Metodología ETL para el procesamiento de datos en repositorios de proyectos de software usando ontologías

Moises Gonzalez García
Departamento de ingeniería en software
CENIDET
Cuernavaca Morelos
moises@cenidet.edu.mx

Osvaldo Daniel Fernández Bonilla
Departamento de ingeniería en software
CENIDET
Cuernavaca Morelos
dekar81@cenidet.edu.mx

***Abstract*—Los repositorios de proyectos de software son una fuente valiosa de información para nuevos proyectos. Para obtener información relevante es necesario usar los procesos ETL. Sin embargo muchas herramientas actuales no manejan los metadatos generados durante las tareas de los procesos ETL o no dan soporte semántico a la tarea de integración. El siguiente artículo propone mediante el uso de ontologías, brindar una técnica ETL para pre-procesar almacenes de datos de proyectos de software que maneja los metadatos de las tareas ETL y de soporte semántico a la tarea de integración**

Keywords— ontologies, software repositories, etl process, semantic interoperability

I. INTRODUCCIÓN

Los repositorios de proyectos de software, guardan información de atributos usados en proyectos previos. Esta información, se guarda en un almacén de datos para futuras consultas o uso en nuevos proyectos. Los procesos de ETL (Extract, Transform, Load; extraer, transformar y cargar) son un método para extraer información de estos almacenes. Estos procesos se encargan de extraer y limpiar la información, convertirla al formato que requiera la herramienta de análisis a usar (cubo de datos, minería de datos, métodos estadísticos) y si es necesario reducir la cantidad de atributos que tienen para mejorar en lo posible los tiempos de búsqueda de datos con el mismo resultado que si se usaran los datos completos.

Sin embargo al momento de pre-procesar la información, muchas aplicaciones propietarias tiene problemas de compatibilidad con el manejo de los metadatos, al momento de pasar los metadatos de una tarea ETL a otra.

Otro problema es la interoperabilidad semántica en la tarea de integración, ya que se deja al usuario resolverlos. Esto problemas son: heterogeneidad semántica, redundancia de registros o de atributos, integración de esquemas y conflicto de valores.

Para resolver esto, esta investigación propone el uso de ontologías para el manejo de metadatos e interoperabilidad semántica en la etapa de integración.

II. TRABAJOS RELACIONADOS

Las siguientes investigaciones se seleccionaron debido a su aportación en las áreas de pre-procesamiento de datos en repositorios de software o uso de ontologías para el manejo de información.

En [1], se brinda un método para el pre-procesamiento de repositorios de datos médicos. Se usa una ontología como estructura de soporte formal de los diversos atributos de las bases de datos. Los datos inconsistentes se almacenan en la ontología mediante la supervisión de un experto y después se transformaban automáticamente según lo requiera el usuario. Los métodos de transformación también con una ontología.

En [2], proporcionan una técnica para ingresar reglas de negocio mediante ontologías como manejador del conocimiento. Las ontologías categorizan y etiquetan semánticamente los metadatos y brindan un esquema de datos unificado, mediante un manejo de los niveles léxicos y semánticos para así realizar los procesos de extracción y carga de forma autónoma.

En [3], se desarrolló un modelo conceptual para mejoramiento de procesos de software enfocados al producto. Para implementar este modelo, se realizó una seria de guía, estas guías son mapas conceptuales que el usuario sigue para realizar las diversas tareas de cada etapa del modelo. Posteriormente como resultado adicional, agregó una herramienta que unifica las dimensiones de requerimientos, procesos de desarrollo y métricas de procesos y productos.

III. OBJETIVOS DE LA INVESTIGACIÓN

Como objetivo general se propone crear una metodología para procesos ETL que utilice ontologías para resolver el problema de interoperabilidad semántica y manejo de metadatos y a su vez brinde una serie de atributos base para la búsqueda de conocimiento dentro de los repositorios de proyectos de software.

Como objetivos específicos se tienen los siguientes:

- Identificar y desarrollar las ontologías necesarias para: el manejo de interoperabilidad semántica en la tarea de integración de datos. Reusar si es posible, ontologías previamente desarrolladas de otras investigaciones previas.

- Identificar y desarrollar las ontologías necesarias para pre-procesamiento de datos y definir los axiomas correspondientes de dichas ontologías.

- Utilizar y adecuar los atributos de proyectos previos para comparación entre proyectos [4],[5], para que brinden soporte al proceso de pre-procesamiento de la información.

- Formalizar un metodología para procesos ETL en la cual se usen las ontologías desarrolladas

IV. METODOLOGÍA

Acorde a lo investigado hasta el momento, no hay una metodología para pre-procesar datos en repositorios de proyectos de software que maneje tanto la interoperabilidad semántica durante las tareas de la etapa de integración y los metadatos generados durante todas las tareas del proceso ETL. Así, las aportaciones de este proyecto son:

- Proporcionar una metodología de procesos ETL que utilice ontologías como herramienta para el manejo de la interoperabilidad semántica en la tarea de integración de datos y manejo de los metadatos durante todo el proceso ETL.

