ETL Process Modeling Conceptual for Data Warehouses: A Systematic Mapping Study

L. Muñoz, J. N. Mazón and J. Trujillo

Abstract— **BACKGROUND: A data warehouse (DW) is an integrated collection of subject-oriented data in the support of decision making. Importantly, the integration of data sources is achieved through the use of ETL (Extract, Transform, and Load) processes. It is therefore extensively recognized that the appropriate design of the ETL processes are key factors in the success of DW projects. OBJECTIVE: We assess existing research proposals about ETL process modeling for data warehouse in order to identify their main characteristics, notation, and activities. We also study if these modeling approaches are supported by some kind of prototype or tool. METHOD: We have undertaken a systematic mapping study of the research literature about modeling ETL processes. A mapping study provides a systematic and objective procedure for identifying the nature and extent of the available research by means of research questions. RESULTS: The study is based on a comprehensive set of papers obtained after using a multi-stage selection criteria and are published in international workshops, conferences and journals between 2000 and 2009. CONCLUSIONS: This systematic mapping study states that there is a clear classification of ETL process modeling approaches, but that they are not enough covered by researchers. Therefore, more effort is required to bridge the research gap in modeling ETL processes.**

Keywords— **ETL process, modeling conceptual, data warehouse, systematic mapping studies.**

I. Introducción

En los años 90, Inmon [4] definió el término Almacén de Datos como: *"una colección de datos orientados por temas, integrados, variables en el tiempo y no volátiles para el apoyo de la toma de decisiones"*. Un AD es integrado, porque los datos que se introducen en el almacén se obtienen de una variedad de fuentes de datos (sistemas heredados, bases de datos relacionales, archivos COBOL, etc.). Para lograr la integración de esa variedad de fuentes se utilizan los procesos ETL. Dichos procesos son los responsables de la *extracción* de los datos a partir de las diversas fuentes de datos heterogéneas, de la *transformación* de estos (conversión, limpieza, etc.) y su *carga* en el AD. Por otro lado, se reconoce ampliamente que el diseño y mantenimiento de los procesos ETL son factores claves en el éxito de proyectos de AD [8], [21].

L. Muñoz, Universidad Tecnológica de Panamá, Panamá, lilia.munoz@utp.ac.pa
J. N. Mazón, Universidad de Alicante, España, jnmazon@dlsi.ua.es
J. Trujillo, Universidad de Alicante, España jtrujillo@dlsi.ua.es

Por su parte, un proceso ETL es extremadamente complejo, propenso a errores y consume mucho tiempo [19]. Además, se ha argumentado ampliamente, en la literatura, que los procesos ETL son costosos y que son una de las partes más importantes del desarrollo de un AD [4], [23]. En [17], se reporta que los costos de herramientas ETL y de limpieza de datos se estiman en al menos la tercera parte de los gastos del presupuesto de un AD. Por otro lado, en [2] se menciona que los procesos ETL representan el 80% de los recursos de desarrollo de un proyecto de AD. En este sentido, los procesos ETL son un componente clave de los ADs, porque los datos incorrectos producirán decisiones incorrectas, por esto un esquema correcto en la fase de diseño del AD es absolutamente necesario. Para especificar un proceso ETL existen diferentes maneras. Desde el desarrollo manual de programas específicos hasta el empleo de herramientas especializadas, como las de los proveedores tradicionales de base de datos desarrollados en los Sistemas Gestores de Base de Datos (SGBD) como: [3], [9], [15]. Estas herramientas se diseñan desde la perspectiva lógica y en la mayoría de los casos, presentan algunos inconvenientes: (i) falta de especificidad, (ii) dependencia de la plataforma destino, (iii) la configuración y estructura son muy complejas, (iv) requieren grandes requerimientos de hardware y, por lo tanto, no son factibles para pequeños y medianos proyectos. Estas características pueden hacer difícil la integración en ambientes heterogéneos de ADs. Por estas razones y, además, por la larga curva de aprendizaje, y el alto costo de adquisición y mantenimiento, muchas organizaciones prefieren desarrollar sus propios procesos ETL.

