UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
POSGRADO EN BIBLIOTECOLOGÍA Y ESTUDIOS DE LA INFORMACIÓN
FACULTAD DE FILOSOFÍA Y LETRAS
INSTITUTO DE INVESTIGACIONES BIBLIOTECOLÓGICAS Y DE LA INFORMACIÓN

BIG DATA PARA BIBLIOTECAS ACADÉMICAS

TESIS
QUE PARA OPTAR POR EL GRADO DE:
MAESTRA EN BIBLIOTECOLOGÍA Y ESTUDIOS DE LA INFORMACIÓN

PRESENTA:
BRENDA ISABEL REYES PAEZ

TUTOR DR. JUAN VOUTSSÁS MÁRQUEZ
INSTITUTO DE INVESTIGACIONES BIBLIOTECOLÓGICAS Y DE LA INFORMACIÓN

Ciudad Universitaria, CD. MX., FEBRERO, 2025.

Contenido

Índice de ilustraciones / figuras

Índice de tablas

Introducción

Desde el surgimiento de las primeras formas de escritura hasta los centros de información actuales, los humanos no han dejado de recopilar información. El nacimiento de la Bibliotecología formó parte de ese proceso: la necesidad de recabar, resguardar y ordenar información parte de esos principios.

Con las metodologías de investigación, que estandarizaron las formas de estudiar fenómenos desde diversas perspectivas y el crecimiento del sector tecnológico, se ha provocado tal aumento en la producción de datos, que son necesarios sistemas de almacenamiento, de organización y de análisis más sofisticados. Se requieren nuevos conceptos, modelos y teorías, aplicados a problemáticas actuales y que su alcance sea el de antaño: organizar y resguardar información para la posterior consulta.

En la era de la información, el volumen de datos generados y almacenados ha crecido exponencialmente: Statista (IFLA, 2024) reveló que hay alrededor de 64.2 zettabytes de datos creados, capturados, copiados y consumidos en todo el universo digital, y se prevé que esta cifra se triplique en 2025.

El fenómeno anterior ha dado lugar al concepto de *Big data*. ‘Datos masivos’, ‘macrodatos’, ‘inteligencia de datos’, o ‘datos a gran escala’ podrían ser algunas traducciones al español de dicho término, el cual se refiere a conjuntos de datos tan vastos y complejos que requieren tecnologías y métodos avanzados para su procesamiento y análisis.

La constante generación de datos de diversa tipología y mediante plataformas variadas suponen un reto para su procesamiento y análisis y, en concreto, plantean diversos retos para la actualización de las prácticas bibliotecarias. Esto sucede en la inserción de metodologías de *minería de datos* y *Big data* en grandes sistemas bibliotecarios universitarios y académicos.

Las bibliotecas académicas, como custodios y facilitadores del acceso a la información, no deberían ser ajenas a la revolución de los datos. De hecho, el aprovechamiento del Big data ofrecería a estas instituciones una oportunidad sin precedentes para mejorar sus servicios, optimizar la gestión de sus recursos y contribuir de manera más efectiva el camino hacia el conocimiento.

Las bibliotecas académicas juegan un papel crucial en el apoyo a la educación superior y la investigación, al proporcionar acceso a una vasta cantidad de recursos informativos. Sin embargo, este tipo de bibliotecas enfrentan desafíos significativos en la gestión y optimización del uso de sus recursos debido a la creciente cantidad de datos generados por los usuarios y por las actividades bibliotecarias. La falta de herramientas y métodos eficientes para analizar y aprovechar estos datos puede resultar en una subutilización de recursos, ineficiencias operativas y una menor satisfacción del usuario.

En un mundo donde la información se ha convertido en un recurso esencial, las bibliotecas académicas tienen la oportunidad de liderar la innovación mediante el uso de Big data. Por ello, lo que se busca con esta investigación es proporcionar un marco teórico y práctico que guíe a estas instituciones en su camino hacia la transformación digital, con lo que se asegura que continúen siendo pilares fundamentales en la generación y difusión del conocimiento.
El problema principal que se aborda en esta investigación es ¿cómo puede el análisis de Big data mejorar la gestión y el uso de los recursos en las bibliotecas académicas al optimizar su operación y mejorar la experiencia del usuario?

Las bibliotecas académicas adquieren una gran cantidad de recursos que son puestos a disposición a través del Catálogo Público en Línea (OPAC, por sus siglas en inglés, Online Public Access Catalogue). Sin embargo, con la manera tradicional es complicado evaluar qué tanto corresponden los resultados de búsqueda que proporciona el OPAC a las necesidades de los usuarios y en qué

medida estas son satisfechas. Entonces, ¿pueden las metodologías de análisis de datos brindar mayor precisión sobre qué se consulta y cómo se consulta?

Para resolver este problema, se plantean estas preguntas de investigación: ¿qué son los datos masivos? ¿Qué es la minería de datos? ¿Cómo son las metodologías de datos masivos? ¿Es posible extraer y analizar los datos recopilados por un software de biblioteca, con las metodologías de datos masivos? ¿Es posible optimizar los recursos y servicios de las bibliotecas académicas utilizando metodologías de Big data aplicados al OPAC? ¿Es un problema de Big data o de minería de datos? ¿Cómo aplicar metodologías de análisis de datos para identificar patrones generados en el OPAC? ¿Es posible implementar cambios en la catalogación y en las respuestas del OPAC a partir de los resultados?

La información derivada de las actividades de los usuarios permitiría a los bibliotecarios obtener, en tiempo real, una enorme cantidad de datos que en sí mismos son difíciles de procesar. Pero que de hacerlo e innovar en las metodologías y herramientas de procesamiento análisis —tal y como ocurrió a mediados del siglo pasado—, podrían arrojar resultados favorables para establecer patrones, generar indicadores, conocer a los usuarios y a la comunidad a la que la biblioteca pertenece, así como dar soluciones inteligentes en las que se pueda garantizar un resultado favorable. Es decir, se crearía un ecosistema producto de los análisis de datos.

Por todo lo anterior, el objetivo del presente trabajo es diseñar una metodología con la que se lleve a cabo una extracción de datos provenientes de un catálogo para analizarlos y, a través de ello, realizar una propuesta de mejora para el OPAC.

Los objetivos específicos son los siguientes:

- Definir el concepto de Big data y minería de datos, así como explicar su aplicación en las bibliotecas.

- Identificar las características del análisis de datos para la elaboración de una metodología con el fin de hacer una extracción de datos provenientes del OPAC.

- Identificar cuáles son los datos que se recopilan en los softwares bibliotecarios a fin de procesarlos mediante las metodologías de sistemas expertos y datos masivos.

- Verificar la viabilidad de aplicación de las metodologías de Big data como estrategias viables para conocer mejor la biblioteca y mejorar los servicios.

Respecto a la metodología que se usa en esta investigación, se eligió la de naturaleza analítico-sintética, centrada en el análisis de la literatura especializada en metodologías de análisis de datos aplicadas a bibliotecas académicas. A continuación, se describe detalladamente cada uno de los componentes de esta metodología.

1. **Fase analítica:**

- **Revisión de la literatura:** consiste en una exhaustiva revisión de los textos existentes sobre metodologías de análisis de datos en el contexto de bibliotecas académicas. Se incluirán fuentes académicas como artículos de revistas, libros, tesis, conferencias y reportes técnicos.
- **Selección de fuentes:** se utilizarán bases de datos académicas reconocidas como Scopus, IEEE Xplore, Google Scholar, entre otras, para identificar y seleccionar las fuentes más relevantes y actualizadas. Los criterios de inclusión estarán basados en la relevancia, el impacto académico y la aplicabilidad práctica de los estudios.

