Normalización de los LACCEI Proceedings a través de un proceso ETL

Jose Texier, Dr. en Ciencias Informáticas[1], Alberto Riba, Ing. en Sistemas[2], y Jusmeidy Zambrano, Especialista en Promoción de la Lectura y Escritura[3]

[1]*Universidad Nacional de Chilecito, Argentina, jtexier@undec.edu.ar, dantexier@gmail.com*

[2]*Universidad Nacional de Chilecito, Argentina, ariba@undec.edu.ar,* [3]*Universidad Nacional de Chilecito, Argentina, jzambrano@undec.edu.ar*

Resumen– ***La producción académica/científica de las instituciones se hace cada día más visible y accesible gracias a la dinámica era de la información en la que vivimos. En ese contexto, LACCEI con base en la Conferencia Anual recopila y publica los Proceedings (desde 2004 hasta 2016) en diferentes páginas web. Los LACCEI Proceedings están clasificados por autor, institución, país y área, y, a su vez, están almacenados como páginas web con diferencias en sus estructuras. Por tanto, en este trabajo se aplicó un proceso ETL (Extract, Transform and Load) para normalizar la información de los trece LACCEI Proceedings existentes, es decir, se extrajo la información, se transformó y se preservó. Gracias al desarrollo realizado, LACCEI tiene toda su producción (1625 artículos) en una base de datos que le permitirá ofrecer servicios a partir de los títulos, autores (con sus correos electrónicos), instituciones, palabras clave, resúmenes y PDF. La información normalizada puede presentarse como un repositorio institucional, desarrollarse un mapa conceptual, generar estadísticas u otras acciones que favorezcan la visibilidad y preservación de los Proceedings en el tiempo.***

Palabras clave– ***LACCEI, proceedings, ETL, extract, transform, load, PDF.***

I. Introducción)

La revolución de la investigación científica en el siglo XX fue un signo del cambio de época que vivimos hace algún tiempo, llamada, la "Era de la Información" [1]. El mayor impacto de los cambios atribuidos a esta "Era" se acentúan a partir de 1990 con la llegada de Internet: por primera vez en la historia el flujo de la información producida se volvió más rápido que el movimiento físico [2]. Adicionalmente, la filosofía del Acceso Abierto (en inglés *Open Access* -OA-) ha permitido que la producción científica de diversas instituciones sea cada día más visible y accesible sin restricciones legales, técnicas y de acceso [3]. De tal manera que se hace necesario (re)pensar el lugar que las universidades ocupan al generar conocimiento. Habría que preguntarse si las universidades y organizaciones educativas están dentro o fuera de los círculos de producción, visibilidad y difusión que la sociedad requiere.

Digital Object Identifier (DOI): http://dx.doi.org/10.18687/LACCEI2017.1.1.166
ISBN: 978-0-9993443-0-9
ISSN: 2414-6390

En este sentido, surge el interés por analizar cómo la producción y difusión del conocimiento en el área de la Ingeniería en una organización como LACCEI[1] -*Latin American and Caribbean Consortium of Engineering Institutions*- impacta en el ámbito educativo universitario. LACCEI, nació en el 2003, es una organización sin fines de lucro de instituciones de América Latina y el Caribe que ofrecen programas académicos en Ingeniería y Tecnología (universidades, colegios, escuelas y empresas) y con interés en otras partes del mundo. Su misión es ser una organización líder que aporte innovaciones en la educación y la investigación en ingeniería. A su vez, impulsa asociaciones entre la academia, la industria, el gobierno y organizaciones privadas para el beneficio de la sociedad y de las naciones. Por ello, LACCEI organiza una conferencia anual donde se presentan trabajos académicos/científicos definidos de acuerdo con la misión estratégica del consorcio que se publican en los ***LACCEI Proceedings***[2] como las colecciones anuales clasificadas por autor, institución, país y área.