Dada la importancia de los procesos ETL dentro del marco de desarrollo de ADs, en este artículo se pretende brindar una aproximación a través de un mapeo sistemático de estudios, de cuáles son los diferentes enfoques de modelado de procesos ETL, características relevantes de cada uno de ellos, actividades, notaciones y problemática presentada en cada enfoque. Es con esta motivación que el estudio actual ha surgido de nuestro trabajo para recopilar, mapear y resumir los estudios primarios sobre modelado conceptual de procesos ETL para AD de manera precisa [7]. Un mapeo sistemático ofrece una visión sistemática de un área de investigación y permite evaluar la cantidad de pruebas existentes sobre un tema de interés [7] (véase, por ejemplo de Bailey et al. mapeo de estudios [1]).

El mapeo sistemático de estudios es una metodología que se utiliza con frecuencia en la investigación médica. El objetivo principal de un mapeo sistemático de estudios es

proporcionar una visión general de un área de investigación y de terminar la cantidad y tipo de investigación y los resultados disponibles. Requiere menos esfuerzo, que las revisiones sistemáticas, mientras que proporciona una visión de más alto nivel de granularidad.

El resto de este artículo está organizado de la siguiente manera. En la Sección II se incluye una breve descripción de lo que se entiende por procesos ETL. La Sección III describe el proceso de mapeo sistemático, incluida la estrategia de búsqueda y selección de los estudios. En la Sección IV, se describen los resultados. En la Sección V se presentan la validación de la evidencia, mientras que en la Sección VI se presentan las conclusiones.

II. Procesos ETL

Para Kimball [6] un proceso ETL es el fundamento de los ADs. Un proceso ETL bien diseñado extrae datos de las fuentes de datos, hace cumplir estándares de calidad de datos, a fin de que los datos puedan ser utilizados por los desarrolladores para las aplicaciones y los usuarios finales puedan tomar decisiones estratégicas. Es decir, los datos son extraídos de los sistemas fuentes, los cuales pasan por una secuencia de transformaciones antes de que se carguen en el AD. El repositorio de los sistemas que contienen las fuentes de datos para un AD puede variar desde hojas de cálculo hasta sistemas mainframe. Las transformaciones complejas son usualmente implementadas en programas procedimentales, ya sea fuera de bases de datos como por ejemplo (en C, Java, PASCAL, etc.) o dentro de base de datos (usando cualquier 4GL). El diseño de un proceso de ETL se compone generalmente de seis tareas definidas en [22]:

- Seleccionar los datos para la extracción: se definen los datos de las fuentes (generalmente provienen de diversas fuentes heterogéneas).
- Transformar las fuentes: una vez que los datos se hayan extraído de las fuentes de datos pueden ser transformados o esos nuevos datos pueden ser derivados. Algunas de las tareas más comunes de este paso son: filtración de datos, conversión de códigos, cálculos de valores derivados, transformación entre diversos formatos de datos, generación automática de números secuenciales (llaves derivadas), etc.
- Unir las fuentes: las diversas fuentes pueden unirse para ser cargadas al almacén como una sola fuente.
- Seleccionar el destino para la carga: el destino o los destinos son seleccionados para cargar los datos posteriormente.
- Unir los atributos de las fuentes de datos con los atributos del destino: los atributos (campos) que se obtuvieron de las fuentes de datos pueden ser mapeados con los correspondientes destinos.
- Cargar los datos: el almacén en poblado con los datos transformados.

III. Proceso del Mapeo Sistemático

El objetivo principal del mapeo sistemático de estudios llevado a cabo es dilucidar los enfoques de modelado conceptual de procesos ETL en ambientes de ADs. De tal manera que podamos tener un panorama amplio de los científica. No sólo para identificar las principales aproximaciones en esta área, sino también sus puntos fuertes y debilidades y, por supuesto, el trabajo futuro que puede llevarse a cabo para solventar posibles debilidades. Después de haber presentado el objetivo principal ahora vamos a definir las siguientes preguntas de investigación (PI):

(*PI1*) *¿A qué nivel de diseño son modelados los procesos ETL?* El modelado de procesos ETL es importante en el desarrollo de almacenes de datos, ya que permite tener una representación abstracta de alto nivel.

(PI2) ¿Qué lenguaje se utiliza para desarrollar el metamolelo de los procesos ETL? Elegir el lenguaje del metamodelo es importante para saber hasta qué punto el enfoque puede facilitar el diseño de tareas. Obviamente, lenguajes como el estándar UML tienen una menor curva de aprendizaje.