- **Análisis crítico:** cada fuente seleccionada será analizada críticamente para identificar tanto las metodologías de análisis de datos propuestas, las técnicas utilizadas, los resultados obtenidos como las limitaciones encontradas. Además, se prestará especial atención a cómo estas metodologías pueden aplicarse o adaptarse al análisis de datos del OPAC en bibliotecas académicas.

2. **Fase sintética:**

- **Síntesis de información:** a partir del análisis crítico de la literatura, se sintetizarán los hallazgos clave y se agruparán las metodologías en categorías, según su enfoque —minería de datos, aprendizaje automático, análisis predictivo, etcétera—. Esta síntesis permitirá construir un panorama comprensivo de las metodologías disponibles y su potencial aplicación en el contexto específico de esta investigación.
- **Desarrollo de un marco conceptual:** con base en la síntesis de la literatura, se desarrollará un marco conceptual que integrará las metodologías de análisis de datos más relevantes y efectivas. Este marco servirá como guía para la aplicación práctica del análisis de datos del OPAC, con lo que se considerarán las particularidades y necesidades de las bibliotecas académicas.
- **Propuesta de metodología aplicada:** finalmente, se propondrá una metodología específica para la recuperación y el análisis de datos del OPAC. Se detallarán los pasos a seguir, las técnicas a emplear y las herramientas recomendadas. Esta propuesta estará fundamentada en la evidencia recopilada y sintetizada durante la investigación.

La implementación de una metodología de recuperación y análisis de datos del Catálogo Público en Línea (OPAC) permite identificar patrones en el comportamiento de la comunidad de usuarios, lo que contribuye a una mejora significativa en los servicios ofrecidos por el catálogo de la biblioteca académica.

Ahora bien, este trabajo se divide en tres capítulos: el primero abordará los conceptos de biblioteca académica, automatización, Sistemas Integrales de Automatización de Bibliotecas, organización documental y los OPAC.

Luego, el capítulo dos desarrolla el marco conceptual, teórico y técnico de datos masivos, minería de datos, inteligencia artificial y sistemas expertos.

Finalmente, el capítulo tres presentará una propuesta teórica sobre cómo tendría que ser la metodología para la recolección de los datos y la aplicación de alguna de las dos metodologías de análisis de datos a una biblioteca académica. No obstante, también se tratará de ejemplificar algunos pasos de manera práctica al final de este trabajo.

Algunos de los alcances de este trabajo son:

1. Se mencionará la posibilidad de aplicarse en bibliotecas académicas, pero en realidad podría ser aplicado en cualquier tipo de biblioteca que tenga un Sistema Integral de Automatización de Bibliotecas (SIAB) y un OPAC.
2. Se utilizarán herramientas de software libre por su facilidad de acceso.

Y algunas limitaciones:

1. La propuesta de este trabajo es teórica.
2. El diseño de recolección de datos deberá variar, según cada caso y acorde a lo que se pretenda analizar.

Capítulo 1: Conceptos: bibliotecas, automatización y OPAC

El universo de las bibliotecas es amplio y sus macroprocesos complejos.
Las bibliotecas académicas han sido tradicionalmente centros de conocimiento y recursos fundamentales para el desarrollo de la educación superior y la investigación. Con la evolución tecnológica y la digitalización de la información, las bibliotecas han adoptado sistemas automatizados para gestionar sus vastas colecciones y mejorar los servicios ofrecidos a sus usuarios. En este contexto, el Catálogo Público en Línea (OPAC) se ha convertido en una herramienta esencial que permite a los usuarios buscar, acceder y gestionar recursos bibliográficos de manera eficiente.

En este capítulo se proporciona un marco conceptual sobre los elementos clave de esta investigación: las bibliotecas académicas, la automatización de bibliotecas y el OPAC. Se analizarán las características, las funciones y la evolución de estos componentes, estableciendo la base teórica necesaria para comprender la aplicación de metodologías de análisis de datos en el OPAC y en la mejora de los servicios bibliotecarios.

1.1. Concepto de biblioteca académica

Antes de explicar el concepto del tipo de biblioteca que se analiza en este trabajo, mencionaré el enfoque con el que se conceptualiza.

George Eberhart (ALA, 2013) define la biblioteca como

> Una colección de recursos en diversos formatos que (1) están organizados por profesionales de la información y otros expertos quienes (2) proporcionan acceso físico, digital, bibliográfico o intelectual y (3) ofrecen servicios dirigidos y programas (4) cuya misión es educar, informar o entretener una variedad de audiencias (5) con el propósito de estimular el aprendizaje individual y el avance de la sociedad en su conjunto (párr. 2).[1]

[1] Las citas originales del inglés son traducción propia.

Este autor también menciona los *recursos de información organizados* y resalta que debe haber acceso a ellos; además, da la posibilidad de considerar estos recursos en distintos formatos y soportes. Debido a estas características, este concepto puede definirse como “tradicional”, debido a que suele ser el enfoque común que se menciona en la literatura especializada.

Dentro del enfoque “tradicional”, la biblioteca es una institución dedicada a la adquisición, conservación, organización y difusión de información y conocimiento en diversas formas y formatos. Asimismo, las bibliotecas han sido vistas como lugares físicos donde se almacenan libros, manuscritos y otros documentos impresos. Sin embargo, en la era digital, estas han evolucionado para incluir colecciones electrónicas y servicios en línea, con lo que se han adaptado a las necesidades cambiantes de los usuarios.

Sin embargo, existe otro enfoque desde el que puede estudiarse el concepto de biblioteca. Orera (s.f.) menciona que una biblioteca puede definirse como un sistema para la transmisión de información. Esta propuesta es destacable porque presenta a la biblioteca desde una visión holística y relacional, donde cada uno de sus macroprocesos se interconecta. Además, interpreta a la biblioteca como un ente que no solo proporciona información a los usuarios, sino que también recibe y procesa la información que estos le dan para optimizar sus servicios y recursos.

Enciso (1997) afirma que "dentro de la evolución del conocimiento y de la información, la biblioteca debe ser el proceso dinámico del fenómeno pregunta-respuesta y el análisis del contexto en el que tiene lugar" (p. 87). Esta perspectiva sugiere que la función de la biblioteca va más allá de un simple almacén de información; debe ser vista como un proceso dinámico que responde a las necesidades específicas de los usuarios en un contexto determinado.

Por otro lado, el enfoque denominado "sistémico", concibe a la biblioteca como un conjunto de elementos interrelacionados. De acuerdo con Enciso (1997), el mínimo de elementos de un sistema es dos, y cada uno de ellos se conecta con los otros directa o indirectamente. En este sentido, la biblioteca se configura como un sistema complejo donde cada componente interactúa para cumplir su misión informativa.

Garza (1984) menciona que la “biblioteca es un sistema de información porque adquiere, procesa, almacena y disemina mensajes, pero se distingue de un centro de información propiamente dicho porque ofrece conocimientos y datos a través de las obras que forman parte de su acervo documental. Esto resalta la función dual de la biblioteca como conservadora de conocimiento y proveedora de información actualizada” (p.17).

El conjunto de las definiciones mencionadas resalta elementos como *recursos de información*, *organización*, *acceso* y *servicios*, además de considerar a la biblioteca como un sistema que consume y produce datos, información y conocimiento. Esto da sentido a su función y justifica su existencia y utilidad en la sociedad actual.

Por lo mencionado anteriormente, esta investigación se propone conceptualizar la biblioteca como un sistema con macroprocesos interrelacionados que organiza recursos de información y ofrece diversos servicios y herramientas a los usuarios. Esta definición no pretende ser universal, sino que se ajusta al enfoque específico de este trabajo de investigación sobre el análisis de datos y el catálogo bibliotecario.

