8075	228,945	24.6
9	ENDOCRIN...	ENDOCR REV	0163-769X	7,365	23.0
10	ANNUAL R...	ANNU REV ...	0147-006X	6,023	21.9
11	IMMUNITY	IMMUNITY	1074-7613	7,518	20.8
12	PHARMAC...	PHARMAC...	0031-6997	6,449	19.3
13	PHYSIOLO...	PHYSIOL REV	0031-9333	9,737	19.2
14	ANNUAL R...	ANNU REV ...	1081-0706	5,607	19.0

Figura 76: Vista previa II fichero CORE JCR

- "Columna derivada".

Esta caja funciona en general para hacer operaciones entre los registros durante el flujo de datos, por ejemplo algún tipo de operación aritmética, aplicar alguna función de cadena o número, la reasignación de un tipo de dato, o la asignación de un valor a una variable dada. Para este proyecto utilizamos esta caja para crear una nueva columna con el nombre de "año" que obtiene su valor de variable @[User::AñoMes].

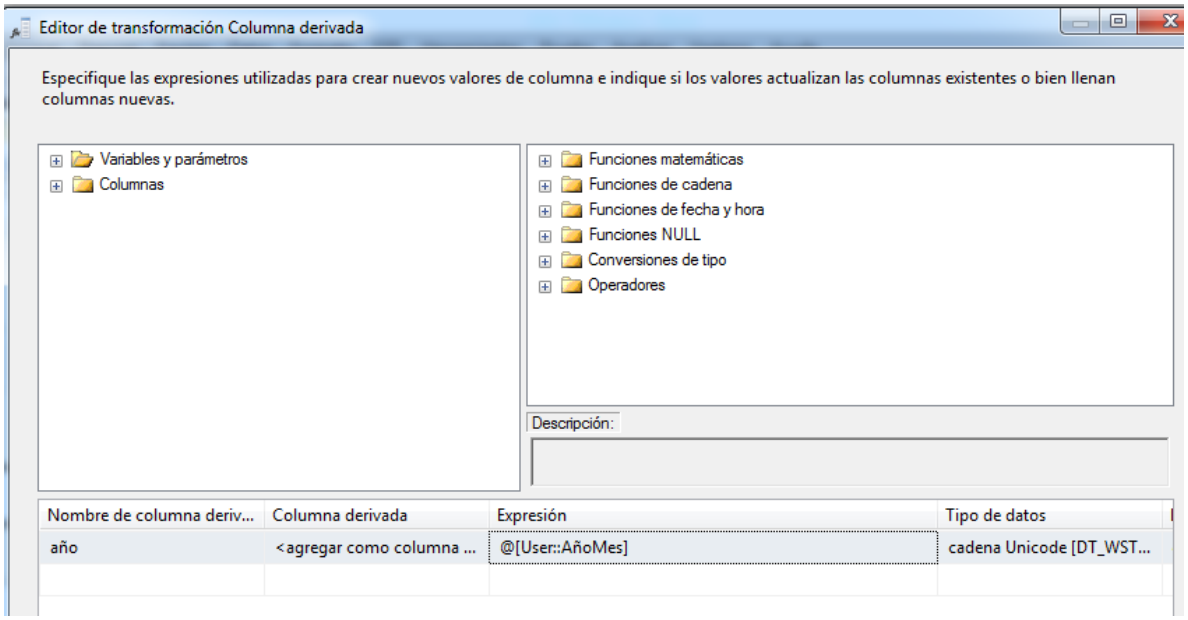


Figura 77: Asignación de variable en la caja de columna derivada

En el apartado de variable se puede ver que este AñoMes es sacado del nombre de fichero de carga a través de la siguiente función: (Substring(@[User::Fichero],17,4)). Del nombre del fichero, toma desde la posición 17 y 4 caracteres más lo que para este ejemplo sería un 1997:

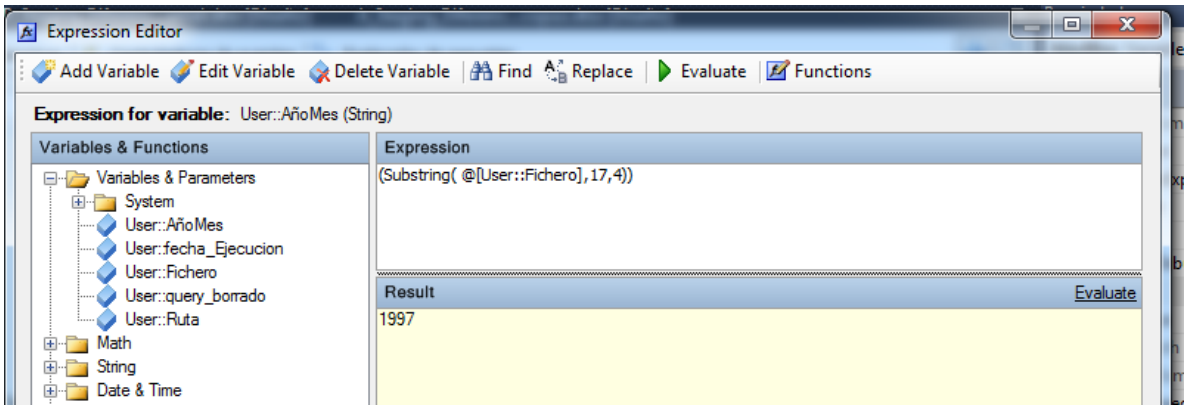


Figura 78: Expresión para la variable AñoMes

- Destino OLE BD “stg_JCR_completa”.

Figura 79: Caja de destino a base de datos

Esta caja toma todos los datos que vienen en el flujo de datos y lo inserta en una tabla destino. Para este ejemplo la tabla destino es [dbo]. [stg_SJR_completa]; al abrir la caja, se puede visualizar la cadena de conexión, el nombre de la tabla destino, y algunas opciones de inserción:

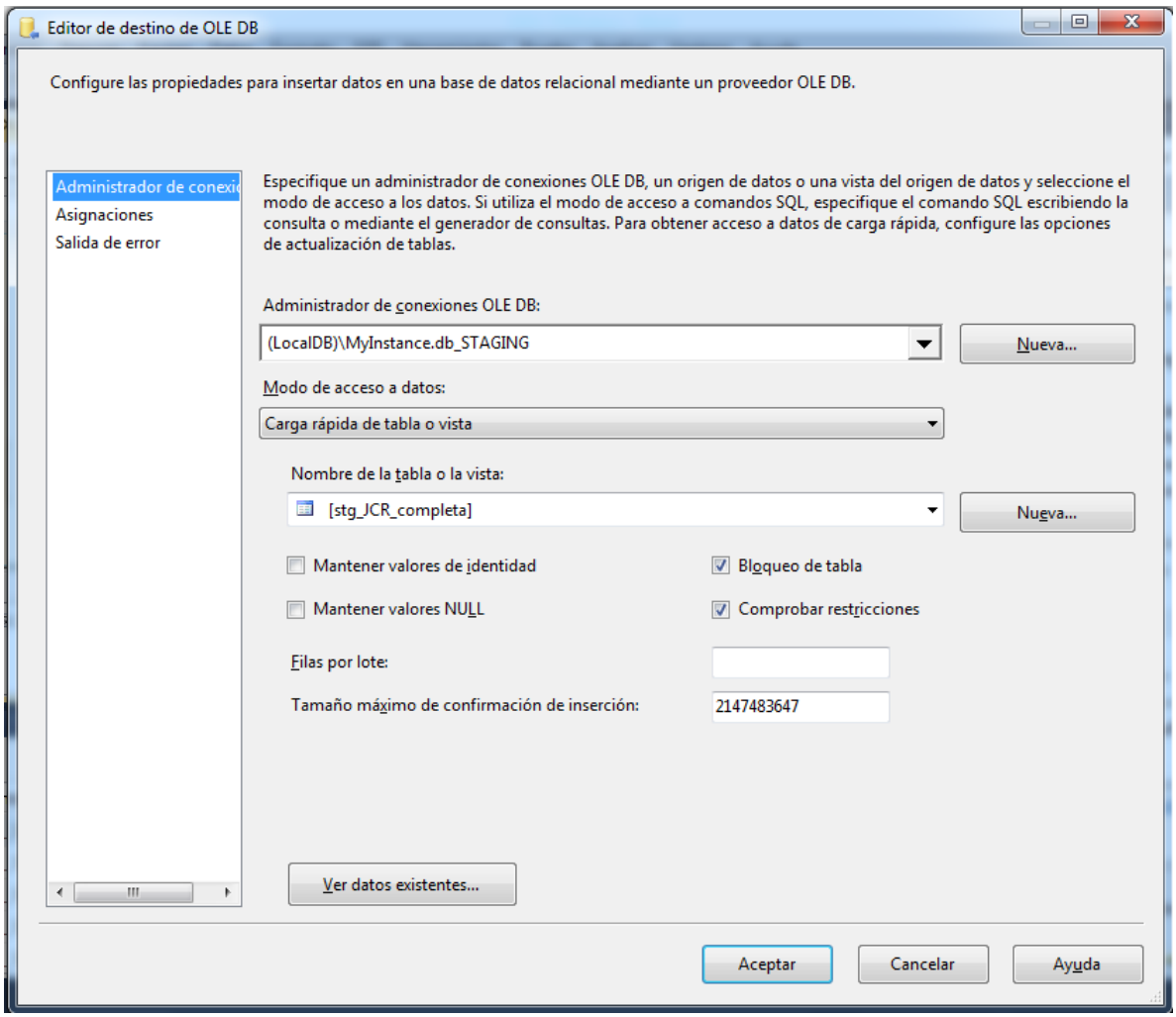


Figura 80: Administrador de conexiones

En la pestaña de asignaciones es donde se configura cada columna de la tabla destino.