- Dar una serie de atributos base formalizados para usarlos en el proceso de extracción de información

- Brindar datos para un sistema de recomendación o de búsqueda de conocimiento orientado a ingeniería en software

Para realizar lo propuesto se define lo siguiente:

- Una serie de ontologías que definen los atributos de proyectos de software. Estas ontologías se definen como ontologías de áreas disciplinarias.

- Ontologías que definen las diversas formas para realizar las tareas ETL, preseleccionadas, en base a su utilidad para pre-procesar información de repositorios de proyectos de software. Estas ontologías se definen como ontologías de pre-procesamiento

Para construirlas las ontologías de las áreas disciplinarias, se propone lo siguiente:

- Para gestión de productos, se definirá acorde a los atributos vistos en [4],[5],[11]

- Para los procesos de desarrollo de software se unificaran conceptos de CMMI, SPEM, ISO/IEC 15504 e ISO/IEC 12207

- Para requerimientos, dado que cualquier proyecto tendría una serie de atributos muy específicos, se propone el uso de la ontología de lenguaje de requerimientos orientado a metas [176].

- Para el dominio del desarrollo de software y dominio del negocio se propone también una serie de atributos base, ya que no se pueden hacer de manera muy específica.

- Para las métricas de productos se usaran los estándares ISO/IEC 14598 y el ISO/IEC 9126 para la calidad del producto de software.

- Los elementos de los procesos de negocios se capturarán manualmente y deberán tener afinidad con los dominios de desarrollo y del negocio.

Para construir las ontologías de pre-procesamiento se pretender realizar lo siguiente:

- Realización de pruebas con los diversos métodos que existen para realizar cada tarea del proceso ETL. Obtener la información correspondiente y capturarla en la ontología para que el usuario pueda consultarla y usar los métodos que desee acorde a la tarea ETL que va a realizar.

La siguiente figura muestra de manera abstracta la ontología de pre-procesamiento.

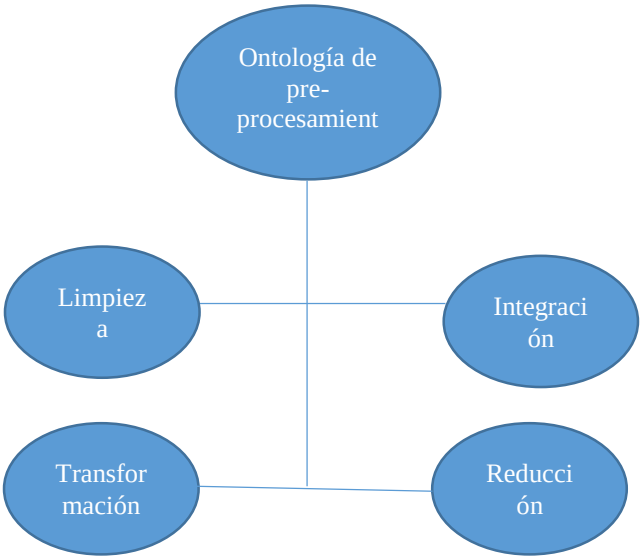

Figura 4.- Ontología de pre-procesamiento

Cada sección de la ontología brindaría una descripción de los conceptos, relaciones y reglas para las diversas tareas de pre-procesamiento que tiene el ETL acorde al tipo de dato que tiene de entrada.

Axiomas de la ontología de pre-procesamiento

La ontología de pre-procesamiento propuesta, usa los axiomas son para impedir fallos en recuperación de datos, ya que los ejemplares no reconocidos por los axiomas podrían producir una incorrecta implementación de los algoritmos encargados de resolver las diversas etapas de las tareas ETL. Se describen algunos tipos axiomas a desarrollar para la ontología propuesta.

Axiomas de limpieza

Aseguran que debe usarse un solo método de limpieza acorde a las condiciones que requiera dicho método

Axiomas de integración

Asegura que los ejemplares deban tener un solo tipo de valor es decir evitar multi-formatos

Asegurar que se tengan un solo tipo de escala numérica para los valores numéricos

Axiomas de transformación

Asegurar que se aplique un solo método de transformación según se requiera

Axiomas de reducción

Asegurar que se aplique un solo método de reducción según se requiera

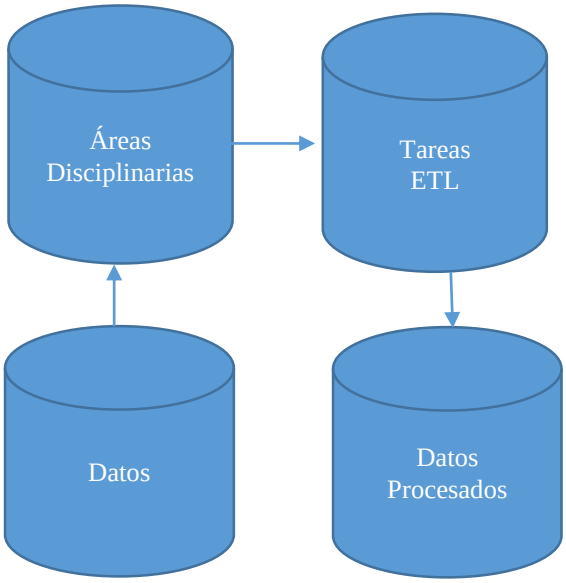

Figura 3.- Arquitectura propuesta para desarrollo de procesos ETL basados en ontologías.