Los *LACCEI Proceedings* se publican los Proceedings desde el 2004 en páginas web, con un formato innovador para el momento de su creación, pero con el pasar del tiempo presenta problemas para ser encontrados por los diferentes buscadores de producción académica/científica como Google Scholar, Microsoft Academic Search, entre otros. Por tanto, surge la necesidad de organizar los trabajos presentados en un formato que garantice el acceso, gestión, reutilización, difusión y preservación de los trabajos. De esta manera se podrían realizar análisis bibliométricos, establecer indicadores propios, identificar fortalezas, consultas, etc.

En este contexto el objetivo central de este trabajo es la normalización de la información de los *LACCEI Proceedings* a través de un proceso ETL (en inglés *Extract, Transform and Load)*. Este proceso de normalización consiste en almacenar la información de las trece conferencias realizadas (desde 2004 hasta 2016) en una base de datos normalizada. Para tal fin, en este artículo se explica el diseño del proceso de ETL; el modelo de datos usado; las fases de extracción, transformación y carga y los trabajos futuros que plantean las opciones de visualización para los *Proceedings*. El impacto de esta normalización conllevará una visibilización sistemática

[1] http://www.laccei.org/

[2] http://www.laccei.org/index.php/publications/laccei-proceedings

de los proceedings anuales para garantizar la preservación en el tiempo de la información.

II. Arquitectura ETL

Desde el 2004 hasta el 2016 se realizaron trece (13) *LACCEI Proceedings*. Cada uno de ellos está en una página web distinta generada cada año a partir de archivos excel y adaptada a una estructura que los clasifica por país, autor, área e institución. El problema radica en que al tratar de extraer en forma automática la información surgen muchas dificultades por la estandarización.

La propuesta de este trabajo fue extraer la información de cada uno de los artículos y preservarlos en una base de datos normalizada, que luego se pueda exportar con la información deseada a formatos específicos de acuerdo con las necesidades que requiera el consorcio. Se adaptó el patrón de arquitectura de software ETL a todo el proceso de extracción de la información, normalización y preservación de cada una de las conferencias de LACCEI. ETL es un patrón arquitectónico del software para la integración de datos [1, 2]. El proceso consta de tres etapas funcionales consecutivas: la extracción de datos de diferentes fuentes, las transformaciones necesarias por la heterogeneidad de la información a través de reglas, y la carga final en un sistema de base de datos que, en muchos casos, es conocido como Data Warehouse [3]. En la Figura 1, se observa un resumen del diseño desarrollado. Esta arquitectura se utiliza principalmente en software empresarial para unificar la información utilizada para los procesos de inteligencia de negocios que conducen a la toma de decisiones [4]. El diseño se basa en las siguientes premisas:

- Permitir la recolección desde múltiples tipos de fuentes de datos.
- Llevar los recursos a una representación abstracta con el fin de normalizar la información, es decir, lograr el almacenamiento de la misma.
- Poder generar información estadística sobre el estado de la información recolectada.

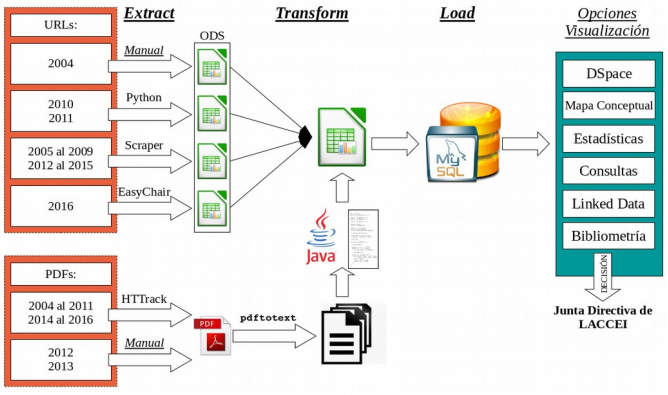

Figura 1. Resumen del escenario ETL desarrollado

III. Modelo de datos

El modelo de base de datos determina la estructura lógica del almacenamiento de la información. El paradigma elegido es el relacional por su amplia difusión en la actualidad. Al definir el modelo de datos es recomendable disponer dentro del sistema ETL de una nomenclatura que permita una mejor gestión y comprensión de todo el sistema. La Figura 2 muestra gráficamente la estructura del modelo de datos diseñado para nuestro sistema. El principal problema al desarrollar una herramienta ETL está cuando los datos poseen distintos formatos, por eso, el proceso de integración requiere de una correcta identificación y análisis de estas fuentes de datos que sirvió como punto de partida para definir un modelo de datos para el sistema ETL.