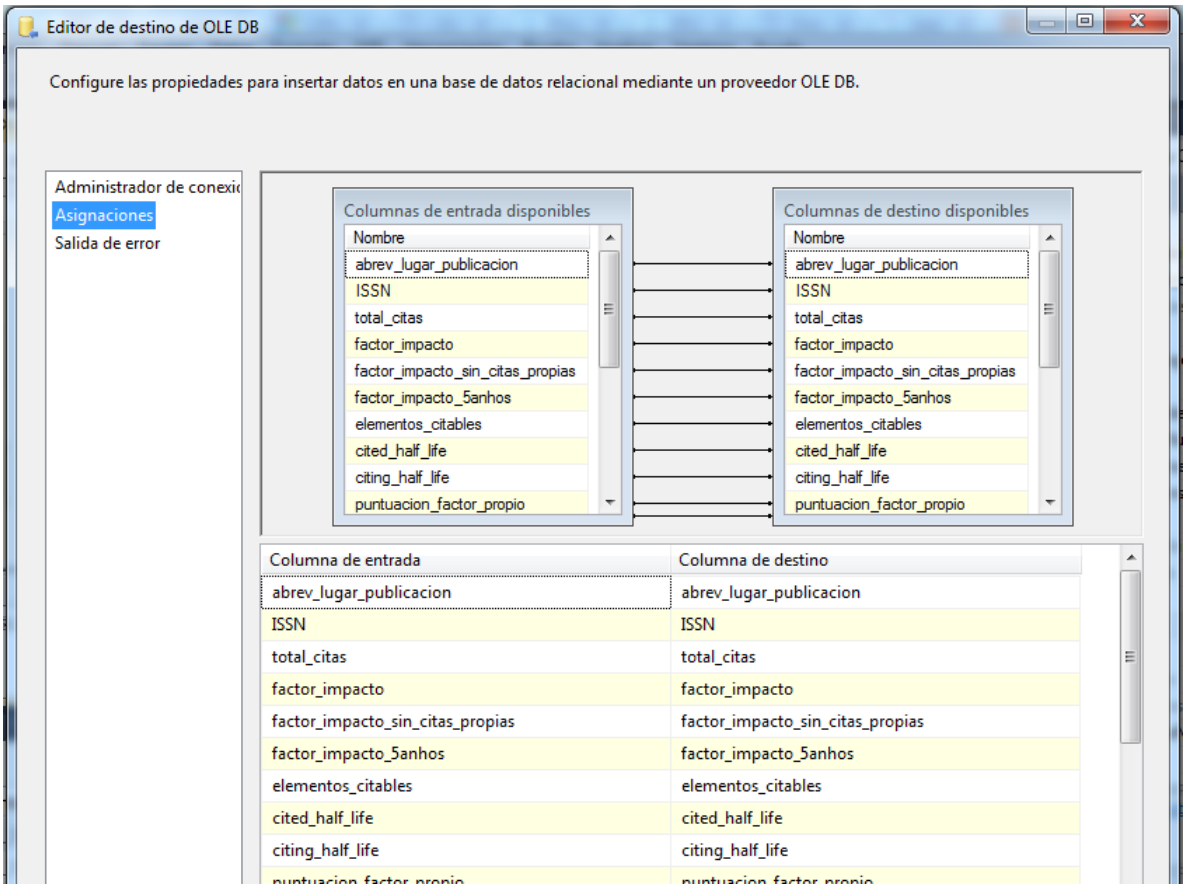


Figura 81: Asignaciones a la tabla Staging

En general, la carga de las demás fuentes de datos es muy similar, aunque puede cambiar la estructura del fichero, el tipo de archivo, la cantidad de columnas y el tipo de dato.

Es importante mencionar que normalmente por cada fuente de información puede existir más de un fichero de carga. Para solucionar esto, se crea un proceso ETL que lo que hace es llamar la carga del fichero de información por cada fichero que se encuentre en la ruta y cumpla con las características configuradas. Estos procesos ETL tienen en su nombre la palabra _completa.

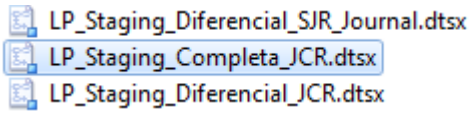


Figura 82: Ejemplo procesos ETL carga diferencial y completa

La estructura de estos procesos es la siguiente:

Figura 83: Estructura de una carga completa

En primer lugar trunca la tabla destino, luego entra a un ciclo (una "ejecución" por cada fichero en la ruta) y por cada entrada al ciclo llama la ejecución del proceso ETL diferencial (para este ejemplo, "LP_Staging_Diferencial_JCR.dtsx") de la carga del fichero JCR.

Solución ETL de Tablas Auxiliares

El siguiente paso de la solución ETL es la carga de las tablas auxiliares. Estas tablas son, en general son muy parecidas a las ETL de staging pero formateadas, es decir, sin nulos, sin caracteres extraños (por ejemplo, los caracteres que pone en los acentos cuando se carga con una codificación diferente a la que es el fichero). Estas tablas están limpias, por decirlo de alguna manera.

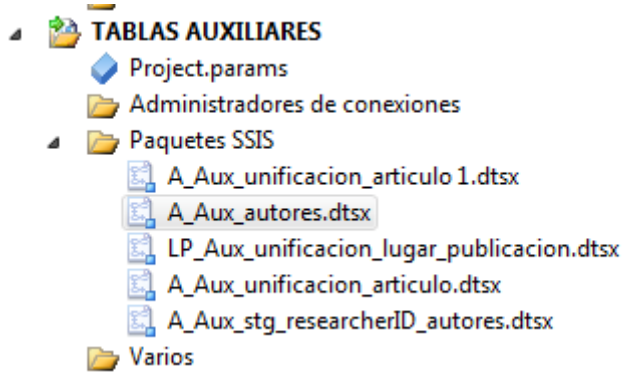


Figura 84: Ejemplo tablas auxiliares del proyecto

Los dos procesos ETL más importantes de esta fase son los de "A_Aux_unificacion_articulo.dtsx" y "LP_Aux_unificacion_lugar_publicacion.dtsx", puesto que son las tablas donde de agrupa la información de las fuentes de información de artículos y los lugares de publicación en dos únicas tablas. Los orígenes de dichos procesos son las tablas de staging explicadas en la sección anterior. A continuación se explicará una de ellas, la de unificación de artículos.