(PI3) ¿El enfoque define una metodología para la definición del proceso ETL? El diseño de procesos ETL es complejo y consume tiempo, por tanto resultaría una tarea más fácil si contempla algún tipo de directrices para garantizar que la ejecución final sea satisfactoria.

(PI4) ¿Está integrado el enfoque con un marco global para el desarrollo de AD? Es importante conocer si el modelado de procesos ETL está integrado con un marco global de desarrollo de AD o si son modelos de manera aislada.

(PI5) ¿Qué actividades ETL se describen en el enfoque? Los datos procedentes de fuentes de datos tienen que ser extraídos y transformados con el fin de garantizar su posterior carga en el AD de manera correcta. Para llevar a cabo estas tareas se pueden utilizar varias actividades como por ejemplo: agregación de datos, conversión de datos a un formato común, filtración de datos, y así sucesivamente. En términos generales, cuantas más actividades contemple el enfoque, sería mucho mejor.

A continuación se presentan las etapas del mapeo sistemático de estudios:

3.1 Definición del alcance, selección de la estrategia y criterios de selección.

El alcance de este estudio fue el siguiente: *Población*. Conjunto de artículos que describen los estudios sobre modelado conceptual de procesos ETL para ADs en la industria, la academia o en reportes de gobierno. *Intervención*. Cualquier estudio con métodos, metodologías, lenguajes de especificación sobre modelado conceptual de procesos ETL para ADs. *Resultados*. Cantidad y tipo de evidencia relativas al modelado conceptual de procesos ETL para ADs. *Diseño del estudio*. Experimentos, estudios de casos, relatos de experiencia, la investigación-acción.

La estrategia de búsqueda consistió en expresiones booleanas formadas por las siguientes palabras claves (en

inglés): *modeling, UML, ETL, processes, Extraction, Transformation, Load, data warehouse*, las cuales se generaron a partir de las preguntas creadas. La cadena de búsqueda básica se construyó a partir de las palabras claves, la misma fue *"modeling ETL processes"*. Algunos de los términos fueron desglosados en expresiones booleanas de tipo OR y AND, formadas por los sinónimos, como por ejemplo: *data warehouse, data warehousing*. También se utilizaron otras cadenas de búsqueda que incluyen otras palabras claves como por ejemplo: *Extraction Transformation Load, modeling ETL processes for data warehouse, designing ETL processes*.

Con respecto a la selección de las fuentes todas fueron digitales. Se seleccionaron estas fuentes, ya que incluyen motores de búsqueda y los artículos que ofrecen son de calidad, además son accesibles vía Web. Las fuentes a partir de las cuales se ejecutó el mapeo sistemático son las siguientes: *IEEE Digital Library, ACM Digital Library, DBLP, ScienceDirect (área de computación), SpringerLink*.

Para seleccionar los artículos, en primera instancia utilizamos los *criterios de inclusión* para hacer el análisis sobre el *título, resumen y las palabras claves*, obteniendo de esta manera el mayor número de trabajos que aportan contribuciones significativas sobre el modelado de procesos ETL. En segunda instancia utilizamos el *criterio de exclusión* donde nos centramos principalmente en el *resumen, introducción y conclusiones*, analizando un poco más aquellos trabajos que lo requerían para asegurarnos de que realmente eran relevantes para el campo de estudio. Con este criterio podemos ver con más detalle de que trata cada documento, ver la relación real que presenta con el objetivo buscado y si, verdaderamente es relevante para nuestro mapeo sistemático, seleccionarlo como estudio primario.

3.2 Selección de los estudios

Los estudios primarios proporcionan pruebas directas acerca de las preguntas de la investigación. El proceso de selección consta de cuatro iteraciones: las tres primeras se llevaron a cabo por tres revisores, mientras que la última iteración la realizó un evaluador. En la primera iteración, cada parte se examinó de forma independiente por dos revisores. Cada revisor aplicó los criterios de inclusión para cada trabajo, basado en el título, resumen y palabras claves. En la siguiente iteración, los artículos que fueron considerados por los revisores, son revisados nuevamente, ahora incluyendo la introducción y conclusiones. En la tercera iteración, dos revisores compararon los resultados y cuando no estaban de acuerdo sobre la inclusión de un documento, discutieron sus posiciones hasta llegar a un consenso. Esos documentos, que son considerados por los dos revisores son evaluados por un tercer revisor, en este caso revisando el artículo completo. La cuarta iteración está destinada a reducir la amenaza a validez interna de nuestros resultados. En la Tabla 1 se pueden apreciar los resultados de los estudios primarios y sus respectivas referencias.