Existen diversos tipos de bibliotecas: escolares, públicas, universitarias, académicas, especializadas, nacionales, entre otras. Cada una cumple con distintos objetivos y, de acuerdo con estos, planea colecciones, desarrolla

catálogos y pone a disposición tanto servicios como programas para apoyar a un perfil específico de personas usuarias.

Cada tipo de biblioteca tiene distintos objetivos y características, pero todas comparten algunos elementos que las hacen funcionar, como los que se mencionan a continuación:

1. La organización de los recursos de información que albergan,
2. un catálogo que permite a los usuarios consultar los recursos con los que cuenta la biblioteca y que da pie al punto 3,
3. servicios para que las personas puedan utilizar las colecciones.

Ahora bien, en la literatura especializada existe una discusión sobre las semejanzas y las diferencias entre el concepto de *biblioteca de educación superior* y *biblioteca académica*.

Quijano (2007) explica que “desde la perspectiva anglosajona, se distingue entre bibliotecas que apoyan la educación de licenciatura (*college*) y aquellas que satisfacen las necesidades de investigación de posgrado. Sin embargo, en México, los términos ‘biblioteca universitaria’ y ‘biblioteca académica’ se utilizan de manera indistinta” (p. xi). Para fines técnicos, en este documento se empleará el término *biblioteca académica*.

Según Varela-Prado y Baiget (2012) las bibliotecas académicas tienen la misión de alinearse con aquella de la institución a la que pertenecen, cuyas funciones son la educación y la investigación. Así pues, se “requieren capacidades profesionales y prácticas que contribuyan a mejorar la investigación y la enseñanza en todas las disciplinas” (p.116).

Garza (1984) sostiene que “la biblioteca académica tiene como objetivo conservar, difundir y transmitir el conocimiento, apoyando las funciones de docencia, investigación y difusión de la institución a la que pertenece” (pp. 25-

26). Además, compara la biblioteca académica con la escolar, ya que ambas comparten el propósito de apoyar programas de enseñanza, y también con la biblioteca especializada, debido a su respaldo a programas de investigación.
Este tipo de biblioteca tiene un impacto significativo en el desarrollo del conocimiento y de la innovación en diversas disciplinas, ya que refuerza y profundiza el aprendizaje de los contenidos curriculares, permite la exploración de aspectos específicos de las disciplinas y proporciona la oportunidad de adquirir habilidades valiosas para la vida profesional y académica de estudiantes y egresados.

Para sintetizar estas definiciones, se presenta el siguiente cuadro comparativo que permite visualizar las diferencias y similitudes entre la biblioteca y la biblioteca académica:

Tabla 1 Cuadro comparativo de definiciones de biblioteca y biblioteca académica

Autor/Año	Definición de Biblioteca	Definición de Biblioteca Académica
Eberhart (ALA, 2013)	Una colección de recursos organizados por profesionales de la información que proporcionan acceso y servicios para educar, informar o entretener a diversas audiencias.	No menciona específicamente el concepto de biblioteca académica.
Orera (s.f.)	Un sistema para la transmisión de información, en el que los macroprocesos se interconectan y donde la biblioteca recibe y procesa información de los usuarios para optimizar sus servicios.	No menciona específicamente el concepto de biblioteca académica.

Enciso (1997)	Un proceso dinámico del fenómeno pregunta-respuesta que responde a las necesidades de los usuarios dentro de un contexto específico.	No menciona específicamente el concepto de biblioteca académica.
Garza (1984)	Un sistema de información que adquiere, procesa, almacena y disemina mensajes a través de su acervo documental.	Conserva, difunde y transmite conocimiento para apoyar la docencia, investigación y difusión dentro de la institución a la que pertenece.
Quijano (2007)	No menciona específicamente una definición general de biblioteca.	En México, los términos "biblioteca universitaria" y "biblioteca académica" se utilizan indistintamente.
Varela-Prado y Baiget (2012)	No menciona específicamente una definición general de biblioteca.	Debe alinearse con la misión de su institución, apoyando la educación y la investigación en todas las disciplinas.

Fuente: Elaboración propia

Como se observa en el cuadro, las definiciones de biblioteca enfatizan la organización de la información, el acceso a los recursos y la función de ofrecer servicios a los usuarios. Mientras que algunas definiciones adoptan un enfoque más tradicional (Eberhart, 2013), otras consideran a la biblioteca como un sistema dinámico e interconectado (Orera, s.f.; Enciso, 1997). En cuanto a la biblioteca académica, los autores coinciden en que su función principal es apoyar la enseñanza y la investigación dentro de su institución. Esta definición es la que se tomará como base en este trabajo, ya que permite analizar cómo los datos pueden optimizar su funcionamiento y sus servicios.

La importancia de la biblioteca académica en la actualidad radica en su papel fundamental como centro de recursos e información para la comunidad educativa y de investigación. A continuación, se detallan algunas razones clave que subrayan su relevancia:

- **Apoyo a la educación continua:** en un mundo que experimenta cambios rápidos y constantes, la educación continua se vuelve cada vez más crucial. Las bibliotecas académicas proporcionan acceso a recursos actualizados y materiales organizados de aprendizaje en línea, lo que permite a estudiantes, profesores y profesionales mantenerse al día con los avances en sus campos de estudio y su desarrollo profesional.

- **Fomento de la investigación innovadora:** las bibliotecas académicas son centros de investigación que ofrecen acceso a una amplia gama de bases de datos, revistas académicas y libros especializados. Facilitan la investigación innovadora al proporcionar a los investigadores los recursos necesarios para explorar nuevas ideas, desarrollar proyectos y publicar resultados.

- **Promoción del pensamiento crítico y la alfabetización informacional:** en un entorno digital saturado de información, es crucial desarrollar habilidades de pensamiento crítico y alfabetización informacional. Las bibliotecas académicas ofrecen programas y recursos que enseñan a los usuarios a evaluar la calidad y la fiabilidad de la información, así como a utilizar herramientas digitales de manera eficaz y ética.

- **Fomento de la colaboración y la comunidad académica:** las bibliotecas académicas son espacios de encuentro, donde estudiantes, profesores e investigadores pueden colaborar, compartir ideas y trabajar en proyectos conjuntos. Estos espacios fomentan la creación de redes y

el intercambio de conocimientos entre miembros de la comunidad académica.

- **Preservación del patrimonio cultural y científico:** las bibliotecas académicas desempeñan un papel crucial en la preservación del patrimonio cultural y científico. Conservan y protegen materiales raros y únicos, como manuscritos antiguos, archivos históricos y colecciones especiales, con lo que aseguran que estén disponibles para las generaciones futuras.

- **Adaptación a las nuevas tecnologías**: en un entorno tecnológico de constante evolución, las bibliotecas académicas deben adaptarse y aprovechar las nuevas tecnologías para satisfacer las necesidades cambiantes de los usuarios. Esto incluye la implementación de sistemas innovadores de gestión de bibliotecas, el desarrollo de plataformas digitales y la integración de herramientas de análisis de datos para mejorar los servicios bibliotecarios.

La biblioteca académica sigue siendo un pilar fundamental en el panorama educativo y de investigación en la actualidad. Su capacidad para proporcionar acceso a recursos actualizados, fomentar la innovación y el pensamiento crítico, promover la colaboración y la comunidad académica, preservar el patrimonio cultural y adaptarse a las nuevas tecnologías la convierten en un recurso invaluable para la sociedad en general.

La necesidad de gestionar grandes volúmenes de información y proporcionar servicios eficientes ha impulsado la adopción de diversas tecnologías. Una de las más significativas es la automatización, que ha transformado radicalmente la manera en que las bibliotecas operan.

