Diseño de un data warehouse para medir el desarrollo disciplinar en instituciones académicas

Salvador Gorbea Portal
María de Jesús Madera Jaramillo*

Artículo recibido:
15 de enero de 2014.
Artículo aceptado:
13 de octubre de 2015.

Resumen

Se presentan los resultados sobre la experiencia adquirida en el diseño de un *data warehouse*, orientado a la medición del desarrollo disciplinar en las ciencias bibliotecológica y de la información en instituciones académicas de Iberoamérica, mediante la explicación de una síntesis gráfica de este proceso de diseño, orientada a la descripción y análisis de sus características, estructura, arquitectura y operaciones de funcionamiento del entorno en el cual se implementa.

* UNAM, Instituto de Investigaciones Bibliotecológicas y de la Información, México.
portal@unam.mx ramadera@yahoo.com.mx

Palabras clave: Data Warehouse; Desarrollo Disciplinar; Ciencias Bibliotecológica y de la Información; Indicadores Científicos

Abstract

Design of a data warehouse to measure disciplinary development in academic institutions

Salvador Gorbea-Portal and *María de Jesús Madera-Jaramillo*

Results drawn from the knowledge acquired in the experience of designing a data warehouse, erected to measure the degree of disciplinary development in Library and Information Science in academic institutions of Ibero-America, are presented through an explanation of a graphical synthesis of the design process, which is focused on a description and analysis of features, structure, architecture and operational functioning of the environment in which it is implemented.

Keywords: Data Warehouse; Development Discipline; Library and Information Science; Scientific Indicators

Introducción

El desarrollo disciplinar en las ciencias bibliotecológica y de la información ha sido irregular en el ámbito mundial y su distribución geográfica presenta un panorama bastante desigual entre las economías desarrolladas y las que se encuentran en vías de desarrollo. Esta desigualdad, aparentemente obvia, propicia la falta de interés por medir los niveles de desarrollo que han alcanzado los países de uno y otro bloque en estas disciplinas y más aún por la creación de mecanismos e instrumentos que permitan medir su comportamiento, así como la realización de investigaciones tendientes a revelar sus regularidades por países.

La falta de interés por medir el desarrollo disciplinar en este campo de conocimiento, como en cualquier otro, genera un círculo vicioso en el que la

DOI: http://dx.doi.org/10.22201/iibi.0187358xp.2017.72.57828

ausencia de medición impide tomar decisiones fundamentadas en torno a la contribución de su desarrollo, y si no se toman medidas sobre su comportamiento difícilmente se podrán mejorar sus condiciones, al menos de forma planificada y consciente.

Para la solución de este problema se deberán tomar decisiones orientadas en dos sentidos: el primero, concientizar a los directivos y líderes de instituciones académicas de que una de las principales formas de incidir en el desarrollo disciplinar de su campo es conociendo su comportamiento; segundo, que la mejor forma de conocer las regularidades que lo describen es midiéndolo mediante un adecuado sistema de indicadores.

Una de las dificultades que se presentan en el intento por medir el desarrollo disciplinar se relaciona con la falta de datos estructurados y normalizados para este propósito. Durante años las instituciones han acumulado grandes volúmenes de datos desarticulados en forma de series cronológicas y reportes estadísticos para resolver las necesidades de información orientadas a satisfacer los requerimientos de los niveles superiores sobre los resultados primarios de su desempeño, los cuales finalmente quedan sintetizados en los informes anuales sobre el desempeño de la institución y de sus académicos con fines evaluativos. Pero ¿qué sucede con el desarrollo de la disciplina para la cual supuestamente van orientados los esfuerzos de una institución académica? ¿En qué medida los resultados académicos de una institución y los recursos que para ello invierte se traducen en crecimiento del desarrollo de su disciplina?

Muchas veces las respuestas a estas interrogantes parecen tan obvias que la necesidad de medir la magnitud y el comportamiento de las regularidades sobre el desarrollo disciplinar de sus campos de conocimientos se diluye ante el interés por conocer la dinámica y funcionamiento de la institución, limitación sustentada en el criterio de que las instituciones con mayor producción, impacto y visibilidad de sus resultados son aquellas que mayor contribución hacen al desarrollo del campo disciplinar en el cual se enmarcan. Tal razonamiento conduce a una suerte de determinismo geográfico en el cual las principales aportaciones a las disciplinas provienen de instituciones académicas procedentes de países de economías desarrolladas, soportadas en grandes volúmenes de recursos financieros y materiales y por consiguiente altos niveles en sus resultados científicos.

Una de las formas que pudiera contribuir con la ruptura de esta inercia parte del interés por conocer cuál es el panorama objetivo de las potencialidades de investigación y docencia de las instituciones académicas, según países seleccionados, *versus* los resultados científicos que estas instituciones tienen y en qué medida las correlaciones entre estos dos aspectos indican una

tendencia hacia su desarrollo y por consiguiente hacia la disciplina a la cual pertenecen, para de esta forma aportar información objetiva que incida y mejore su tendencia hacia el desarrollo.

Lo anterior no resulta una tarea fácil, se requiere contar con un repositorio de datos de gran volumen, de instrumentos, métodos y metodologías que conformen un sistema cooperativo de información orientado hacia estos fines, desde una perspectiva histórica y que aporte información de forma oportuna para la toma de decisiones en materia de medición sobre el desarrollo institucional y disciplinar.

Uno de los campos de conocimiento que en las últimas décadas ha aportado las herramientas necesarias para la medición y evaluación de fenómenos tan complejos como el antes descrito es el Knowledge Discovery in Database (KDD)- Descubrimiento de conocimiento en bases de datos. El KDD se define como una metodología que agrupa procesos no comunes de identificación de patrones válidos, novedosos, potencialmente útiles y finalmente comprensibles en los datos (Fayyad, Piatetsky-Shapiro y Smyth, 1996; Han y Kamber, 2001). Esta metodología tiene entre sus etapas esenciales a la minería de datos, entendida como el proceso de descubrimiento de conocimiento sobre repositorios de datos complejos mediante la extracción oculta y potencialmente útil en forma de patrones globales y relaciones estructurales implícitas entre los datos (Kopanakis y Theodoulidis, 2003; INFOVIS, 2006).