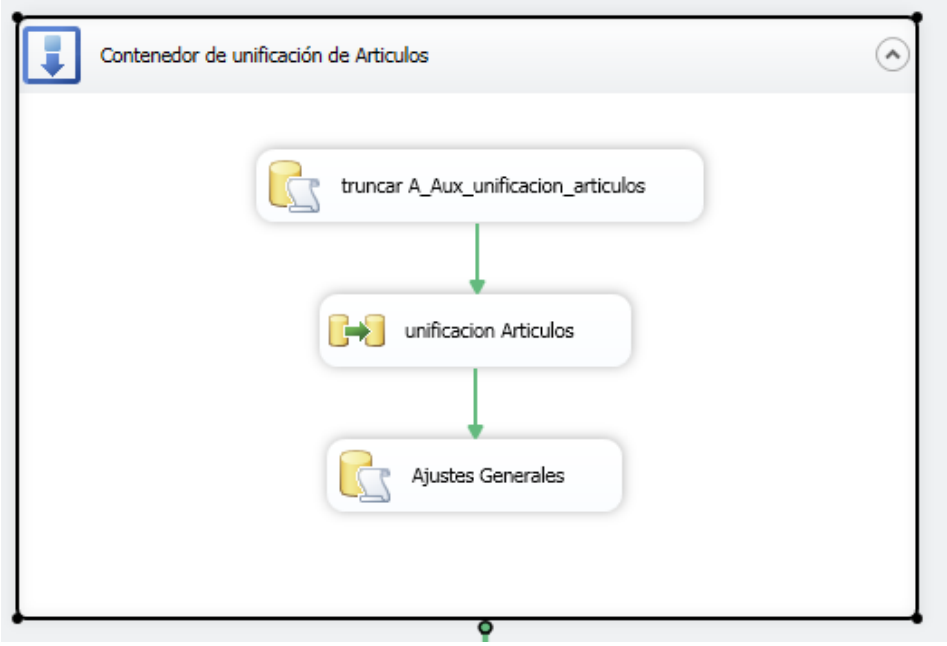


Figura 85: Estructura unificación artículos

El primer paso del contenedor de secuencias es truncar la tabla de unificación de artículos El segundo paso es el proceso de unificación, y el tercero es de ajustes generales. Estos ajustes son del estilo de reemplazar nulos, quitar espacios al inicio y final de los campos, etc.

El proceso de unificación de artículos es el siguiente:

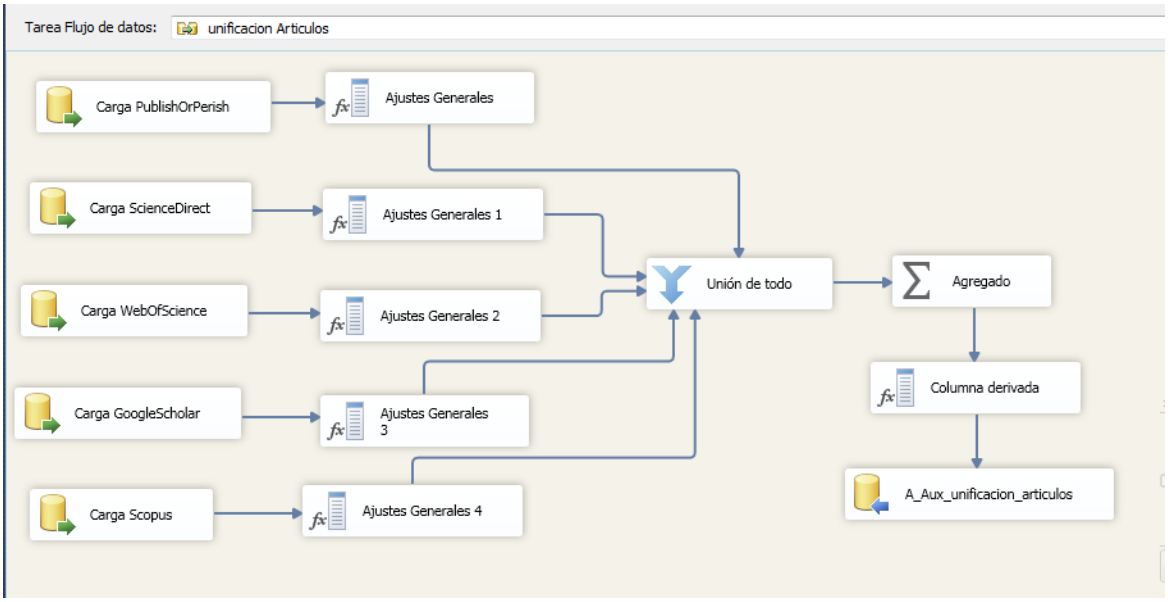


Figura 86: Unificación de las fuentes de información de artículos

En la primera parte se hace una carga de las tablas de *staging* de cada una de las fuentes de los artículos. En la segunda parte, de ajustes generales, se cambian los tipos de datos y se incorpora información adicional como por ejemplo la fuente del fichero, para saber de dónde proviene cada registro. Al abrir una caja de "Ajustes Generales" (una herramienta de columna derivada), tiene una apariencia como ésta:

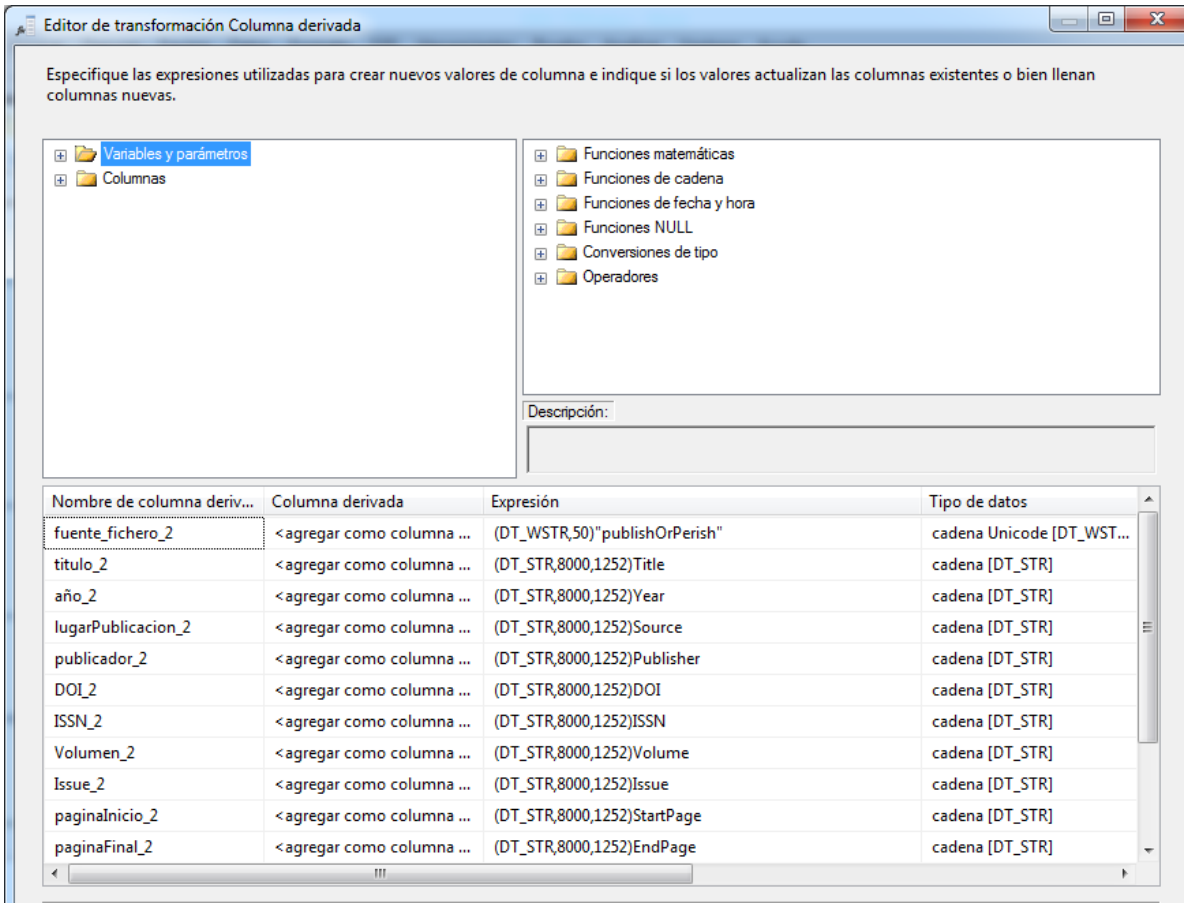


Figura 87: Columna derivada - ajustes generales

En la tercera fase está una caja de unión, que se encarga de unir toda la información de acuerdo a la columna de entrada, Por ejemplo, une todas las columnas de "lugarPublicacion" de las fuentes de información, dejando solamente una. Es importante garantizar que las columnas a unir, tengan el mismo tipo, de lo contrario la herramienta arroja error.

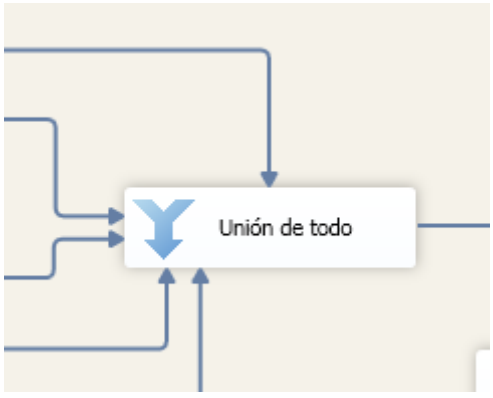


Figura 88: Caja unión de todo

Finalmente se encuentra la fase de agregado que es una agrupación de datos ya sea haciendo un "Group By", un Max(), Min(), etc. Por ejemplo, si la columna de "lugarPublicacion_2" tiene dos o más registros con el mismo contenido, lo agrupa dejando solo una. Finalmente, se inserta toda la información en la tabla de destino, para el ejemplo la tabla de destino es la tabla auxiliar de unificación de artículos.

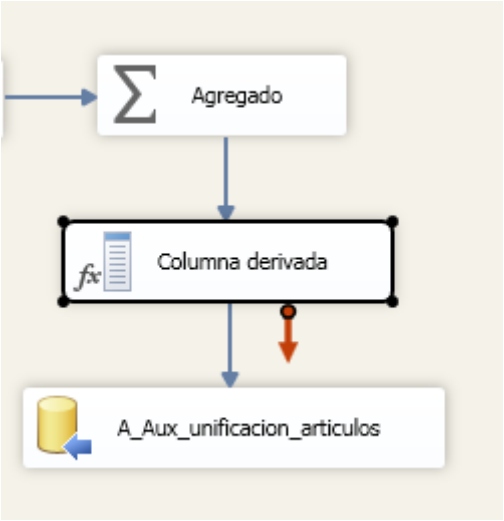


Figura 89: Caja de agregado, columna derivada y destino a base de datos

Dentro de los procesos ETL del trabajo fin de máster, está el proceso de *matching*, este proceso se ha explicado en el punto 3.4.1 Matching – Emparejamiento.