La automatización en bibliotecas no solo se refiere a la mecanización de procesos administrativos, sino que también abarca la implementación de sistemas inteligentes para la catalogación, la recuperación de información y la personalización de servicios para los usuarios. Esta evolución tecnológica permite a las bibliotecas académicas optimizar sus recursos, mejorar la accesibilidad de sus colecciones y ofrecer una experiencia más ágil y adaptada a las necesidades específicas de estudiantes e investigadores.

1.2. Automatización de bibliotecas

La tecnología se desarrolla de manera paulatina y por etapas, no sucede como surgimiento espontaneo. Para poder implementar otras metodologías de métricas de datos o incluso inteligencia artificial, se necesitan datos sistematizados y organizados que puedan ser utilizados con distintos propósitos y es en este sentido que surge la automatización.

Por un lado, Voutssás (2019) dice que “se entiende a la automatización como el uso de un dispositivo —mecánico, eléctrico, electrónico, etcétera— para minimizar o sustituir en un proceso a un operador humano” (p. 1).
Ese es un concepto interesante porque habitualmente se considera que para que la automatización pueda llevarse a cabo se requiere que estén inmersas las tecnologías de la información, pero desde la perspectiva que propone este autor, puede decirse que “existieron múltiples dispositivos no computacionales que fueron usados en las bibliotecas a lo largo del tiempo’’ (Voutssás, 2019, p. 1).

Por otro lado, Buonocore define la automatización como un proceso que tiende a la mecanización de actividades, lo cual desempeña una función importante en la técnica documentalista (como se cita en Bravo, 2022).

Además, Tomás Saorín define a la automatización de bibliotecas como la aplicación de la tecnología en estas con el fin de automatizar los procesos y los

instrumentos de información, creando así un corpus de conocimientos y experiencias profesionales bastante amplio (como se cita en Bravo, 2022).

Bronsoiler Frid (1989) la define como

> La aplicación de las computadoras a los servicios que se ofrecen y a las operaciones que se realizan en una biblioteca, contribuyendo a su efectividad. La automatización de bibliotecas es la implementación de herramientas que agilizarán las tareas bibliotecarios en menor tiempo y mejor precisión (p. 22).

En algunas ocasiones se piensa que las bibliotecas digitales y la automatización es lo mismo, pero esta idea no es correcta. Una biblioteca digital es la biblioteca virtual que es la que mayor énfasis detalla en su obra, gran parte de su colección documental es en formato digital y, por lo tanto, el acceso a la información será universal, sin límite alguno. No cuentan con la presencia de espacios físicos, para que se agilicen los servicios bibliotecarios con mayor rapidez y en tiempo real.

A continuación, se mencionarán algunos antecedentes de la automatización que fueron publicados en la obra *Antecedentes de la automatización en México* de Juan Voutssás (2019) y en los que podemos seguir el recorrido que han tenido que atravesar las bibliotecas para llegar hasta la época actual.

Voutssás resalta que la automatización de las bibliotecas ocurrió mucho antes de la implementación de equipos de cómputo, dado que existen diversos tipos de herramientas que se han empleado para la mejora de la administración, organización y servicios bibliotecarios.

Por ejemplo, el autor menciona a Calímaco de Cirene, quien creó los *pinakes* o tablas, que fueron metadatos con registro temático en 120 volúmenes (esto hacia el año 265 a.C.).

Voutssás (2019) también explica que los sistemas de organización temática son ejemplos de tecnología asociada a bibliotecas. Frederick Rostgaard se preguntaba ya en 1697 si ordenar los libros por orden alfabético sería lo óptimo o si se debería preferir una clasificación temática, lo que dio paso posteriormente a la clasificación del Vaticano, al Sistema Decimal de Melvil Dewey, al de la Biblioteca del Congreso de los Estados Unidos de Cutter y al de Ranganathan (p. 2).

De acuerdo con la perspectiva de este autor, podría considerarse que el código de catalogación angloamericano, que unificaba las reglas para realizar ese proceso, también fue un avance tecnológico, aunque “su manifestación física no fue evidente”, dado que toda automatización es tecnología, aunque no toda tecnología es automatización (Voutssás, 2019, p. 3).

Desde 1775, se tiene documentado el uso de tarjetas catalográficas en lugar de volúmenes manuscritos para el registro de las obras. El abate Rozier (autor y miembro de la Academia de Ciencias de París) incluso consignó que “las tarjetas ofrecen una gran facilidad para estos índices, ya que a través del ordenamiento que permiten, pueden sustituir a los índices en volúmenes manuscritos y eliminan la necesidad del recopiado frecuente (Voutssás, 2019, p. 5).

Por otra parte, Ezra Abbot diseñó cajones de tamaño estandarizado para colocar las tarjetas en forma vertical, pero fue Melvil Dewey quien le da al catálogo una forma estandarizada. Hasta 1900, la empresa Library Bureau era la encargada de comercializar las cajoneras gabinete, así como cualquier otro mueble tipo archivero que se utilizara para el manejo de tarjetas y otros accesorios.

Voutssás (2019) menciona que la “Ciclopledia anual y registro de acontecimientos importantes del año 1894 Appletons”, describe una máquina visualizadora de catálogos de biblioteca del siglo XIX:

> Una caja con tapa de cristal que llega hasta el pecho, en su interior se encuentran dos prismas hexagonales girando sobre sus ejes, a uno de los cuales se encuentra acoplada una manivela. Alrededor de estos y cayendo casi hasta el fondo de la caja se encuentra una cadena sin fin formada por ligeros marcos metálicos en los que se pueden insertar tarjetas con las descripciones de los libros en la biblioteca. Mirando a través de la tapa de cristal, el usuario ve cuatro larga páginas de tarjetas ordenadas alfabéticamente y puede girar la manivela en cualquier dirección para traer a la vista cualquier otra página que desee (p. 6).

Más tarde, en 1898, James Rand creóun sistema racionalizado de archivado usando tarjetas, índices, divisores, pestañas de carpetas, etiquetas. Más tarde, pudo fabricar su propio sistema de índices tras fundar la Rand Ledger Company. Y después, en 1915, su hijo lo perfeccionó con las cajoneras desplegables y le puso el nombre de American Kardex Company (Voutssás, 2019, p. 6).

Luego, en 1883, Joseph C. Rowell envió una carta a la reunión de la conferencia de Bibliotecas Americanas en Nueva York. En dicho escrito se anexaba una tarjeta catalográfica de muestra que había realizado con la ayuda de una máquina de escribir (inventada en 1873), así comunicaba que desde ese momento pensaba hacer todas las tarjetas de su biblioteca con ese modelo (Voutssás, 2019, p. 8).

En 1885, Dewey pidió que se le fabricara una máquina de escribir especial para bibliotecas. El encargado para hacerla fue Hammond. Con la creación de la Hammond Card Cataloger se podía realizar cambios en el tamaño de la letra y en el alfabeto que se iban a utilizar. Para 1902, “65 de 66 bibliotecas encuestadas ya usaban la máquina de escribir como medio para fabricar sus tarjetas catalográficas” (Voutssás, 2019, p. 9).

Ahora bien, sobre la automatización utilizando computadoras, la literatura menciona que el desarrollo de las computadoras surge durante la Segunda Guerra Mundial y posterior a ella.

Respecto a las bibliotecas, tenemos como primer ejemplo el experimento llevado a cabo por Harley Tillet, en 1954 (Voutssás, 2019, p. 57) en el que realizó una recuperación bibliográfica usando el método de indización coordinada Uniterm. También se hizo uso de las colecciones “de informes con 15 mil términos en la computadora IBM-701 de la estación de pruebas de Ordenanzas Navales de la Marina de Estados Unidos”. Este sería considerado como el primer programa de recuperación de información hecho en una computadora de propósito general.