Como se puede apreciar, la metodología del KDD en general y el proceso de minería de datos en particular requieren contar como etapa previa del uso de datos estructurados y orientados con estos propósitos. De ahí que uno de los primeros pasos de esta metodología incluye el diseño de un *data warehouse* o repositorio de datos, el cual tiene como finalidad almacenar los datos provenientes de una o más bases de datos y fuentes operativas, proceso mediante el cual se requiere "limpiar" (eliminar los datos que no serán usados como soporte de decisiones), normalizar y desheredar los datos de las estructuras anteriores de las que proceden para integrar un nuevo ordenamiento y disposición orientados de acuerdo con la lógica del nuevo uso o análisis a los que se van a destinar.

El *data warehouse* y la minería de datos surgieron a finales de la década de los 80 como consecuencia del desarrollo y evolución de la tecnología de bases de datos (Han y Kamber, 2001: 2). Desde entonces y hasta la fecha el desarrollo teórico y el perfeccionamiento de estos dos procesos han sido continuos, así como su extensión a casi todas las esferas del mundo empresarial y de las organizaciones, incluyendo el novedoso campo de la tecnología y los dispositivos de detección de las telecomunicaciones inalámbricas (Raffaetá *et al*., 2013). En este desarrollo muchas de las etapas de diseño del

DOI: http://dx.doi.org/10.22201/iibi.0187358xp.2017.72.57828

data warehouse han recibido una atención considerable en la literatura especializada; sin embargo, la atención a la prueba de los datos contenidos en los *data warehouse* ha sido poco tratada, aunque este problema ya empieza a aparecer en la literatura como parte del perfeccionamiento y atención que se le presta a este campo (Golfarelli y Rizzi, 2013). Este vertiginoso desarrollo que presentado el diseño del *data warehouse* muestra dos etapas: se habla de una primera generación en la que por una variedad de razones los metadatos no fueron considerados una parte significante en su generación, y una segunda generación, conocida como DW 2.0, en la que se revalora, entre otras cosas, el rol de los metadatos (Inmon, Strauss y Neushloss, 2008)

El diseño del *data warehouse* se ha convertido en el punto crucial de la metodología del KDD y del proceso de minería de datos debido a que su diseño, característica, arquitectura, estructura y operaciones están determinados por el fin último de los procesos de análisis y medición que se pretendan realizar con los datos que lo integran. En la actualidad esta tecnología se ha convertido en un importante sistema de soporte para la toma de decisiones de grandes empresas, consorcios y de instituciones científicas y universidades porque optimiza y orienta los datos de las organizaciones para su mejor explotación; es precisamente su orientación como soporte de la toma de decisiones la que lo diferencia de los sistemas de bases de datos operacionales.

El avance de un proyecto de investigación encaminado a medir el comportamiento métrico del desarrollo disciplinar en instituciones académicas iberoamericanas en ciencias bibliotecológica y de la información motivó el interés por el diseño e implementación de un *data warehouse* con estos fines. El presente artículo tiene como propósito presentar una síntesis gráfica de los resultados obtenidos en este proceso de diseño mediante la explicación de sus características, estructura, arquitectura y operaciones de funcionamiento.

Material y método

Fuente

Las fuentes principales de información a partir de las cuales se alimenta el sistema operacional del *data warehouse* parten de la implementación de un sistema de cuestionarios encaminado a la obtención de información y de indicadores sobre las potencialidades en investigación y docencia en esta disciplina y región. El levantamiento de esta información permite elaborar un diagnóstico nacional para cada país así como otro regional, en el que se presenta una panorámica general del comportamiento de este problema en la zona.

A partir de estos cuestionarios se diseña la estructura de la base de datos Potencialidades (incluye datos de potencialidades y bibliométricos), la cual ha sido desarrollada en un sitio web para facilitar su llenado desde cualquier país participante. Más detalles al respecto pueden ser consultados en los cuestionarios que se encuentran en el enlace de este proyecto o la página http://132.248.242.212/~gorbea/observatorio.html

Unidades de análisis y observación

En la fuente de información anterior se podrán identificar tres unidades de análisis principales a partir de las cuales se identificarán las variables y los indicadores que se obtendrán para cada una de ellas:

- Instituciones
- Investigadores o docentes
- Proyectos de Investigación
- Otras unidades de observación relacionadas con el diseño de los indicadores compuestos son indicadores de potencialidades e indicadores bibliométricos de producción y comunicación científica

Las limitaciones espaciales y temporales para el ingreso de la información quedan definidas como países de la región iberoamericana en los que se imparte cualquier nivel de enseñanza y/o se realizan investigaciones científicas en ciencias bibliotecológica y de la información durante el periodo comprendido entre 2007 y 2013.

A partir de estas definiciones se obtiene un sistema de medición en el que se integran más de 40 indicadores de potencialidades y bibliométricos a partir de los cuales el data warehouse genera en forma resumida dos matrices de indicadores desglosados por año para cada institución y país, cuyas estructuras pueden ser consultadas en un artículo anterior publicado sobre este sistema (Gorbea-Portal y Piña-Pozas, 2013: 159).

Mediante la correlación de los indicadores integrados en ambas matrices se generan índices e indicadores compuestos a partir de las consultas realizadas al data warehouse y mediante el empleo de la estadística multivariada: análisis de correspondencia, escalamiento multidimensional, análisis factorial y análisis de componentes principales, además de otros modelos como los números índices y los indicadores compuestos.

Los resultados obtenidos con esta metodología permiten medir el comportamiento métrico del desarrollo disciplinar en la temática y región de referencia por instituciones y países, además de establecer tendencias que

DOI: http://dx.doi.org/10.22201/iibi.0187358xp.2017.72.57828

permitan aportar información para la toma de decisiones en materia de investigación, docencia y desarrollo disciplinar.