IV. RESULTADOS Y ANÁLISIS

Luego de haber extraído y analizado la información relevante de cada estudio, en esta sección se presenta un resumen de cada uno de los estudios seleccionados, los resultados de cada uno de ellos, además de la comparación de las propuestas de acuerdo a las características y actividades presentadas.

TABLA I
RESULTADOS DE LOS ESTUDIOS PRIMARIOS Y SUS REFERENCIAS

Fuentes	Resultados	Estudios Primarios	Referencias
ScienceDirect	3	1	[10]
DBLP	6	2	[18], [22]
ACM Digital Library	111	1	[20]
IEEE Digital Library	100	1	[23]
SpringerLink	334	1	[11]

1) *Vassiliadis, P., Skiadopulos, S., Sellis, T., Conceptual Modeling of ETL Processes, in Proceeding of DOLAP. 2002: United States*.

El modelo presentado por los autores se centra en la definición de actividades ETL, se caracteriza por diferentes capas de instanciación y generalización. Para ello, se define una notación gráfica propia que permite capturar la semántica de los procesos ETL, además de presentar los principales elementos para el modelado de los procesos ETL (*concept, attribute, transformation, ETL-Constraint, etc*.). La propuesta es enriquecida con una *paleta* de actividades ETL frecuentemente utilizadas, es personalizada para la búsqueda de relaciones entre los atributos y las actividades ETL. Además, se propone un metamodelo propio que involucra un pequeño conjunto de constructos genéricos. La utilización de notación propia, permite el tratamiento de atributos "*first class citizens",* sin embargo esto podría resultar un problema, ya que un AD por lo general contiene cientos de atributos y por lo tanto un modelo de procesos ETL puede llegar a ser extremadamente complejo si cada atributo individual es representado como un elemento del modelo.

2) Simitsis, A., Vassiliadis, P., A Methodology for the Conceptual Modeling of ETL Processes, in CAiSE Workshops. 2003.

El trabajo propone una metodología basada en el enfoque propuesto en [23], la idea principal es completar el modelo a través del diseño de un conjunto de pasos, que conducirán a la meta básica, es decir la especificación de atributos interrelacionados. Estos pasos constituyen una metodología para el diseño conceptual de procesos ETL. El objetivo de la metodología es delinear un análisis de la estructura y contenido de las fuentes de datos existentes y de mapear éstas al almacén de datos. Los pasos de la metodología son los siguientes: i) identificar los almacenes de datos apropiados, ii) identificar candidatos y candidatos activos para participar en los datos almacenados, iii) mapear los atributos entre los proveedores y consumidores, iv) describir el diagrama con

restricciones en tiempo de ejecución. Esta propuesta muestra la misma notación gráfica propia para modelar procesos ETL presentada en [23].

3) Trujillo, J.L., S., A UML Based Approach for Modeling ETL Processes in Data Warehouses, in 22nd International Conference on Conceptual Modeling. 2003: USA, Chicago.

En el trabajo se identifica como objetivo principal permitirle al diseñador descomponer un proceso ETL complejo en un conjunto de procesos simples, lo cual facilitara el diseño y mantenimiento de procesos ETL. Para ello, proponen el uso de Diagramas de Clases UML para el modelado de procesos ETL, las actividades de los procesos están representadas por medio de clases estereotipadas, relacionadas entre sí por medio de dependencias UML. Desarrollan una paleta con las actividades que representan las tareas más comunes de un proceso (*aggregation, convesion, filter, incorrect, join, loader, log, merge, surrogate, wrapper).* Por otro lado, el uso de mecanismos de agrupación (paquetes UML [13]) facilita la creación y mantenimiento de procesos ETL complejos. Se destaca que la propuesta está totalmente integrada en un enfoque global de desarrollo de ADs basado en UML. Una debilidad de esta propuesta es que la presentación del modelo no es tan clara y fácil de entender para personal con diferente tipo de conocimiento (gerentes de empresas, administradores de bases de datos), sobre todo en las primeras etapas del modelado de ADs. Esto se debe principalmente al problema del encapsulamiento de clases, elementos muy cruciales en los procesos ETL, como los atributos y las relaciones entre estos.