Ya para 1952, Jesse Shera (bibliotecólogo pionero de la automatización de bibliotecas en Estados Unidos) fundó el Centro de Investigaciones en Documentación y Comunicación, siendo el primero en investigar y desarrollar de forma sistemática proyectos de recuperación de información apoyándose en computadoras, que influyó durante las siguientes tres décadas. El trabajo e investigación de su centro contribuyó en el surgimiento de OCLC en 1967 (Voutssás, 2019, p. 147).

Voutssás (2019) menciona otro ejemplo: la Biblioteca Nacional de Medicina de Estados Unidos, en la cual se compilaban las fichas bibliográficas del Index Medicus manualmente. Luego, en 1958, se cuestionó si se podría mecanizar el índice. General Electric concursó para desarrollar el sistema, que entró en servicio en 1964 y, en su momento, fue el mayor sistema de almacenamiento y recuperación del mundo. En 1971, se instaló una versión en línea al que se le llamó “Medline”. En un inicio, podía dar servicio a 25 usuarios simultáneamente; para principios de los noventa, podía atender varios miles al mismo tiempo (p. 65).

L. R. Bunnow realizó un documento para la empresa aeronáutica Douglas en 1960. En ese informe se hacía la recomendación de un sistema de recuperación informatizado, con el cual la producción de tarjetas de catálogo se incluía, por primera vez. Esta propuesta consideraba que, de un único registro catalográfico legible por máquina, se podían obtener múltiples productos: tarjetas de catálogo impresas, bibliografías, etc. Aunque para ese momento aún no existía el formato MARC (Voutssás, 2019, p. 66).

El acrónimo MARC significa Machine Readable Cataloging, o catalogación legible por máquina, y fue desarrollado por la Library of Congress junto con otras bibliotecas seleccionadas de 1965 a 1968. Hay que resaltar que MARC surge como un esfuerzo de la Biblioteca del Congreso con el fin de diseñar e implementar los procedimientos para automatizar las funciones de catalogación, indizado, búsqueda y recuperación de información y, a su vez, convertir las tarjetas catalográficas existentes a un formato electrónico (Voutssás, 2019, p. 93).

En 1967, se fundó el Ohio College Library Center (OCLC) con el propósito de fabricar y distribuir tarjetas a todas las bibliotecas de las escuelas del estado y así reducir costos. Con el tiempo, también brindaron el servicio de recuperación de información bibliográfica en línea (Voutssás, 2019, p. 95).

Todo lo antes mencionado, son ejemplos muy precisos de la integración de métodos (Sistemas de Organización), materiales (tarjetas catalográficas) o herramientas (la máquina de escribir) que dieron pie a lo que surgiría después.
De esa manera, la evolución de la automatización nos permite entender el surgimiento de cada uno de los elementos que la integran hoy en día, así como los procesos que pueden resultar de utilidad para la aplicación de otros métodos. Además, estas tecnologías han mejorado significativamente la experiencia del usuario, por lo que conocer la evolución de estas permite a las bibliotecas diseñar y ofrecer servicios más intuitivos y personalizados.

Los usuarios modernos esperan rapidez y facilidad de acceso a la información. Las bibliotecas que entienden la evolución de la automatización pueden satisfacer estas expectativas y mantener su relevancia como centros de información.

1.2.1. Sistemas Integrales de Automatización de Bibliotecas

La automatización aplicada a bibliotecas ha evolucionado significativamente los Sistemas Integrales de Automatización de Bibliotecas, que permiten gestionar eficientemente diversas operaciones, desde la catalogación hasta el préstamo de materiales. Estos sistemas no solo optimizan los procesos internos, sino que también generan grandes volúmenes de datos sobre el uso de los recursos, las preferencias de los usuarios y las tendencias de consulta.

Al integrar Big data en este contexto, las bibliotecas académicas pueden analizar estos datos de manera avanzada para obtener resultados valiosos, mejorar la toma de decisiones y personalizar los servicios ofrecidos. La combinación de automatización, Sistemas Integrales de Automatización de Bibliotecas (en lo subsecuente SIAB) y datos masivos no solo transforma la gestión de las bibliotecas, sino que también potencia su capacidad para adaptarse a las necesidades cambiantes de los usuarios y contribuir más efectivamente al ámbito académico.

Pero ¿qué son los SIAB? José Luis Herrera Morrilla (2010) menciona que hay diferentes objetivos de la aplicación [de Sistemas integrales de automatización], dependiendo el objetivo específico de la biblioteca.
Aunque de forma general, los SIAB son programas o software de computadora que organizan un fondo bibliográfico y tienen la capacidad de desarrollar sistemas multifuncionales a través de distintas operaciones.

Además, permiten (Herrera Morrilla, 2010):

- Servicios a los usuarios: mayor número de consultas en menor tiempo, consultas más específicas o generales de acuerdo con las necesidades e inquietudes del usuario.
- Seguridad: al aumentar la precisión en las consultas, disminuye el número de erratas, por lo que los resultados son confiables
- Cooperación: constitución de redes, préstamo interbibliotecario, tanto a nivel local-institucional como a nivel estatal o nacional, etcétera.
- Normalización de la documentación: se establecen diversos estándares y se utilizan diversas reglas para una descripción más homologada acorde a los reglamentos y procesos de catalogación y clasificación.
- La producción de todo tipo de estadísticas, listas, catálogos: es decir, la información contenida en el catálogo puede ser visualizada en diversas modalidades conforme a las necesidades del bibliotecario o de la comunidad.

  Liberación y control de la circulación de las publicaciones: tener colecciones especificas dentro del acervo, que puedan ser utilizadas por los usuarios, según lo establecido por la biblioteca.

Puede observarse que la aplicación de sistemas de automatización facilita las tareas del bibliotecario, además de permitir otras que de no existir sería casi imposible de cumplir, como el análisis de datos.

Ningún ser humano tiene la capacidad de procesar miles y millones de datos a un ritmo imparable y de diversas fuentes, por lo que la implementación de este tipo de software de bibliotecas no solo apoya a las tareas existentes, también permite la expansión de ellas de una manera más inteligente.

Ahora bien, los softwares fueron diseñados para facilitar la organización mediante las normas de descripción (que daban pie a las bases de datos), así como para elaborar estadísticas de algunos servicios o áreas de la biblioteca.

Aunque existen diversas denominaciones para este tipo de sistemas, para este trabajo uasaremos la de Sistemas Integrales de Automatización de Bibliotecas (SIAB).

De acuerdo con Lourdes David, un SIAB

> Es aquel que tiene una base de datos en común para realizar todas las funciones básicas de una biblioteca. Un sistema integrado de bibliotecas permite a la biblioteca vincular las actividades, por ejemplo, la circulación con la catalogación, gestión de publicaciones seriadas, etc., en un momento dado. Hace uso de un servidor de archivos y clientes en una red de área local. La mayoría de los sistemas de gestión de bibliotecas tienen los siguientes módulos: catalogación y OPAC, circulación, adquisiciones, gestión de publicaciones seriadas y el módulo de préstamo interbibliotecario (David, 2001, como se cita en Arriola y Montes, 2014, p. 48).

Además, Arriola y Montes (2014) mencionan que

> Un sistema de automatización de bibliotecas es aquel que posee un conjunto de módulos que abarcan las actividades bibliotecarias más importantes, los cuales están relacionadas entre sí, ya que comparten una misma base de datos, aunque dichos módulos tienen funciones distintas están unificados para facilitar su control, y de esta manera ayudar a mejorar la eficiencia y eficacia de los procesos, servicios y de la gestión general de la biblioteca (p. 49).