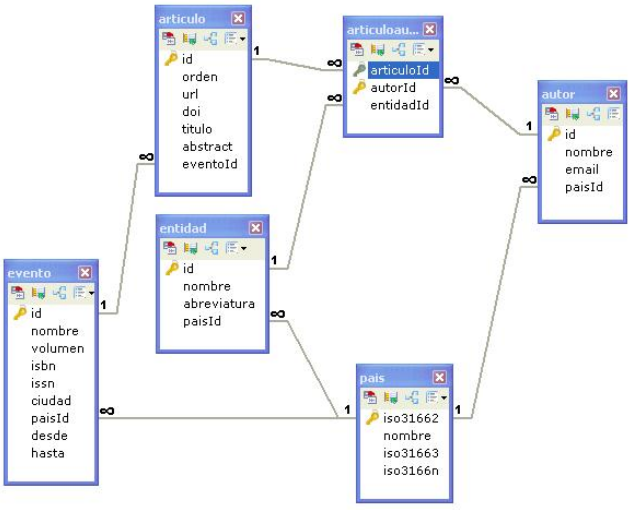

Figura 2. Modelo de datos diseñado

El modelo de datos provee la estructura lógica a un único repositorio centralizado que almacena los datos provenientes de las distintas fuentes luego de su reestructuración y limpieza. Esta base de datos puede ser utilizada para la generación de reportes, los procesos de análisis de la información, toma de decisiones y también puede actuar como una fuente para diversas aplicaciones.

IV. Extracción

El primer paso en cualquier escenario ETL es la extracción, ya que es la responsable de obtener los datos de los sistemas fuentes. Cada fuente de datos tiene un conjunto distinto de características que necesitan ser gestionadas eficazmente para el proceso ETL. Durante la extracción de datos de los 13 *LACCEI Proceedings* se aplicaron seis subprocesos:

- Para la conferencia del 2004, la extracción se realizó manualmente porque las 131 páginas web que representan los 131 trabajos presentados, en su gran mayoría, no contaban con una estructura definida. La información obtenida fue: título, palabras clave, autor, correo electrónico, resumen, institución, URL e ID de la conferencia. Se almacenó en un archivo ODS[3] (Open Document Spreadsheet).

[3] http://www.opendocumentformat.org/

- Para las conferencias del 2005 al 2009 y del 2012 al 2015, se usó un "Google Chrome Extension" llamado Web Scraper[4]. Este componente permite extraer información que está bien estructurada en código HTML. La información obtenida fue: título, autor, institución, URL e ID de la conferencia, que se almacenó en un archivo ODS.
- En las conferencias del 2010 y 2011, se usó un script realizado en Python [5]. La información obtenida fue: título, autor, institución, URL e ID de la conferencia. Se almacenó en un archivo ODS.
- La información de la conferencia del 2016 se obtuvo directamente de EasyChair[5], el software administrador de la conferencia. Los siguientes campos obtenidos: título, autor, institución, URL e ID de la conferencia. Se almacenó en un archivo ODS.
- Los PDF que representan cada una de las conferencias del 2004 al 2011 y del 2014 al 2016, se obtuvieron vía HTTrack Website Copier[6], un software que permite bajar todo el sitio web a una computadora local de forma recursiva.
- Los PDF para las conferencias del 2012 y 2013, se obtuvieron de forma manual, ya que los enlaces no llevan directamente al PDF sino a otras páginas web.