4) Skoutas, D., Simitsis, A., Designing ETL processes using semantic web technologies, in DOLAP. 2006. p. 67-74.

Se propone un modelo conceptual de procesos ETL a través de tecnologías de la Web Semántica en este caso en particular Ontologías. Para la creación de la ontología se utilizó *Web Ontology Language (OWL).* Luego de creada la ontología que describe el dominio de aplicación y el mapeo entre la ontología y las fuentes de esquemas, se describen apropiadas transformaciones ETL para la integración de la data de las fuentes de datos y la carga de estas al AD, las cuales pueden ser derivadas semiautomáticamente. Se presenta un conjunto de operadores que son comúnmente utilizados en procesos ETL (*filter (σ), project (π), join (J), aggregation (γ), function (convert)).*

El primer paso de la propuesta es determinar qué información se necesita extraer de las fuentes, para poblar cada atributos/relación del AD. El siguiente paso es determinar que transformaciones son requeridas para integrar los datos de las relaciones de las fuentes de datos a las relaciones destino. Finalmente las transformaciones producidas se ordenan para que los esquemas de entrada puedan ser poblados exitosamente. Aunque la propuesta derive semiautomáticamente la transformación ETL, la misma no ha sido desarrollada para ser integrada en un marco formal de desarrollo de ADs, por lo que pudiera generar problemas de interoperabilidad e integración en las diferentes capas de un AD.

5) Luján, S., Vassiliadis, P., Trujillo, J., Data Mapping Diagrams for Data Warehouse Design with UML, in 23nd International Conference on Conceptual Modeling (ER). 2004. p. 191-204.

Los autores presentan una propuesta de modelado conceptual que intenta solucionar el problema de tratamiento de datos a niveles de granularidad muy bajos, que incluyen la definición de reglas de transformación a nivel de atributo, para ello proponen una extensión UML con un nuevo diagrama denominado diagrama de mapeo de datos (*Data Mapping*), el cual permite representar las reglas de transformación entre atributos necesarias, para modelar los procesos ETL a nivel conceptual. Para facilitar su manejo, se usó diagramas de paquetes de UML [13], obteniendo una propuesta que permite especificar los procesos ETL con diferentes niveles de detalle. En este sentido, los *Data Mappig* están organizados por niveles, i) en el Nivel 1 cada esquema del AD (p.e. fuentes de datos en el nivel conceptual en el esquema conceptual de origen (SCS), esquema conceptual del AD en el esquema conceptual del AD (DWCS)) son representados por paquetes, ii) en el Nivel 2 se describen las relaciones de entre las tablas de origen incluidas en el esquema y el destino en el AD, iii) en el Nivel 3 se describen las transformaciones intermedias y los controles a nivel de tabla, que tienen lugar durante el flujo, iv) en el Nivel 4 los Diagramas *Data Mapping* capturan el mapeo entre atributos. Una contribución de esta propuesta es que con el mecanismo extensión que permite UML se pueden modelar atributos como "*first class citizens"*. Por otro lado, la principal desventaja de esta propuesta es que no se ocupa de la derivación automática del mapping y las transformaciones ETL.

6) Li, Z., Sun, J., Yu, H., and Zhang, J., CommonCube-based Conceptual Modeling of ETL Processes, in International Conference on Control and Automation. 2005: Budapest, Hungria.

El trabajo de los autores se centra en el modelado conceptual de procesos ETL, proporcionando definiciones formales para entidades ETL (p.e. fuentes de datos, metadatos, funciones y tareas ETL, etc.). Para ello emplean *CommonCube*, permitiendo describir los esquemas de los cubos de datos, que son más compatibles con las estructuras de los ADs de datos para representar cubos, comparada con las tablas relacionales de otros modelos ETL. La utilización de *CommonCube* libera al diseñador de procesos ETL de la excesiva dependencia del esquema físico de los ADs. Basados en las funciones de restricción sobre los atributos de las fuentes y las operaciones de transformación de los atributos destino, se define el *mapping* ETL, que captura la semántica de las relaciones de los atributos de las fuentes y los atributos destino. Por otra parte, las entidades básicas (*Data source, CommonCube, Simplified data warehouse, Data target, Constraint function, Transforming operation, Correspondence y ETL mapping*) y actividades (*ETL task y ETL sesion*) son

formalmente definidas y organizadas de manera sistemática. Una de las desventajas de la propuesta es que no cuenta con una notación gráfica para representar los elementos de un proceso ETL, además esta propuesta no está integrada a un marco formal de desarrollo de ADs.