Análisis y discusión de los resultados

La literatura especializada sobre data warehouse es abundante en la reseña de los aspectos teóricos que sustentan todo este novedoso campo de conocimiento y por consiguiente en definiciones que describen el concepto de data warehouse. Inmon, por ejemplo, la define como "una colección de bases de datos integradas, de carácter temático diseñadas para apoyar la función del sistema de soporte de decisiones, donde cada unidad de datos es relevante en algún momento en el tiempo" (2005: 29. Traducción propia).

Kimball proporciona una definición más simple al señalar que un data warehouse "es una copia de la data transaccional (también llamada información de las transacciones) específicamente estructurada para realizar consultas y análisis" (1996: 310. Traducción propia). La data transaccional son los datos que describen un evento, es decir, el cambio como resultado de una transacción. La información transaccional siempre tiene una dimensión de tiempo, y un valor numérico, y se refiere a uno o más objetos, llamados datos de referencia. Éstos se pueden resumir en un data warehouse para ayudar a la accesibilidad y el análisis de los datos. Sin embargo, ambas definiciones se complementan y aportan mayor conocimiento sobre su comprensión.

Las etapas de diseño del *data warehouse* que aquí se presentan toman como base las definiciones anteriores y la estructura propuesta por el propio Inmon, quien señala que hay dos principales aspectos en su construcción: "el diseño de la interfaz procedentes de sistemas operativos y el diseño del propio *data warehouse*" (2005: 71. Traducción propia).

Por lo anterior, en la explicación de la estructura que aquí se presenta se distinguen las dos partes: la operacional de la base de datos relacional de la cual provienen los datos y las estructuras simplificadas que conforman el *data warehouse*.

Características del data warehouse

Entre las características distintivas de los *data warehouse* se encuentra la *orientación al tema*, la *integración*, la *no volatilidad* de los datos y su denominación de *tiempo variante*, esto es, que facilitan el uso de datos históricos para el análisis de tendencia debido a que el horizonte de tiempo es significativamente más largo que en el sistema operacional (Inmon, 2005). Asimismo,

la información se clasifica con base en los temas que interesan a la organización, lo que significa que los datos no se encuentran en función de las aplicaciones, sino que se orientan a los sujetos. El diseño de este *data warehouse* tiene como propósito la obtención de indicadores bibliométricos y de potencialidades, así como la conformación de un indicador compuesto que sintetice la correlación entre ambos, con la finalidad de medir el desarrollo disciplinar en instituciones académicas, de ahí que los sujetos a los que se orienta son directores de escuelas, facultades, institutos y centros de investigación que participan en el proyecto, quienes podrán tener acceso en la entrada y actualización de la información del sistema operacional, así como la consulta al *data warehouse* en la obtención de información para la toma de decisiones que permitan orientar el desarrollo de sus instituciones académicas y de su propia disciplina.

La integración de los datos en este diseño se manifiesta a partir de la normalización de las variables procedentes de la estructura operativa, además de la unificación y síntesis de las medidas o magnitudes en las que éstas se presentan. La integración de los datos elimina la falta de normalización que tienen los datos en el sistema de base de datos relacionales y contribuye a la consistencia del sistema. Con el resultado de este proceso de integración se generan los catálogos, directorios y otras informaciones sobre los datos que permiten la integración de los metadatos, parte esencial de los *data warehouse*, sobre todos los de la segunda generación (DW 2.0).

Los datos en el sistema operativo se pueden actualizar constantemente y permiten el cálculo de indicadores de potencialidades y bibliométricos en tiempo real en el momento de la consulta, mientras que en el *data warehouse* la información que se obtiene es de carácter histórico y acumulativo; refleja una fotografía del momento de la última actualización masiva que recibió del sistema operacional por lo que su consulta aporta información para el análisis de tendencias de desarrollo en las distintas instituciones que participan en el proyecto.

Otra característica del data warehouse se refiere a la no volatilidad de sus datos, su única actualización proviene de la base de datos operativa con la última información que ingresó al sistema. Para efectos de este diseño la actualización del data warehouse se prevé una vez cada dos años, tiempo durante el cual se realizan las encuestas y se actualizan los sistemas de cuestionarios y de indicadores. Estas actualizaciones no sustituyen sus datos sino que se acumulan con los que ya existen cargados de años anteriores.

DOI: http://dx.doi.org/10.22201/iibi.0187358xp.2017.72.57828

Estructura del data warehouse

La estructuración del *data warehouse* se realiza a partir de diferentes niveles de detalles. En su estructura se identifican los *detalles de los datos actuales y antiguos (históricos)*, los *datos ligeramente resumidos* y los *altamente resumidos.* La estructuración de estos niveles parte de la que se encuentra en la estructura operacional y se asocia con el nivel de granularidad que alcanzan los datos, por ejemplo, los datos actuales son más voluminosos y menos granulados mientras que los datos históricos son más integrados y sintéticos. El aspecto más importante de un *data warehouse* es el tema de la granularidad, ésta se refiere al nivel de detalle o de resumen que tienen unidades de datos en el *data warehouse* (Inmon, 2005).

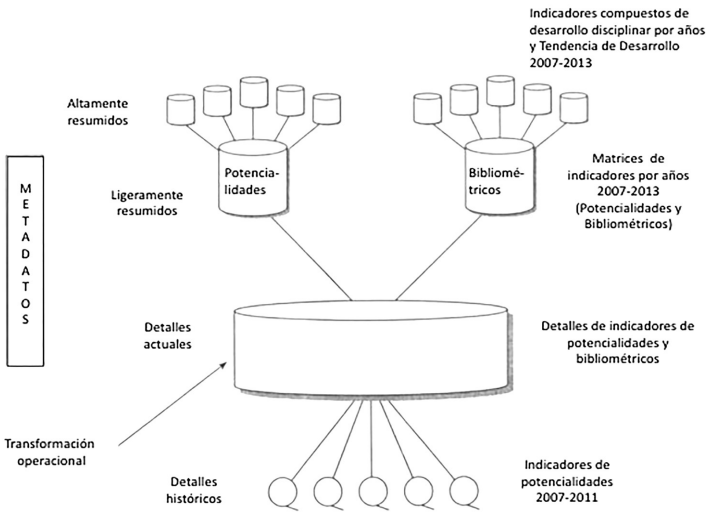

Figura I. Estructura de data warehouse para medir el desarrollo disciplinar en instituciones académicas a partir del esquema presentado por Inmon (2005)

Además de la especificación de estos niveles de detalles, en la estructura se incluyen los metadatos, en este caso se refiere a datos acerca de los datos, la descripción de las estructuras, contenidos, llaves, índices, catálogos, directorios, etc. Es decir, toda la información que le dé contenido a los datos y que en su análisis ayude a su interpretación y su conocimiento. Para la representación de la estructura de este diseño se presenta en la *Figura 1* un esquema adaptado a partir del presentado por Inmon (2005: 34).