V. Transformación

El segundo paso en un escenario ETL es la transformación de datos obtenidos en la extracción. El paso de transformación tiende a realizar algo de limpieza y consolidar los datos entrantes de diferentes fuentes, de tal manera de contar con datos precisos que sean correctos, completos, consistentes y sin ambigüedad. De la fase anterior se obtuvieron cuatro archivos ODS, los cuales se consolidaron en uno solo (en formato ODS). Solamente para la conferencia del 2004, se obtuvieron los correos electrónicos, palabras clave y los resúmenes. Entonces, a partir de los PDF se extrajo todo el texto con el comando "`pdftotext`" en Linux, pero a través de un Bash[7] todo se consolidó en un archivo de texto, que luego es la entrada de un código en JAVA para poder extraer los correos electrónicos, palabras clave y resúmenes de todos los trabajos del 2005 al 2016. En un primer análisis para extraer el texto de los archivos PDF se evaluaron las herramientas: KEA[8], Mr. Dlib[9], Alchemy[10] (ahora en IBM Watson) y ParsCit[11], pero finalmente se decidió realizarlo como se explicó anteriormente [6, 7], es decir, hacerlo a medida de las necesidades planteadas.

Luego de la consolidación en un solo archivo de los 1625 trabajos presentados en todas las conferencias de LACCEI, se inició un proceso de depuración de la información para los siguientes campos:

- El campo autores, que en los casos que existiera más de uno, se realizaba un separación de las cadenas de autores.
- Para cada autor, la afiliación o institución se encuentra entre paréntesis, entonces se procedió a eliminarlo y extraerlo. De la misma forma se realizó con los países.
- Para cada trabajo (menos los trabajos del 2004), se inició un *match* para agregar los correos electrónicos, palabras clave y resúmenes, los cuales se obtuvieron del programa en JAVA con el texto generado del comando de Linux. El programa en JAVA es un desarrollo propio [5].

VI. Carga

Cargar los datos a un sistema de bases de datos es el paso final del proceso ETL. En este paso, los datos extraídos y transformados se adaptan a la estructura definida en una base de datos MySQL para que los usuarios finales y los sistemas de aplicación diseñados puedan acceder. Esta actividad incluye una gran variedad de acciones de acuerdo con los requerimientos planteados. En esta fase se interactúa directamente con la base de datos destino y al realizar esta operación se aplican todas las restricciones y triggers definidos para respetar valores únicos, integridad referencial, campos obligatorios, chequeo de rangos de valores. Es muy importante tener en cuenta estas restricciones y triggers se hacen para garantizar la integridad y calidad de los datos en el proceso ETL.

En la fase de transformación se generó un archivo de tipo ODS que integra toda la información proveniente de las distintas fuentes de datos, donde cada registro contiene la información referente a cada artículo (título, abstract, palabras claves, autores, instituciones, países, url, conferencia, etc.). Estos datos fueron cargados en la base de datos donde se normalizaron, proceso en el que se dividen los datos en distintas tablas, se generan claves principales y foráneas y se establece la integridad referencial, procedimiento implementado utilizando procedimientos almacenados y triggers con la aplicación phpMyAdmin[12]. El detalle actividades realizadas fue el siguiente:

- Importación de los datos resultantes de la fase de transformación disponibles en un único archivo con formato ODS a la base de datos.
- Ejecución de procedimientos almacenados y triggers para la división de los datos en las tablas correspondientes.
- Creación de las claves principales.
- Chequeo y eliminación de valores duplicados de las tablas de autor e institución.

[4] http://webscraper.io/

[5] http://easychair.org/

[6] https://www.httrack.com/

[7] http://manpages.ubuntu.com/manpages/wily/man1/pdftotext.1.html

[8] http://www.nzdl.org/Kea/

[9] http://mr-dlib.org/

[10] https://www.ibm.com/watson/developercloud/natural-language-understanding.html

[11] http://wing.comp.nus.edu.sg/parsCit/

[12] https://www.phpmyadmin.net/

- Creación y carga de la tabla de países y sus respectivos códigos ISO 3166[13] con abreviaturas de dos y tres letras.
- Actualización de la tabla intermedia articuloautor.
- Creación de claves foráneas para establecer la integridad referencial y garantizar la integridad y calidad de los datos en el proceso ETL.