V. RESULTADOS LOGRADOS

Los resultados obtenidos mediante el uso de la metodología propuesta, se pueden dividir en la ontología de áreas disciplinarias, la ontología de pre-procesamiento y el resultado obtenido al aplicar la metodología propuesta.

La ontología de áreas disciplinarias daría lo siguiente:

- Una serie de atributos base para extracción de información en repositorios de proyectos de software que puede reusarse en otros proyectos.
- Relaciones formales entre diversos atributos que hay para el desarrollo de proyectos de software.

La ontología de pre-procesamiento:

- Manejar los metadatos generados por cada tarea para pasarlos a la siguiente tarea al momento de realizar pre-procesamiento de datos.
- Brindar apoyo al usuario durante el proceso semántico que implica la tarea de integración de datos.
- Tareas ETL específicas para usar en repositorios de proyectos de software, evitando tener que probar todos los métodos existentes.

Y como resultado principal obtenido se tendría:

- Una metodología para el proceso ETL para pre-procesar repositorios de proyectos de software basados en ontologías que solucione el problema de interoperabilidad semántica y manejo de metadatos.
- Datos que pueden usarse para un sistema de recomendación de ingeniería en software o de búsqueda de conocimiento.

VI. CONCLUSIONES

Los repositorios de software ofrecen mucha información útil para nuevos proyectos. Sin embargo esta debe ser extraída y procesada antes de que pueda utilizarse correctamente en un nuevo proyecto. Varias investigaciones al momento de extraer información, definen una serie de atributos para extraer información pero estos atributos son seleccionados mediante experiencias personales y no con un método formal o investigación previa; otros solamente definen que estos repositorios tienen información útil que puede extraerse pero no definen una metodología para poder extraer dicha información.

Este trabajo mediante los atributos base propuestos y el uso de ontologías de pre-procesamiento y de áreas disciplinarias, propone resolver el problema del manejo de metadatos y el problema semántico de la tarea de integración de datos. Sin embargo aún sigue siendo el problema de relacionar los diversos atributos de las áreas involucradas en el desarrollo de proyectos de software, ya que debe realizarse manualmente; pero si es posible automatizar el manejo de la semántica de los atributos mediante el uso de ontologías y de gestionar los metadatos generados durante las tareas del proceso ETL.

RECONOCIMIENTOS

Agradecimientos para la maestra María del Rosario por su aportación de los repositorios de software y al CONACYT por el apoyo económico de manutención.

REFERENCIAS BIBLIOGRAFÍCAS

[1] David Pérez del Rey Tesis Doctoral “Un modelo de integración y preprocesamiento de información distribuida basado en ontologías” 2007

[2] Joel Villanueva Chávez Noviembre Tesis de Maestría “Marco de trabajo basado en ontologías para el proceso ETL” 2011

[3] Rini van Solingen Tesis Doctoral “Product Focused Software Process Improvement: in the embedded software domain” 2000

[4] Delgado Solís Cindy Tesis de Maestría “Caracterización de Proyectos de Software para Configurar su Desarrollo y Habilitar la Comparación entre Casos Almacenados en la Memoria Organizacional” 2008

[5] Sánchez Santamaría Miriam Tesis de Maestría “Evaluación de Técnicas de Comparación de Diferentes Grupos de Características de Proyectos de Software” 2010

[6] Coral Calero, Francisco Ruiz y Mario Piattini “Ontologies for Software Engineering and Software Technology” Libro Primera Edición Pag 339 2006

[7]Philip Nour, Harald Holz y Frank Maurer “Ontology-based Retrieval of Software Process Experiences” ICSE Workshop on Software Engineering over the Internet Pag 2 2000

[8] Ana C. O. Bringuente, Ricardo A. Falbo y Giancarlo Guizzardi “Using a Foundational Ontology for Reengineering a Software Process Ontology” Proceeding ER '09 Proceedings of the ER 2009 Workshops (CoMoL, ETheCoM, FP-UML, MOST-ONISW, QoIS, RIGiM, SeCoGIS) on Advances in Conceptual Modeling - Challenging Perspectives Pag 179 - 188 2011

[9] Claes Wohlin Per Runeson, Martin Host Magnus C. Ohlsson y Bjorn Regnell Anders Wesslen “Experimentation in Software Engineering” Libro Primera edición 2012

[10]ISO/IEC 9126 disponible en http://www.essi.upc.edu/~webgessi/publicacions/SMEF%2704-ISO-QualityModels.pdf

[11] Hans-Bernd Kittlaus y Peter N. Clough “Software Product Management and Pricing Key Success Factors for Software organizations” Libro Primera Edición