Estructura operacional

Como parte de la estructura general del *data warehouse* (*Figura 1*), a continuación se presenta la estructura general de la base de datos relacional o sistema operacional, así como las estructuras detalladas con sus atributos de los tres elementos que la integran. Esta etapa del diseño operacional es sumamente importante en la construcción del *data warehouse*, debido a que la robustez de ésta depende de cómo hayan sido concebido los datos almacenados en el sistema operacional, de ahí la importancia que le conceden a esta etapa autores como Inmon (2005). Dicho de otra forma, no se puede lograr un buen diseño del *data warehouse* si no se domina a la perfección el diseño de bases de datos convencionales o relacionales.

En la *Figura 2* se muestran de forma general las relaciones que subyacen entre los tres elementos que integran una base de datos operacional. Los datos sobre las instituciones, los recursos humanos y los proyectos de investigación comparten tablas comunes y se relacionan entre sí a partir de las llaves de conexión, identificadas al inicio del nombre del atributo con las letras ID.

En las *Figuras 3*, *4* y *5* se detallan los atributos de cada una de las relaciones que se dan al interior de los tres conjuntos de datos, referidos al desarrollo institucional, los recursos humanos y los proyectos de investigación, respectivamente.

En los esquemas mostrados en las *Figuras 3, 4* y *5* se pueden apreciar las tablas de la base de datos relacional, formadas por campos. Cada tabla está dividida en dos secciones, la sección superior contiene los campos que forman parte de la llave primaria de la tabla. Las interrelaciones entre dos o más tablas se llevan a cabo a través de las llaves primarias y son llamadas llaves foráneas (FK, *foreign key*). Una llave foránea puede formar parte de la llave primaria de la tabla dependiente, es decir, forma parte del conjunto de campos que sirven para identificar de manera única a los registros de la tabla, o ser simplemente un atributo más de la tabla dependiente.

Estructura del data warehouse

De las estructuras anteriores se obtienen nuevas tablas y campos (variables), estas últimas se relacionan entre sí para conformar un indicador. En este sentido se presenta, a modo de ejemplo, la relación de uno de los indicadores de los más de 40 que se obtiene en este diseño, el identificado como producción científica-institución-país y representado en la tabla denominada *instPC* de la *Figura 6.*

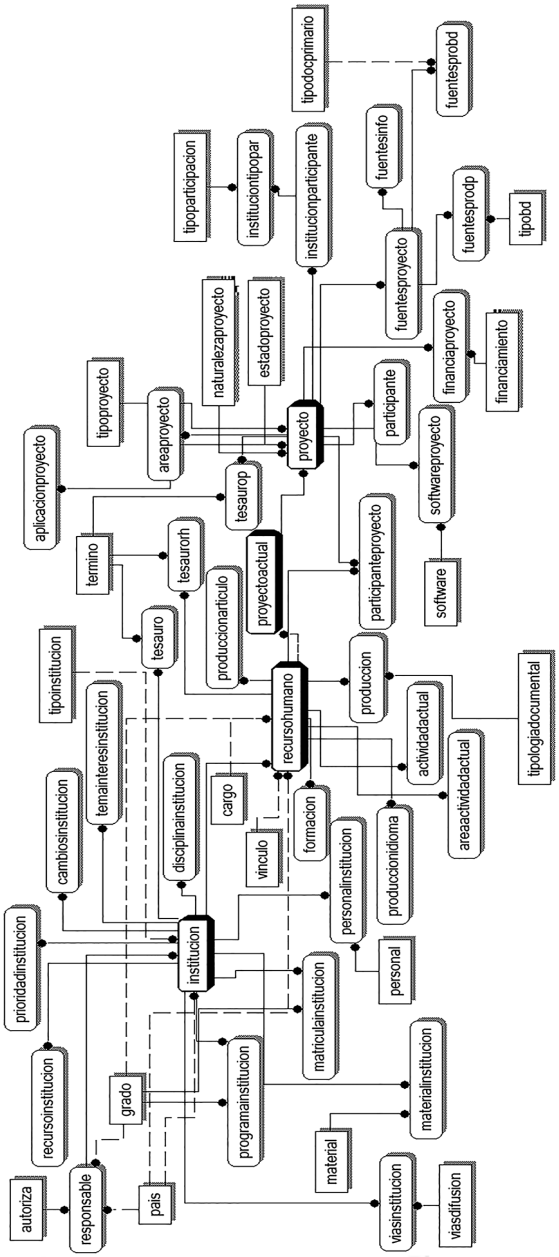

Figura 2. Estructura general simplificada de la base de datos operacional

DOI: http://dx.doi.org/10.22201/iibi.0187358xp.2017.72.57828

INVESTIGACIÓN BIBLIOTECOLÓGICA, Vol. 31, Núm. 72, mayo/agosto, 2017, México, ISSN: 2448-8321, pp. 161-181

