1. Introducción

Hoy en día existen muchas investigaciones relacionadas con el diseño y compilación de corpus lingüísticos electrónicos. Los corpus electrónicos son colecciones de documentos textuales en forma electrónica que constituyen una muestra del uso del lenguaje en algún ámbito general o específico. Es decir, conjuntos de documentos representativos de la forma en que se habla y se escribe algún idioma como el español o algún lenguaje de especialidad como el utilizado en las ingenierías.

Las investigaciones relacionadas con los corpus se hacen desde diversas perspectivas, ya que tienen muchas aplicaciones. Por ejemplo, la lingüística se ha beneficiado enormemente de la disponibilidad de estos recursos y las herramientas computacionales para analizarlos. Otro ejemplo importante está constituido por las investigaciones en computación para el procesamiento de grandes cantidades de textos. Esto último está muy en boga por el auge que vive Internet y los diversos desarrollos de minería de textos para descubrir información importante en grandes cantidades de textos.

En el Instituto de Ingeniería de la UNAM se lleva a cabo investigación para desarrollar estos recursos y hacer investigación básica y aplicada en lingüística y computación para, entre otras cosas, desarrollar tecnologías del lenguaje, tales

como diccionarios electrónicos, analizadores morfológicos y sintácticos, sintetizadores de voz, buscadores de contextos definitorios, etc.

En ese marco y mediante el financiamiento DGAPA PAPIIT de los proyectos IN400905 “Constitución del Corpus Histórico del Español de México” e IN402008 “Glutinometría y variación dialectal”, el Grupo de Ingeniería Lingüística desarrolla varios corpus electrónicos. El que atañe a esta tesis es específicamente el Corpus Histórico del Español en México, que consiste en diversos documentos escritos en los siglos XVI, XVII, XVIII y XIX, tanto en la Nueva España como en el México Independiente.

Los documentos de este corpus están codificados en XML (*Extensible Markup Language*, o en español, lenguaje de etiquetado extensible). Típicamente, existe una base de datos bibliográfica para concentrar en un lugar la información sobre autores, títulos de documentos, etc. de estos documentos. El presente trabajo contribuye a automatizar la construcción de esta base de datos. Específicamente, este trabajo consiste en el desarrollo de un constructor de bases de datos bibliográficas que analiza el esquema XML que valida los documentos del corpus para conocer la estructura de la información bibliográfica que servirá para dicha base.

Para llevar a cabo este trabajo, a continuación se enuncian los objetivos de esta tesis, la definición del problema a resolver y la metodología que se utilizará para llevar a cabo estas tareas.

Objetivo de la tesis

Desarrollar un constructor automático de bases de datos relacionales a partir de esquemas XML para ser pobladas con datos en documentos XML

Definición del problema

Se suele introducir manualmente información en bases de datos sujeta a errores de varios niveles. La ventaja de los documentos XML es que tienen la información pertinente ya codificada. Los esquemas XML, además de validar la estructura de esos documentos, representan la forma de la base de datos. Queremos construir dicha base a partir de estos esquemas y llenarla automáticamente con la información codificada en los documentos, evitando errores y automatizando el proceso.

Metodología

1. Se hará un modelado UML del desarrollo.
2. Se formulará un algoritmo que se irá mejorando iterativamente según se vayan incorporando las características del esquema.
3. Se desarrollará, instalará y comprobará el funcionamiento adecuado del constructor.

En los siguientes capítulos se expondrá la información básica sobre XML en general como lenguaje de codificación de datos en documentos (capítulo 2), la aplicación de la tecnología XML en el *Corpus Histórico del Español en México* (capítulo 3), además se hablará de SQL como lenguaje de consultas de bases de datos (capítulo 4). En el capítulo 5, se presenta la metodología seguida para el desarrollo del presente trabajo. En los siguientes capítulos (6 y 7), se describen

los comandos SQL para la construcción y población de la base de datos. En el capítulo 8 se presentan las conclusiones y en el apéndice al final se incluyen los diagramas UML generados y la bibliografía consultada.