VII. Trabajos futuros

Sobre la base de los resultados obtenidos en la propuesta desarrollada, surgen algunas líneas de trabajo futuros, tales como:

- El trabajo realizado es una solución muy puntual para el consorcio de LACCEI, pero puede replicarse en organizaciones con necesidades similares, de manera tal que puede realizarse un análisis comparativo entre ellos.
- El esquema propuesto por basarse en ETL, resulta muy amplio y abarcativo; por ello, se decidió dejar a futuro el desarrollo de componentes de software que fortalecerán tal iniciativa (con licenciamiento libre), por ejemplo: realizar el proceso de extracción de información de artículos en páginas web estáticas y/o de los PDF de los artículos.
- El trabajo realizado permite tener una base de datos normalizada, la cual puede dar comienzo a la elaboración de un ranking bibliométrico para las instituciones miembros de LACCEI, a partir de datos como: medida de investigadores / profesores en la universidad con producción, medida de producción por investigador, relación de la producción con el factor de impacto y el índice Scopus (SJR o SNIP), etc., archivos ricos (PDF, doc/docx, ppt/pptx, ps/eps), enlaces entrantes a los sitios web de las universidades, documentos en Google Scholar, entre otros datos que se puede generar en la base de datos.
- Desarrollar un conjunto de estrategias necesarias para optimizar la visibilidad científica de LACCEI.
- La visibilidad web de trabajos académicos y científicos están siendo determinados por indicadores relacionados con, entre otros aspectos, la disponibilidad web de los artículos en Open Access en Repositorios Institucionales y en revistas [8]. Resulta lógico pensar que esta propuesta contribuye con la comunicación de la ciencia generado en el contexto de LACCEI, además, permite la preservación de la información en el tiempo y el reconocimiento por parte de los pares de los avances y estudios que se hacen en las diferentes instituciones del consorcio. Por ello, a partir de una base de datos normalizada, LACCEI puede visualizar la información de varias formas:
  - En un repositorio digital realizado en DSpace[14], donde cada trabajo o ítem estará asociado con colecciones y comunidades. Entonces, se puede hablar del repositorio institucional de LACCEI o de la biblioteca digital de LACCEI, ya que con el pasar de los años se han entrelazado estos dos conceptos [9, 10], por tanto, es es indistinto el término que se use.
  - Desarrollar un mapa conceptual de toda la producción científica de LACCEI, gracias a la extracción del título, resumen y palabras claves de cada producto de la producción científica relevada, la cual puede estar organizada por áreas de interés, por países, por instituciones, etc [11]. Adicionalmente, se puede hacer una análisis de procesamiento de lenguaje natural [12] con otros fines.
  - Generar estadísticas clasificadas por año, país, institución (afiliación), por tema, e incluso un análisis sintáctico para generar nubes de palabras. Ver Figura 3.
  - Implementar una interfaz de consulta, una alternativa diferente al DSpace, la cual puede estar en la página web principal de LACCEI y disponible para miembros especiales del consorcio.
  - Realizar un proyecto de Linked Open Data (LOD), donde pueden estar relacionados los trabajos de los Proceedings con información proveniente de las Universidades, de los países (factbook de la CIA[15] o del World Bank[16]) o con términos establecidos en el catálogo WorldCat[17], VIAF[18], Wikipedia[19], por ejemplo.