Esto puede ser visualizado en la ilustración 1:

Ilustración 1 Estructura del SIAB

Fuente: De Arriola y Montes, 2014.

Algunas características que deben tener los SIAB para garantizar su funcionalidad son los que se menciona a continuación de acuerdo con Open Source Software, 2005, como se cita en Arriola y Montes (2014):

> Fiabilidad: el tiempo que un sistema puede permanecer en operación sin intervención del usuario.
> Calidad: el número de errores en un número fijo de líneas de código.
> Seguridad: lo resistente que el software es para no autorizar acciones fuera de protocolo (por ejemplo, virus).
> Flexibilidad: la facilidad con que el software puede ser personalizado para satisfacer las necesidades específicas y que se puedan ejecutar en diferentes tipos de dispositivo.
> Gestión de proyectos: la capacidad de organizar los proyectos en desarrollo.
> Estándares abiertos: los documentos creados con un tipo de software deben de ser leídos y trabajados en cualquier otro software.
> Los costos de cambio: el costo de pasar de un sistema a otro.
> Costo total de propiedad: la totalidad de los gastos durante la vida útil del software.
> Facilidad de uso: lo fácil y amigable que es usar el software (p. 52).

Asimismo, tal como se mencionó en las definiciones, los SIAB tienen una estructura modular que controla cada tarea, tal como se mencionó en la figura 1. Al inicio de su creación, estos módulos no estaban interconectados entre sí como se muestra en la figura, pero con la evolución de los SIAB (que se mencionará más adelante) se detectaron las ventajas de la interconexión entre las bases de datos.

En realidad, existen diversos tipos de SIAB: de código abierto —que puede ser adaptado y programado de acuerdo a las necesidades de la institución en cuestión—, con licencia por tiempo limitado o para un número determinado de computadoras —es decir, que se realiza el pago por el uso del software— y con diferentes características y herramientas. Para elegir un SIAB, es necesario considerar las características y necesidades de la biblioteca en cuestión.

Por ejemplo, algunos de los sistemas más utilizados en México son Aleph (de licencia), Janium (de licencia), SIABUC (de licencia) y Koha (software libre). La figura 2 muestra los módulos con los que cuenta el software libre Koha.

Ilustración 2. Koha desde la interfaz del bibliotecario

Fuente: *captura realizada al sistema KOHA el día 20 de septiembre 2024.*

Ahora bien, una forma de conocer a mayor profundidad los SIAB para poder elegir el más apropiado, es a través de la comparativa. A continuación, se presenta la tabla 1, en la que se analizan cuatro softwares que se utilizan en el territorio

mexicano. Se consideraron los módulos con los que cuentan, si es de código abierto o cerrado, idioma y las instituciones que lo utilizan.

Tabla 2 Comparativa de los módulos de los SIAB

SIAB	Módulos	Código	Idioma	Instituciones que lo utilizan
Janium	• OPAC • Control bibliográfico • Control de autoridades • Circulación • Control de suscripciones • Adquisiciones • Reportes y estadísticas • Seguridad avanzada • Administración • OAI-PMH	Cerrado	Español	○ Universidad Autónoma del Estado de México ○ Universidad Juarez del Estado de Durango ○ Sistema de Bibliotecas del Subsistema de Universidades Tecnológicas
Aleph	• Adquisiciones • Administración • Catalogación • Préstamo • Ejemplares • Búsqueda (OPAC) • Revistas • Utilidades	Cerrado	Español Inglés	○ Universidad Nacional Autónoma de México ○ Secretaría de Cultura– Dirección General de Bibliotecas Públicas ○ Intituto Politécnico Naciona
SIABUC	• Adquisiciones • Análisis • Consultas • Publicaciones periódicas • Inventario • Préstamos • Estadísticas • Publicaciones en web	Cerrado	Español	○ Universidad de Colima ○ Biblioteca Nacional de Colombia

Koha	• Administración • Adquisiciones • Catalogación • Autoridades • Publicaciones periódicas • Circulación • OPAC • Informes • Herramientas	Abierto	Español Inglés	• Instituto Nacional de Cancerología • Universidad La Salle • Universidad de Monterrey • Universidad Autónoma Chapingo, División de Ciencias Económico Administrativas • Universidad Autónoma de la Ciudad de México

Fuente: elaboración propia.

Es preciso aclarar que dependerá de la versión del SIAB, del tamaño de la biblioteca, del número de usuarios, entre muchas otras cuestiones, para saber los requerimientos técnicos específicos que requerirá el software para operar correctamente.

La importancia de un Sistema Integral de Biblioteca radica en su capacidad para optimizar la gestión de los recursos de información y conectarlos con los servicios bibliotecarios de manera eficiente y efectiva. Algunas razones por las que un sistema integral de biblioteca es crucial se presentan en los siguientes párrafos.

Eficiencia en la gestión de recursos:

- Automatización de procesos: un sistema integral permite la automatización de tareas rutinarias como la catalogación, el préstamo y la devolución de materiales, con lo que se reduce el tiempo y esfuerzo del personal bibliotecario.

- Gestión de colecciones: facilita la gestión de inventarios y colecciones, ayudando a mantener tanto registros precisos y actualizados sobre el estado como la ubicación de los materiales.

Mejora de la experiencia del usuario:

- Acceso fácil a la información: proporciona a los usuarios acceso rápido y sencillo a los catálogos en línea, permitiéndoles buscar y localizar materiales desde cualquier lugar.
- Autoservicio: implementa funciones de autoservicio, como la renovación de préstamos y la reserva de materiales, lo que mejora la fluidez del servicio.

Optimización de la información:

- Bases de datos integradas: un sistema integral puede gestionar múltiples bases de datos y recursos electrónicos, con lo que se permite un acceso más eficiente a una variedad de fuentes de información.
- Análisis de datos: facilita la recopilación y el análisis de datos sobre el uso de la biblioteca, lo que puede informar sobre decisiones estratégicas y mejorar la planificación de servicios.

Mejora en la comunicación y colaboración:

- Interoperabilidad: permite la integración con otros sistemas y redes bibliotecarias, y así se facilita la cooperación interbibliotecaria y el intercambio de recursos.
- Comunicaciones automatizadas: mejora la comunicación con los usuarios a través de notificaciones automatizadas sobre préstamos, devoluciones, reservas y eventos.
- Inclusión digital: facilita el acceso a recursos digitales y servicios en línea, promoviendo la inclusión digital y el acceso equitativo a la información para todos los usuarios, independientemente de su ubicación o condición socioeconómica.

- Personalización del servicio: permite personalizar la experiencia del usuario, adaptándose a sus necesidades y preferencias individuales.

Adaptabilidad y escalabilidad:

- Flexibilidad: un sistema integral puede adaptarse a las necesidades cambiantes de la biblioteca y sus usuarios, al permitir la incorporación de nuevas funcionalidades y tecnologías a medida que evolucionan.
- Escalabilidad: es capaz de crecer junto con la biblioteca, al manejar un aumento en el volumen de materiales y usuarios sin comprometer la eficiencia.

En resumen, un sistema integral de biblioteca es esencial para mejorar la eficiencia operativa, optimizar el acceso y la gestión de la información, mejorar la experiencia del usuario, y asegurar la sostenibilidad y relevancia de las bibliotecas en el entorno digital actual.

Uno de los elementos fundamentales para los SIAB es contar con una interfaz para mostrar a los usuarios las colecciones que alberga la biblioteca, sea desde la institución o fuera de ella. A esto último se le llama OPAC o catálogo público, el cual se convierte en la ventana que proporcionan los bibliotecarios para que los usuarios puedan explorar y buscar en las colecciones.