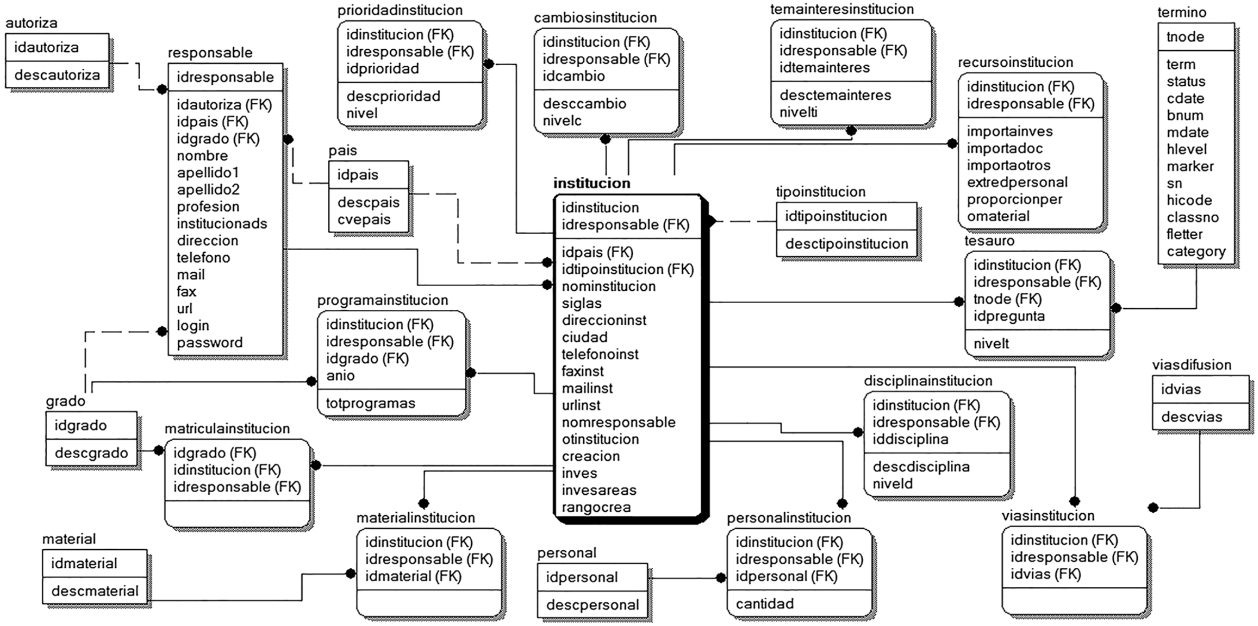

Figura 3. Estructura particular extendida con los atributos del segmento correspondienteal desarrollo institucional

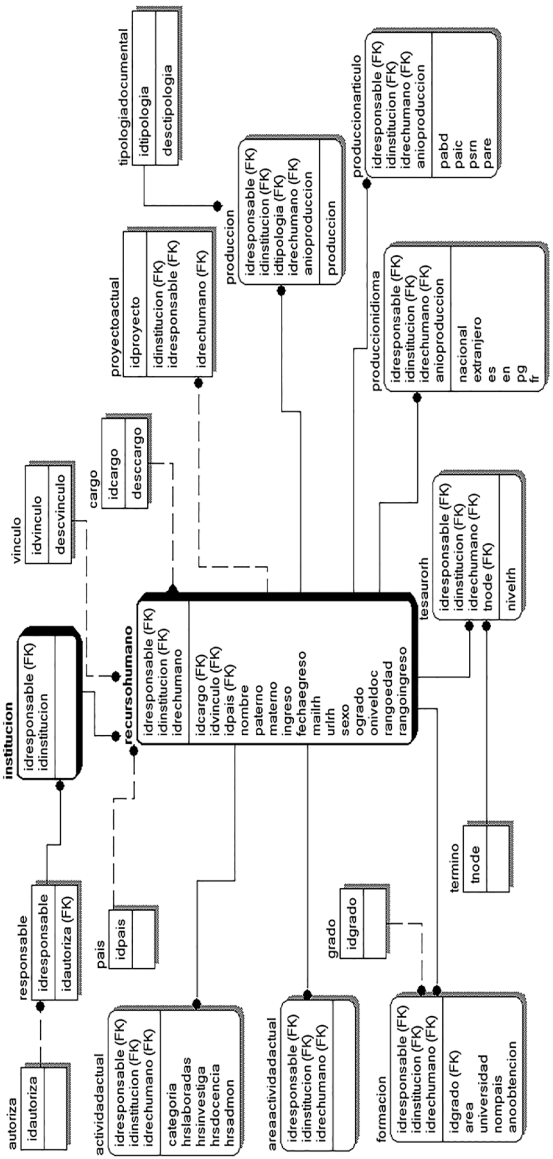

Figura 4. Estructura particular extendida con los atributos del segmento correspondiente a los recursos humanos

DOI: http://dx.doi.org/10.22201/iibi.0187358xp.2017.72.57828

INVESTIGACIÓN BIBLIOTECOLÓGICA, Vol. 31, Núm. 72, mayo/agosto, 2017, México, ISSN: 2448-8321, pp. 161-181