Por otra parte, en la Tabla II podemos observar una comparación con base a las características que presentan los estudios primarios. A partir de ella podemos inferir, que dos tipos de metamodelo están presentes en las propuestas: metamodelo estándar y metamodelo propietario. Sólo dos enfoques ([11,22]), utilizan estándares de desarrollo para sus propuestas, específicamente UML. A pesar de que las Ontologías permiten la formulación de un exhaustivo y riguroso esquema conceptual, y posibilitan una mejor comunicación, reutilización e inferencia computacional, solo una de las propuestas ([20]) utiliza este marco de modelado. Los demás estudios primarios utilizan metamodelo propietario, lo que puede generar problemas de interoperabilidad e integración en las diferentes capas de un almacén de datos. Por otro lado, solo dos propuestas están integradas en un marco formal de desarrollo de ADs. Nuestro punto de vista es que esta es la razón por la cual sólo las propuestas basadas en UML se integran en un marco de desarrollo formal de ADs. Otra característica importante, es que las propuestas de investigación para el modelado ETL proceso deben considerar la aplicación de una herramienta. Dicha herramienta debe apoyar el modelado del proceso de ETL. Sin embargo, sólo dos enfoques utilizan herramientas CASE para el modelado conceptual.

Un complemento para *Rational Rose* se desarrolla en [22] para el modelado de procesos ETL en el modelado conceptual. Este complemento se centra en la documentación de los procesos ETL.

TABLA II
COMPARACIÓN DE LAS CARACTERÍSTICAS DE LOS MODELOS ANALIZADOS

Características	***[23]***	***[22]***	***[18]***	***[20]***	***[11]***	***[10]***
Paleta	✓	✓	✓			
UML		✓			✓	
Ontología				✓		
Actividades	✓	✓	✓	✓	✓	✓
Metodología			✓			
Notación propia	✓		✓			✓
Herramienta		✓			✓	

En la Tabla III se presenta una comparación de las diferentes actividades de procesos ETL que contemplan las diferentes propuestas de modelado. De esta tabla podemos deducir, que las propuestas [20] y [22], realizan una mayor especificación de actividades. Además, de tener una gran cantidad de actividades en común. La propuesta de Li es la que realiza la especificación más pobre de actividades. Por otro lado, solo las actividades de *aggregation* y *filter* son tomadas en cuenta en todas las propuestas de modelado.

TABLA III
COMPARACIÓN DE LAS ACTIVIDADES DE PROCESOS ETL EN LOS MODELOS ANALIZADOS

Actividades	***[23]***	***[22]***	***[18]***	***[20]***	***[11]***	***[10]***
Aggregation	✓	✓	✓	✓	✓	✓
Join	✓	✓	✓	✓		
Conversion		✓		✓		
Filter	✓	✓	✓	✓	✓	✓
Surrogate	✓	✓	✓			
Merge		✓				
Concatenate				✓		
Load		✓				
Union	✓		✓	✓		
Distribute				✓		
Select				✓		
Log	✓	✓				
Wrapper	✓	✓				

V. VALIDACIÓN DE LA EVIDENCIA

Las principales amenazas a la validez de este mapeo sistemático se relacionan al sesgo en la selección de los estudios que deben incluirse y, en algunos casos, a las posibles inexactitudes en la extracción de datos.

Hemos considerado las cinco fuentes digitales mencionados. Con respecto a este punto, quizás el principal problema que enfrenta la validez del mapeo sistemático es la de si hemos podido encontrar toda los estudios primarios, aunque el alcance de las conferencias y revistas cubiertos por el mapeo es lo suficientemente amplio en el ámbito estudiado. No obstante, somos conscientes que es imposible lograr la integridad total. Puede que algunos documentos que existan no se hayan incluido, aunque la amplia revisión que se desarrolló y el conocimiento de este tema nos han llevado a la conclusión de que, si existen, probablemente no son muchos.

La información fue validada a través de la comparación de los resultados del análisis independiente entre los autores de los artículos. Por último, un criterio de inclusión y exclusión bien definido fue utilizado para seleccionar sólo los documentos más adecuados.