Al elemento descrito arriba se le conoce como OPAC, por sus siglas en inglés, Online Public Access Catalog (catálogo de acceso público en línea) y es una de las herramientas de los SIAB en la que este trabajo va a centrarse. Sin embargo, antes de profundizar en el análisis de los catálogos públicos, es importante entender que su funcionamiento depende de los registros bibliográficos que presentan al público. Estos registros son el resultado de un cuidadoso proceso de catalogación, el cual garantiza la organización y accesibilidad de la información que el OPAC pone a disposición de los usuarios

1.2.1.1. Organización documental: catalogación y registros catalográficos

La catalogación y clasificación aseguran que los datos bibliográficos sean precisos, coherentes y estandarizados. Esta calidad de datos es esencial para cualquier análisis de Big data, ya que permite obtener resultados fiables y útiles. Los registros bien organizados y exactos facilitan la integración de múltiples fuentes de datos y mejoran la capacidad de realizar análisis complejos.

Pero para que se puedan procesar bajo el enfoque de otras metodologías, se requiere garantizar la calidad de los datos que, en este caso, se lleva a cabo a través de la organización bibliográfica y documental. Este es un proceso intelectual en el que, a través de la catalogación descriptiva, la catalogación por materia o temática y la clasificación bibliográfica, se busca representar la forma, el contenido y la localización de los recursos de información. Para elaborar esa representación se utilizan reglas, normas y estándares internacionales.

La catalogación descriptiva es la etapa "donde se extraen diversos datos de la obra que se está procesando, para constituir las fichas bibliográficas" (DGIRE-UNAM, 2007, párr. 1). Se caracteriza por describir un recurso de información para hacerlo identificable: título, autor, año de publicación, características físicas. En otras palabras, es toda la serie de elementos que nos permite diferenciar y recuperar un recurso.

Al obtener los datos bibliográficos de cada recurso, se sistematiza la información. Puede ser en las fichas bibliográficas (ya en desuso) o en registros en formato MARC (Machine Readable Cataloging o catalogación legible por máquina, en español). Así pues, al conjunto de registros se le llama catálogo.

Para ese proceso se utilizan distintas normas y estándares internacionales, tales como Reglas de Catalogación Angloamericanas (RCA) o Recursos, Descripción y Acceso (RDA), entre otras. Esto con el propósito de que todas las bibliotecas en

el mundo realicen el proceso de manera homogénea y se facilite tanto la compartición entre bibliotecas como el acceso por parte de los usuarios.
Sus principales objetivos son los siguientes (DGIRE-UNAM, 2007):

- “Permitirle al usuario encontrar un documento a través del autor o título.
- Conocer lo que tiene la biblioteca.
- Orientarse en la selección de un documento a consultar, dependiendo de su edición, idioma, naturaleza literaria” (párr. 4).

La organización documental basada en estándares internacionales (como MARC, RDA, etc.) facilita la interoperabilidad entre diferentes sistemas y bases de datos. Esto es crucial en el análisis de Big data, ya que permite la integración de datos de múltiples fuentes, con lo que amplía el alcance y la profundidad del análisis.
La catalogación temática es la parte del proceso en el que se le asigna uno o varios puntos de acceso temáticos al registro del recurso de información en proceso, con el fin de que se representen los temas o asuntos de los que trata.

En la catalogación temática se busca representar el contenido del recurso.
Esta tiene algunos objetivos (DGIRE-UNAM, 2007):

- Identificar en el catálogo las obras que tratan sobre la misma temática.
- Acomodar en la estantería el material de acuerdo con el tema principal que traten.

Para llevar a cabo este proceso, se pueden seguir diversos listados organizados, tales como los encabezamientos de materia, tesauros y los sistemas de clasificación, como el Decimal Dewey o el de la Library of Congress. Utilizando estos dos tipos de catalogación es como se obtiene la información para la elaboración de los registros catalográficos.

La Universidad Complutense de Madrid entiende

"Por ficha o registro catalográfico al resultado de representar formalmente el contenido de un registro en un soporte, bien sea físico o digital. Los elementos descriptivos irán dispuestos en un orden preestablecido y normalizado, para así poder identificarlo de forma unívoca. El conjunto de fichas de una institución dará lugar a su catálogo" (s.f., párr. 1).

Ilustración 3. Registro de catálogo

Detalles de la obra

Guardar

Petición anticipada
Solicitar reproducción
Solicitar en préstamo interbibliotecario (acceso para bibliotecas)
Encontrar más sobre estos temas

Ver signatura/s Índice/Resumen **Registro del catálogo**

Principio y fundamento de la perspectiva [Manuscrito]
Barbaro, Daniele 1513-1570

Autor personal:	Barbaro, Daniele (1513-1570)
Título uniforme:	[La pratica della perspettiva. Español]
Título:	Principio y fundamento de la perspectiva [Manuscrito] / de Daniel Varvaro, Patriarca de Aquileya ; comentado en lengua castellana de la italiana por Phillipe Lázaro de Goyti, maestro de obras y arquitecto, vecino de Madrid
Publicación:	1643
Descripción física:	126 h. ; 38 x 25 cm.
Nota de contenido:	Tratado de astronomía (h. 91-100v). Tratado de reloxes solares, rrecoxidos de varios autores antiguos y modernos (h. 101-126v)
Fuente de adquisiciones:	Compra Francisco Ruiz de Medrano 1763
Nota tít. y men. res:	Tít. tomado de h. 25
Nota tít. y men. res:	Tít. en el tejuelo: Perspectiva, Astronomia y Reloges
N. área desc. fis.:	En blanco las h. 19, 21, 23, 90, 114
Nota sobre ilustrac.:	En h. 1, grab., a modo de portada, con la leyenda: Beati qvi in Domino morivntvr
Nota sobre ilustrac.:	Dos grab. de arquitectura, entre las h. 6-7 y 11-12: Romae Claudii Duchetti formis, 1585; y Claudii Duchetti formis, Romae 1582. Ambrosius Brambilla, f.
Nota sobre ilustrac.:	Dibujos geométricos y planos intercalados en el texto, o a página doble y entera
Encabez. materia:	Perspectiva
Encabez. materia:	Astronomía
Encabez. materia:	Relojes de sol
Autor personal:	Lázaro de Goiti, Felipe
Título uniforme:	Tratado de astronomía
Título uniforme:	Tratado de relojes solares
Enlace:	Biblioteca Digital Hispánica

Nota. Registro tomado de *Principio y fundamento de perspectiva*, de Daniele Barbaro. Biblioteca Digital Hispánica. Biblioteca Nacional de España
Fuente: Universidad Complutense de Madrid, s.f.

Al conjunto de los registros se le llama *catálogo*. El catálogo representa todo lo que la biblioteca posee. Sin embargo, aquí vale la pena resaltar que, si los registros catalográficos no están elaborados acorde a la normativa internacional y en consideración a los objetivos institucionales considerando su comunidad de usuarios, no representará en su totalidad lo que la biblioteca posee.

En un principio, existieron los catálogos no automatizados: se trataba de una serie de fichas catalográficas en las que venía toda la descripción bibliográfica

de los recursos de información. Podías buscar en las fichas por título, tema o autor.

Posteriormente, y con la evolución de la automatización, los catálogos cambiaron: utilizaron el formato MARC, los puntos de acceso y otros recursos. Entonces era posible que la computadora pudiera recuperar, en segundos, la información solicitada.
En el caso de los SIAB, los catálogos en línea se les denomina OPAC, de los cuales se hablará en el siguiente apartado.

1.2.2. OPAC

En el Glosario de la American Library Association (2000, como se cita en Eserada y Okolo, 2019) definen al OPAC como un catálogo o una base de datos bibliográfica, diseñada para acceder a través de terminales computarizadas, en el que los usuarios puedan buscar y recuperar información de manera directa y efectiva, sin requerir el apoyo de un intermediario humano.