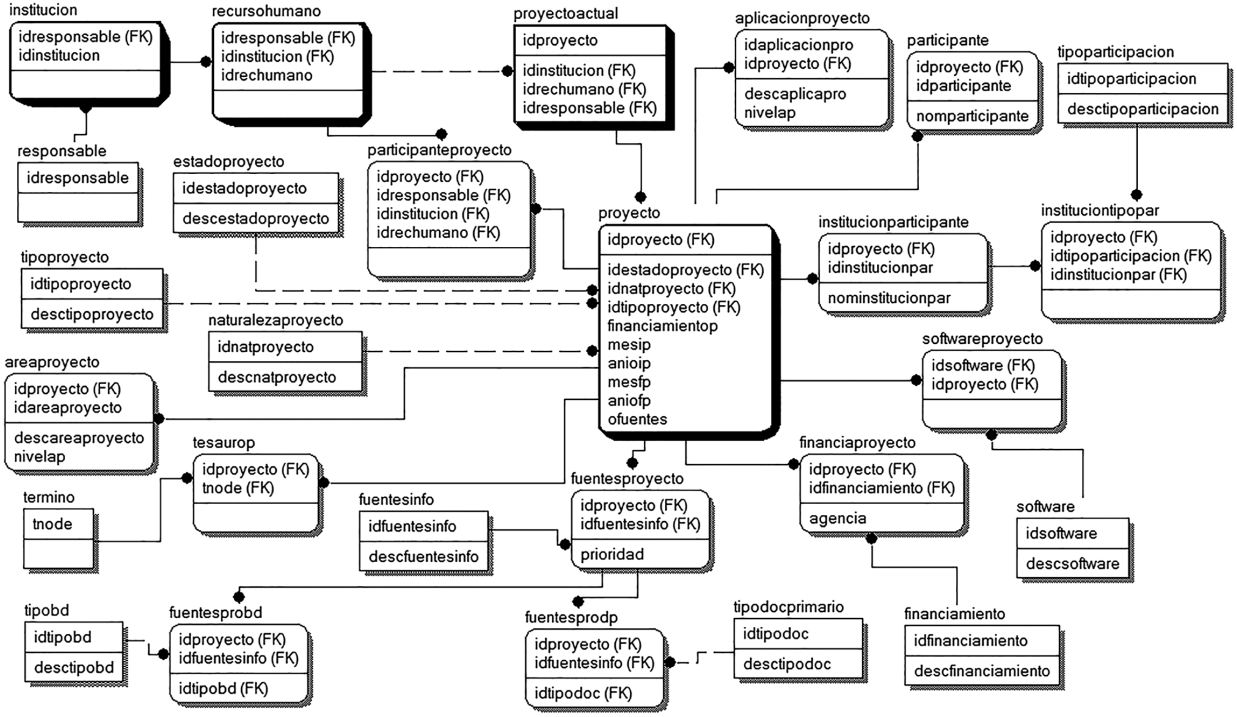

Figura 5. Estructura particular extendida con los atributos del segmento correspondiente a los proyectos de investigación

DOI: http://dx.doi.org/10.22201/iibi.0187358xp.2017.72.57828

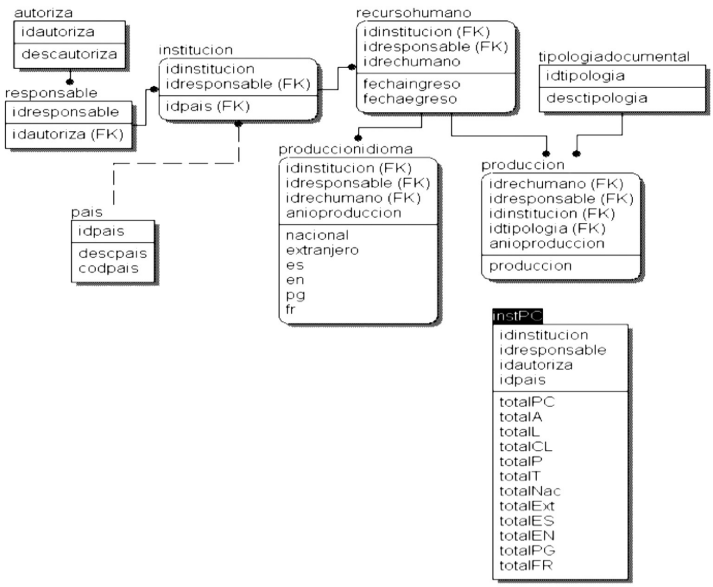

Figura 6. Relación depurada para el cálculo del indicador producción científica-institución-país

En la relación anterior las tablas y campos quedan indicados de acuerdo con lo explicado en las relaciones generales de la base de datos operacional. En la *Tabla 1* del *Anexo* se muestran las definiciones de tablas y campo usadas para la integración y síntesis.

Arquitectura del data warehouse

En la arquitectura se representan los diferentes niveles de acceso a los datos y a la información de toda la estructura: *nivel de base de datos operacional* y de base de datos externo, esta última como complemento de los datos operacionales; *nivel de acceso a la información*, a partir del cual el usuario final dispone de la conversión de los datos en información, incluye el conjunto de herramientas de que dispone el sistema para la visualización de la información gráfica, de reportes, diagramas, etc.; *nivel de acceso a los datos*, se encuentra muy relacionado con el nivel de acceso a la información y funciona como intermediario entre el sistema operacional y el nivel de acceso a la información; *nivel de directorio de metadatos*, el usuario mantiene acceso a los directorios y a la información sobre los datos; *nivel de gestión de procesos*, constituye el conjunto de procesos, programas, rutinas y aplicaciones responsables de mantener actualizado el *data warehouse*; *nivel de mensaje de la aplicación,* es el encargado de transportar

la información por toda la red conectada al sistema; *nivel data warehouse*, en este nivel se procesan los datos actuales de forma dinámica para usos estratégicos del sistema; *nivel de organización de los datos*, representa el último nivel en la arquitectura y resume el conjunto de procesos necesarios, el acceso a la información desde bases de datos operacionales o externas, procesos de selección, editar, resumir, cargar datos en el depósito.

Operaciones del data warehouse

Entre las operaciones que comúnmente se desarrollan en un *data warehouse* se representan las siguientes: *los sistemas operacionales*, constituyen la fuente principal de datos, se estructuran por lo general en bases de datos relacionales; *extracción, transformación y carga* de datos, operaciones que permiten extraer y manipular los datos desde diversos sistemas operacionales para luego limpiarlos, eliminar inconsistencia, transformarlos y cargarlos en el *data warehouse*; *creación de los metadatos,* que constituyen los datos sobre los datos, cuenta con los directorios, catálogos de fácil consulta por los usuarios finales; *acceso de usuario final*, garantiza el acceso del usuario final mediante interfaces gráficas que le permiten generar reportes, hacer consultas, generar reportes mediante el OLAP (procesamiento analítico en línea) de acuerdo con los requerimientos del usuario; *plataforma del data warehouse*, plataforma donde reside el *data warehouse* y que por lo regular es un servidor de base de datos relacionales, de acuerdo con el volumen puede llegar a requerir una batería de servidores; *datos externos*, no siempre son requeridos, depende de la magnitud y volumen de los datos, sirven de complemento a los datos provenientes del sistema operacional.