VI. CONCLUSIONES.

En este artículo hemos presentado un mapeo sistemático de estudios de los principales enfoques propuestos para el modelado conceptual de procesos ETL para ADs. Proporcionando un marco de trabajo actualizado, lo que nos permite formular nuevas actividades de investigación. El marco de revisión y el protocolo utilizados para la realización de esta revisión nos garantiza la completitud de los resultados.

Con relación a las propuestas de modelado conceptual, son pocas las aproximaciones encontradas. Lo que indica la necesidad de dedicar más esfuerzo al desarrollo de propuestas de modelado conceptual. Por su parte, el modelado conceptual de procesos ETL se ha realizado desde la perspectiva de las fuentes de datos, cuyos esfuerzos se centran en la representación y la descripción formal de las transformaciones necesarias. Sin embargo, las propuestas no contemplan una total automatización en la generación de código, como paso previo a la integración de un marco formal

de desarrollo de AD. Además, la representación de los modelos es estática lo que describe sólo las propiedades estructurales de los procesos, no existiendo así una representación dinámica que permita ver el comportamiento de un proceso ETL. Por otro lado, la mayoría de los estudios no utilizan estándares para el modelado de los procesos, lo que puede generar problemas de integración con las diferentes capas de un AD. Como conclusión, la carencia más importante que hemos identificado en el ambiente de modelado conceptual, es la definición de mecanismos formales para la obtención automática de código de procesos ETL, como base de la implementación final de un AD y de la integración de estos procesos en un entorno formal de desarrollo de ADs. Por otro lado, a nivel de desarrollo de software la comunidad científica está apostando por la utilización de estándares que permitan disminuir los tiempos y costos de desarrollo del software, como es el caso de *Archictecture Driven Model (MDA)* [14]. En este sentido, ninguna de las propuestas actuales de modelado de procesos ETL se enmarca en este contexto. En un trabajo futuro nuestro objetivo es presentar una propuesta de modelado de procesos ETL alineada con MDA.

REFERENCIAS

[1] J. Bailey, D. Budgen, M. Turner, B. Kitchenham, P. Brereton, and S. Linkman. Evidence Relating to Object-Oriented Software Design: A survey. In Proceedings of First International Symposium of Empirical Software Engineering and Measurement, IEEE, 2007

[2] M. Demarest. The politics of data warehousing. cited; Available from: http://www.hevanet.com/demarest/marc/dwpol.html.

[3] IBM. WebSphere DataStage. http://www-306.ibm.com/software/data/integration/datastage/.

[4] W. Inmon, Building the Data Warehouse. 1992: Press/John Wiley.

[5] M. Jarke, M. Lenzerini, Y, Vassiliou, and P. Vassiliadis. Fundamentals of Data Warehouses. Springer, 2000.

[6] R. Kimball, J. Caserta. The Data Warehouse ETL Toolkit. 2004: WileyPublishing.

[7] B. Kitcheham, T. Dyba, M. Jorgensen. Evidence-based software engineering, in Proceeding of the 26th Int. Conf. on Software Engineerig(ICSE, 2006), IEEE Computer Society. pp. 273-28

[8] S. March, A. Hevner. Integrated decision support systems: A data warehousing perspective, Decision Support Systems, Volume 43, Issue 3, 2007, 1031-1043.

[9] Microsoft. SQL Server 2005 Integration Services (SSIS). http://technet.microsoft.com/enus/sqlserver/bb331782.aspx

[10] Z. Li, J. Sun, H. Yu, and J. Zhang. CommonCube-based Conceptual Modeling of ETL Processes, in International Conference on Control and Automation. 2005: Budapest, Hungria.

[11] S. Luján, P, Vassiliadis, J. Trujillo. Data Mapping Diagrams for Data Warehouse Design with UML, in 23nd International Conference on Conceptual Modeling (ER). 2004. p. 191-204.

[12] S. Luján, and J. Trujillo. A data warehouse engineering process. In Tatyana M. Yakhno, editor, ADVIS, volume 3261 of Lecture Notes in Computer Science, pages 14-23. Springer, 2004.

[13] Object Management Group. Unifed Modeling Language: Superstructure: version 2.0, formal /05-07-04, 2005

[14] OMG. MDA Guide (draft version 2). http://www.omg.org/docs/omg/03-06-01.pdf(2003).