[13] http://www.iso.org/iso/home/standards/country_codes.htm

[14] http://www.dspace.org/

[15] https://www.cia.gov/library/publications/the-world-factbook/

[16] http://databank.worldbank.org/data/home.aspx

[17] https://www.worldcat.org/

[18] https://viaf.org/

[19] https://www.wikipedia.org/

VIII. Conclusiones

Este trabajo se planteó como objetivo general normalizar los *LACCEI Proceedings* a través de un proceso ETL, de tal manera que para dar respuesta a tal propósito se vincularon diversas áreas de la disciplina de las Ciencias de la Computación, a saber: recuperación de información, procesamiento de lenguaje natural, base de datos, programación, entre otras. Por ello, en las siguientes líneas se presentan las conclusiones con base en lo realizado:

- Se emplearon diversos lenguajes con distintos propósitos (Java, Bash, Python, SQL, XML) que hicieron de la propuesta una alternativa compleja y disímil bajo el enfoque ETL.
- Todas las herramientas y lenguajes de programación usados están en software libre o su uso es libre, tales como: los lenguajes con distintos propósitos enumerados anteriormente, Google Docs, LibreOffice Writer, LibreOffice Calc, Zotero, MySQL, HTTrack Web, Ubuntu 16.01, PhpMyAdmin, etc.
- A partir del trabajo desarrollado, LACCEI puede tener conocimiento exacto de la producción científica desarrollada, personal que la realizó, institución involucrada, etc. Además, puede reconocer el posicionamiento web y hacer seguimiento de la producción científica.

Referencias

[1] M. R. De Giusti, A. J. Lira, and N. F. Oviedo, “Extract, transform and load architecture for metadata collection,” presented at the VI Simposio Internacional de Bibliotecas Digitales (Brasil, 2011), 2011.

[2] “The Management of Conformed ETL Architecture - ProQuest.” [Online]. Available: http://search.proquest.com/openview/ceeacf226e1475d8e29fd5b0bcc33ff7/1?pq-origsite=gscholar. [Accessed: 16-Dec-2016].

[3] M. M. I. Awad, M. S. Abdullah, and A. B. M. Ali, “Extending ETL framework using service oriented architecture,” *Procedia Comput. Sci.*, vol. 3, pp. 110–114, Jan. 2011.

[4] S. H. A. El-Sappagh, A. M. A. Hendawi, and A. H. El Bastawissy, “A proposed model for data warehouse ETL processes,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 23, no. 2, pp. 91–104, Jul. 2011.

[5] J. Texier, A. E. Riba, and J. Zambrano, “ETL - LACCEI 2017,” 2016. [Online]. Available: https://github.com/dantexier/ETL_LACCEI. [Accessed: 06-Dec-2016].

[6] A. Casali, C. Deco, C. Bender, and S. Fontanarrosa, “Extracción Automática de Metadatos de Objetos Digitales Educativos,” *Conf. LACLO*, vol. 5, no. 1, Nov. 2015.

[7] A. G. Ororbia II, J. Wu, M. Khabsa, K. WIlliams, and C. L. Giles, “Big Scholarly Data in CiteSeerX: Information Extraction from the Web,” in *Proceedings of the 24th International Conference on World Wide Web*, New York, NY, USA, 2015, pp. 597–602.

[8] J. Texier, “Notas metodológicas para cubrir la etapa de documentar una investigación,” Proyecto de Enlace de Bibliotecas (PrEBi), Aug. 2011.

[9] J. Xia and D. B. Opperman, “Current Trends in Institutional Repositories for Institutions Offering Master’s and Baccalaureate Degrees,” *Ser. Rev.*, vol. 36, no. 1, pp. 10–18, Mar. 2010.

[10] J. Texier, “Los repositorios institucionales y las bibliotecas digitales: una somera revisión bibliográfica y su relación en la educación superior,” presented at the 11th Latin American and Caribbean Conference for Engineering and Technology - 2013, Cancun, Mexico, 2013, p. 9.

[11] F. Ronzano and H. Saggion, “Knowledge Extraction and Modeling from Scientific Publications,” 2016.

[12] C. D. C. Ta and T. P. Thi, “Automatic Extraction of Semantic Relations from Text Documents,” in *Future Data and Security Engineering*, 2016, pp. 344–351.