En la *Figura 7* se muestra el conjunto de operaciones que se integran en el diseño del *data warehouse*, un servidor en el que se hospeda la base de datos relacional con las estructuras antes descritas gestionada con *Postgresql.* En ella se almacenan los datos proveniente del sistema de cuestionarios de instituciones, recursos humanos y proyectos de investigación, alimentados por la red de instituciones de los países participantes en el proyecto. Mediante el procedimiento de extraer, transformar y cargar se depuran y sintetizan los datos que son incluidos en el *data warehouse* y se crean los metadatos. Finalmente se actualiza en forma masiva el *data warehouse*, residente en un servidor UNIX, para ser consultada por los usuarios finales. No está representado en estas operaciones el acceso a datos externo debido a que, por el momento, este diseño no tiene prevista la integración de datos provenientes de fuentes externas con los del sistema operacional.

DOI: http://dx.doi.org/10.22201/iibi.0187358xp.2017.72.57828

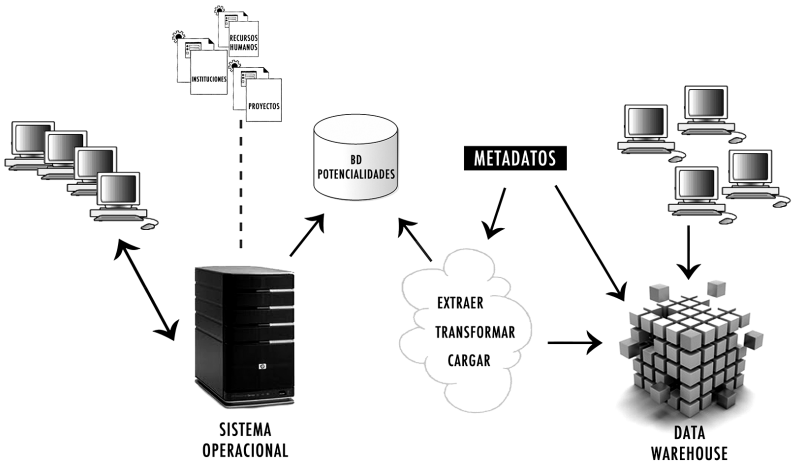

Figura 7. Operaciones del *data warehouse*

Consideraciones finales

El diseño de los *data warehouse* ha alcanzado su mayor desarrollo en las grandes empresas y transnacionales del sector industrial, bancario y financiero; su aplicación en las universidades e instituciones científicas todavía es escasa, sin embargo, ya empieza a aplicarse en estas esferas con éxito.

La experiencia de diseño que aquí se presenta demuestra la utilidad de estas estructuras en el cálculo científico en general y en particular en la metría de comportamientos complejos como éste, orientado a medir el desarrollo disciplinar de instituciones académicas, en el cual la lógica y estructura de las bases de datos y sistemas operacionales no satisfacen la medición de este tipo de problema.

Esta experiencia de aplicación de los *data warehouse* en la metría de la información y del conocimiento científico, específicamente en la metría del desarrollo institucional y disciplinar en las ciencias bibliotecológica y de la información, unido a los esfuerzos encaminados al diseño de los *data warehouse* en las bibliotecas para la aportación de información en la toma de decisiones mediante la bibliominería de datos, alerta sobre la apertura de un campo de conocimiento transdisciplinario de insospechadas dimensiones.

Referencias

Fayyad, U. G., G. Piatetsky-Shapiro y P. Smyth. 1996. "From data mining to knowledge discovery in databases". *AI Magazine* 17 (3): 37-54.

Gorbea-Portal, S. y M. M. Piña-Pozas. 2013. "Propuesta de un indicador para medir el comportamiento del desarrollo disciplinar de las Ciencias Bibliotecológica y de la Información en instituciones académicas". *Investigación Bibliotecológica* 27 (60): 153-180.

Golfarelli, M. y S. Rizzi. 2013. "Data Warehouse Testing", en *Developments in Data Extraction, Management, and Analysis*, editado por Nhung Do, J. Wenny Rahayu y Torab Torabi, 91-108. Hershey, PA: Information Science Reference.

Han, J. y M. Kamber. 2001. *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann Publishers.

INFOVIS. 2006. *Visualización de Información en Documentación. ¿Qué es la minería de datos?* http://infovis.rivarela.com/index.php?q=mieriadedatos

Inmon, W. H. 2005. *Building the Data Warehouse*. Indianapolis: Wiley Publishing Inc.

Inmon, W. H., D. Strauss y G. Neushloss. 2008. DW 2.0. *The Architecture for the Next Generation of Data Warehousing*. UK: Elsevier's Science & Technology.

Kimball, R. 1996. *The data warehouse toolkit: practical techniques for building dimensional data warehouses*. New York: John Wiley & Sons, Inc. https://www.amazon.com/Data-Warehouse-Toolkit-Techniques-Dimensional dp/0471153370#reader_0471153370

Kopanakis, I. y B. Theodoulidis. 2003. "Visual data mining modeling techniques for the visualization of mining outcomes". *Journal of Visual Languages and Computing* 14 (6): 543-589.

Raffaetá, A. et al. 2013. "Visual Mobility Analysis Using T-Warehouse", en *Developments in Data Extraction, Management, and Analysis*, editado por Nhung Do, J. Wenny Rahayu y Torab Torabi, 1-22. Hershey, PA: Information Science Reference.