[15] Oracle. Oracle Warehouse Builder 10g. cited; Available from: http://www.oracle.com/technology/products/warehouse/.

[17] C. Shilakes, J. Tylman. Enterprise Information Portals. Enterprise Software Team http://sagemaker.com/company/downloads/eip/indepth.pdf.

[18] A. Simitsis, P. Vassiliadis. A Methodology for the Conceptual Modeling of ETL Processes, in CAiSE Workshops. 2003.

[19] A. Simitsis, P, Vassiliadis, T, Sellis. State-Space Optimization of ETL Workflows. IEEE Trans. Knowl. Data Eng. 17(10): 1404-1419 (2005)

[20] D. Skoutas, A, Simitsis. Designing ETL processes using semantic web technologies, in DOLAP. 2006. p. 67-74.

[21] M. Solomon. Ensuring A Successful Data Warehouse Initiative. Information Systems Management, 22 (1), 2005, 26-36.

[22] J. Trujillo, S, Luján. A UML Based Approach for Modeling ETL Processes in Data Warehouses, in 22nd International Conference on Conceptual Modeling. 2003: USA, Chicago, 307-320.

[23] P. Vassiliadis, S, Skiadopulos, T, Sellis. Conceptual Modeling of ETL Processes, in Proceeding of DOLAP. 2002: United States.

AGRADECIMIENTOS

Este trabajo es soportado por los proyectos ESPIA (TIN2007-67078) del Ministerio de Educación y Ciencia de España y QUASIMODO (PAC08-0157-0668) de la Consejería de Educación y Ciencia de Castilla-La Mancha, España. Lilia Muñoz dispone de una beca de la Secretaria Nacional de Ciencia, Tecnología e Innovación (SENACYT) y el Instituto para la Formación y Aprovechamiento de Recursos Humanos (IFARHU), de la República de Panamá.

Lilia Muñoz es graduada de Ingeniería de Sistemas Computacionales de la Universidad Tecnológica de Panamá (Panamá) y profesora de la Facultad de Ingeniería de Sistemas Computacionales de la Universidad Tecnológica de Panamá (Panamá). Muñoz actualmente es estudiante del Doctorado de Aplicaciones de la Informática de la Universidad de Alicante (España). Sus actividades de investigación incluyen modelado conceptual de procesos ETL, calidad de datos, validación de medidas. Ha publicado y presentado artículos en conferencias nacionales e internacionales tales como ADI, JISBD, CISTI, CLEI y DOLAP. Su correo de contacto es lilia.munoz@utp.ac.pa

Jose Norbeto Mazón es profesor ayudante doctor en el Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Alicante. Obtuvo su doctorado en Informática por la mismo universidad en el seno del grupo de investigación Lucentia. Ha publicado diversos artículos de investigación sobre almacenes de datos e ingeniería de requisitos en conferencias nacionales e internacionales (como DAWAK, ER, DOLAP, BNCOD, JISBD etc.) y en varias revistas como Decision Support Systems (DSS), SIGMOD Record o Data and Knowledge Engineering (DKE). Ha sido co-organizador del International Workshop on Business intelligencE and the WEB (BEWEB 2010) y del International Workshop on The Web and Requirements Engineering (WeRE 2010). Sus áreas de interés son: inteligencia de negocio, diseño de almacenes de datos, bases de datos multidimensionales, ingeniería de requisitos y desarrollo dirigido por modelos. Su correo es jnmazon@dlsi.ua.es

Juan Trujillo es profesor en la Escuela de Informática de la Universidad de Alicante (España). Trujillo obtuvo su Doctorado en Informática en la Universidad de Alicante (España) el año 2001. Sus intereses de investigación incluyen modelado de bases de datos, diseño conceptual de almacenes de datos, bases de datos multidimensionales, OLAP, y análisis y diseño orientado a objetos con UML. Ha publicado artículos en conferencias internacionales y revistas tales como ER, UML, ADBIS, CaiSE, WAIM, *Journal* de Gestión de Bases de Datos (JDM) e IEEE *Computer*. Participa como miembro de Comité de Programa de varios talleres y conferencias tales como ER, DOLAP, DSS, y SCI. También ha participado como revisor de varias revistas tales como JDM, KAIS, ISOFT y JODS. Su correo es jtrujillo@dlsi.ua.es.