Anexo

Tabla 1. Definición de las relaciones y atributos considerada para el cálculo del indicador producción científica-institución-país

Relación/Atributos	Descripción
autoriza	En la relación autoriza se almacenan los diferentes tipos de acceso a la base de datos que puede tener cada participante del proyecto.
idautoriza	Atributo que identifica el nivel de acceso o rol que desempeña cada participante del proyecto (llave primaria, PK por sus siglas en inglés); es de tipo numérico secuencial asignado automáticamente por el sistema. (PK)

DOI: http://dx.doi.org/10.22201/iibi.0187358xp.2017.72.57828

descautoriza	Este atributo describe el nivel de autorización a la base de datos, los niveles pueden ser responsable regional, coordinador nacional y responsable institucional.
responsable	En esta relación se almacenan los datos de las personas participantes en el proyecto, así como el rol que desempeñan en él. A continuación se definen solamente los atributos que son de interés para la agrupación de datos que aquí se muestra.
idresponsable	Atributo que identifica de forma única a cada registro, es de tipo numérico secuencial asignado automáticamente por el sistema. (PK)
idautoriza	Este atributo es una llave foránea (FK), ver definición en relación *autoriza*. (FK)
Pais	En la relación *pais* se almacenan el catálogo de autoridad de los países.
idpais	Atributo que identifica de forma única a los países, es de tipo numérico secuencial asignado automáticamente por el sistema. (PK)
descpais	Nombre de los países.
codpais	Código de dos letras para identificar a los países.
institucion	En esta relación se almacenan los datos generales de cada una de las instituciones participantes en el proyecto, a continuación se mencionan los atributos necesarios para crear la agrupación de datos utilizada para obtener la producción científica.
idinstitucion	Ver definición en relación *institucion*. (PK)
idresponsable	Ver definición en relación *responsable*. (PK)
idpais	Ver definición en relación *pais*. (PK)
recursohumano	En esta relación se almacenan los datos generales de los recursos humanos de cada una de las instituciones participantes en el proyecto. A continuación se definen solamente los atributos que son de interés para la agrupación de datos que aquí se muestra.
idinstitucion	Ver definición en relación *institucion*. (PK)
idresponsable	Ver definición en relación *responsable*. (PK)
idrechumano	Ver definición en relación *recursohumano*. (PK)
fechaingreso	Fecha de ingreso a la institución.
fechaegreso	Fecha en la que el recurso humano terminó su relación o vínculo con la institución.
tipologiadocumental	En esta relación se almacenan los datos de la tipología documental tomada en cuenta en el cuestionario de recursos humanos.

idtipologia	Atributo que identifica de forma única a cada tipo de documento contabilizado en el cuestionario de recursos humanos, es de tipo numérico secuencial asignado automáticamente por el sistema. (PK)
desctipologia	Se almacena la descripción de la tipología documental, los valores son artículo, libro, capítulo de libro, ponencias y tesis.
produccionidioma	La relación *produccionidioma* se usa para almacenar los datos relacionados con el idioma de los artículos, así como el tipo de publicación de la producción científica en general (publicación nacional o extranjera).
idinstitucion	Ver definición en relación *institucion*. (PK)
idresponsable	Ver definición en relación *responsable*. (PK)
idrechumano	Ver definición en relación *recursohumano*. (PK)
anioproduccion	En el atributo *anioproduccion* se almacena el valor correspondiente al año de la producción científica de la tipología documental producida por cada uno de los recuros humanos de cada institución. (PK)
Nacional, extranjero	En los atributos *nacional* y *extranjero* se almacena el total de la producción científica publicada en medios nacionales y extranjeros por cada recurso humano.
es, en, pg, fr	Estos atributos son usados para almacenar el total de artículos publicados en español, inglés, portugués y francés de los recursos humanos de cada institución.
produccion	La relación *produccion* es utilizada para almacenar los datos relacionados con la producción de artículos, libros, capítulos de libros y ponencias publicadas por los recursos humanos de las instituciones, así como las tesis dirigidas.
idrechumano	Ver definición en relación *recursohumano*. (PK)
idinstitucion	Ver definición en relación *institución*. (PK)
idresponsable	Ver definición en relación *responsable*. (PK)
idtipologia	Ver definición en relación *tipologiadocumental*. (PK)
anioproduccion	En el atributo *anioproduccion* se almacena el valor correspondiente al año de la producción científica de la tipología documental producida por cada uno de los recursos humanos de cada institución. (PK)
produccion	En este atributo se almacena el total de artículos, libros, capítulos de libros, ponencias publicadas y/o tesis dirigidas por cada recurso humano por institución.

DOI: http://dx.doi.org/10.22201/iibi.0187358xp.2017.72.57828

Producción científica agrupada por institución	
instPC	En la relación *instPC* es donde se almacena la agrupación de la producción científica total y por artículo de cada institución.
idinstitucion	Ver definición en la tabla *institucion*. (PK)
idresponsable	Ver definición en la tabla *responsable*. (PK)
idauotriza	Ver definición en la tabla *autoriza*. (PK)
idpais	Ver definición en la tabla *pais*. (PK)
totalPC	En este atributo se almacena el total de la producción científica de cada institución.
totalA	En el atributo *totalA* se almacena de total de artículos publicados por los recursos humanos de cada institución, es decir, se suma la producción de artículos de cada recurso humano para así obtener el total por institución.
totalL, totalCL, totalP	Estos tres atributos están relacionados con la producción de libros, capítulos de libros, y ponencias publicadas por los recursos humanos de cada institución, el cálculo se hace en forma similar al total de artículos publicados.
totalT	El atributo *totalT* se utiliza para almacenar el valor del total de tesis dirigidas por los recursos humanos de cada institución.
totalNac, totalExt	En los atributos *totalNac*, *totalExt*, se almacena el total de la producción científica publicada en medios nacionales y extranjeros respectivamente.
totalES, totalEN, totalPG, totalFR	Estos atributos son usados para almacenar el total de artículos publicados en español, inglés, portugués y francés por institución.

Nota: la falta de acentos en los nombres de los atributos en esta tabla y en las figuras precedentes que representan las relaciones del sistema se debe a que se ha querido respetar la forma original como los aporta el sistema y su lenguaje de programación, en donde no se admiten acentos de ningún idioma.

Para citar este texto:

Gorbea-Portal, Salvador y María de Jesús Madera-Jaramillo. 2017. "Diseño de un data warehouse para medir el desarrollo disciplinar en instituciones académicas". *Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información* 72 (31): 161-181.
http://dx.doi.org/10.22201/iibi.0187358xp.2017.72.57828