También puede ser definido como una base de datos bibliográfica que describe los recursos de una biblioteca, que permite a los usuarios buscar un documento por autores, títulos, temas o palabras clave desde una terminal. Además, permite imprimir, descargar o exportar los registros desde diferentes medios electrónicos y en distintos formatos (Gohain y Siakia, 2013, como se cita en Eserada y Okolo, 2019).

Para Romero (2005), el OPAC es un catálogo bibliográfico automatizado disponible para su consulta interactiva a través de terminales o nodos de computadoras que cuentan con una conexión a una red, ya sea local o internacional (internet). Al tratarse de un catálogo que reúne de forma automatizada los registros bibliográficos de la colección de una biblioteca particular, está vinculado por medio de ciertas herramientas tecnológicas a los catálogos de otras bibliotecas y puede ser consultado por cualquier persona que

tenga una necesidad de información, que cuente con el equipo de cómputo adecuado y que requiera su uso para ubicar un documento específico.

Para Eserada y Okolo (2019), el OPAC es la puerta de entrada a la colección de la biblioteca. Podemos concluir, entonces, que el catálogo de acceso público es una base de datos bibliográfica que brinda la posibilidad de recuperar registros mediante distintos tipos de búsqueda o el acceso completo a diversos recursos de información, desde la biblioteca u otros lugares.

Aunque en la actualidad su empleo es conocido en el ámbito bibliotecológico, su desarrollo e implementación ha pasado por diversas etapas.

1.2.2.1 Antecedentes

Romero (2005) menciona que el desarrollo de estos catálogos va de la mano con la automatización de las tareas bibliotecarias, y resalta tres momentos primordiales:

- 1.° Comienza a principios de los sesenta, cuando las computadoras solo se usan como máquinas rápidas para la elaboración de las fichas catalográficas.
- 2.° Entre los sesenta y principios de los setenta, cuando se descubrieron las importantes ventajas de la catalogación compartida.
- 3.° A mitad de los setenta, cuando los bibliotecarios se dan cuenta de que podían emplear las computadoras para las tareas cotidianas: catalogación, circulación y control de publicaciones periódicas.

Asimismo, Romero cita a Charles Hildreth quien propuso que el desarrollo de los OPAC puede dividirse en tres generaciones. Sin embargo, al haber publicado su artículo en 1987, complementaremos las aportaciones de Hildreth con la investigación publicada por Marek Nahotko, en el 2020, quien menciona las últimas generaciones.

Primera generación

Creados en los setenta y ochenta, los OPAC se utilizaban, principalmente, para buscar artículos, a través de un número limitado de metadatos básicos como autor, título, signatura topográfica o encabezamientos de materia (Nahotko, 2020).

Las opciones de búsqueda eran limitadas, dado que se hacía a partir de frases específicas, por ejemplo, el título del recurso completo y correcto, pues se requería la coincidencia de carácter por carácter entre la consulta del usuario y el registro.

Los primeros OPAC eran únicos; es decir, eran desarrollados a "la medida'' por las bibliotecas que los usaban. Los OPAC comerciales surgen en 1978; los desarrolladores trataron de hacerlos funcionar e, incluso, visualmente los crearon similares a los catálogos de fichas, pero no eran sencillos de utilizar.
Estos OPAC seguían los principios básicos de consulta de los catálogos impresos (Hildreth, 1984, como se cita en Romero, 2005).

Segunda generación

Estos OPAC aparecieron a finales de los ochenta, fueron propuestos por empresas que vendían SIAB para bibliotecas y otros insumos.

En esa época, se mantenía el modelo de búsqueda conocido como el catálogo de tarjetas y la funcionalidad de las bases de datos bibliográficas. También era posible buscar por encabezamientos de materia (lenguajes controlados) o palabras clave.

Se agregaron nuevos campos, como el de búsqueda con operadores booleanos. Esta la innovación que permitió que avanzaran las búsquedas por combinaciones de atributos y diferentes elementos, como las palabras clave con clasificación.

Además, algunos de los catálogos ofrecían la función de “modo de visualización del registro” (corto, medio, completo) y el nivel de soporte de la interfaz de usuario, por ejemplo, diseños y lenguaje distintos dependiendo si eras un usuario principiante o uno experimentado.

Además, se conectaba con el módulo de circulación del SIAB, lo que daba la oportunidad de nuevos resultados y combinaciones.

Para este momento, ya se había logrado rescatar las características del catálogo tradicional y mezclarlo con la flexibilidad de los sistemas de recuperación, empleados hasta entonces solo en los servicios de índices y resúmenes.

Por otra parte, algunos problemas de estos OPAC fueron los errores frecuentes, la pobre o nula asistencia en la resolución de problemas del sistema, los problemas utilizando lenguajes controlados y la pobre organización de los sets de respuesta. También estaban la dificultad para encontrar los términos por materia, pues no arrojaban resultados de errores tipográficos; los problemas con las abreviaturas e iniciales, los problemas con la lógica booleana y que los usuarios no identificaban las diferencias entre los ficheros, índices y campos (Romero, 2005).

Nahotko (2020) menciona que Hildreth resumió las características de la segunda generación de OPAC. Se mencionan algunas:

- Acceso por materias y palabras clave
- Búsqueda booleana
- Búsqueda por términos de índice
- Revisión de existencias en los estantes
- Registros bibliográficos estándar completos
- Múltiples formatos de visualización
- Visualización de resultados de búsqueda o manipulación de impresión

- Opción de ayuda, sensible al contexto
- Mensajes de error informativos
- Indicaciones de opciones de acción y "cómo hacer"
- Término de búsqueda y rutinas de coincidencia aproximada

Tercera generación

Está en operación desde 1996; se suponía que evitaría los problemas típicos de las generaciones previas, al utilizar interfaces, técnicas y herramientas innovadoras.

Los OPAC podían recuperar información a partir de un lenguaje controlado como no controlado de los registros, así que los resultados se presentaban por relevancia. No solo contenían metadatos sobre los libros impresos, también de revistas, recursos audiovisuales y electrónicos.

Además, el módulo de circulación tuvo un gran desarrollo como control de los préstamos, de los datos específicos sobre su localización y de la lista de copias que tenía la biblioteca.

La interfaz tenía un menú y los datos se presentaban a través de ventanas con iconos. Así, el OPAC comenzó a brindar acceso remoto a través de internet y los usuarios perdieron contacto directo con los bibliotecarios.

Las funcionalidades de la tercera generación de OPAC de acuerdo con Hildreth (1987, como se cita en Romero, 2005), son las siguientes:

- Incorporación de expresiones del lenguaje natural
- Conversión o coincidencia automática de términos
- Técnicas de recuperación no booleanas
- Métodos de retroalimentación de relevancia
- Apoyo para la navegación inteligente

- Integración de palabras clave, vocabulario controlado y enfoques de búsqueda basados en clasificación

Cuarta generación

Nahotko (2020) resalta que la cuarta generación de OPAC fue propuesta por Behesthi. Esta generación estuvo representada por el uso de muchas herramientas provenientes de la web, incluso consideraba el lenguaje gráfico para la interfaz, también el compartir recursos utilizando el protocolo Z39.50, y usaba tanto el hipertexto como la habilidad de procesar metadatos en diversos formatos como MARC o Dublin Core.

Las opciones de ayuda se presentaban en forma de chat. Los registros de metadatos contenían los *links* con multimedia y los textos completos, además de *links* con recursos de otras bases de datos.

Las imágenes de las portadas, los resúmenes, las tablas y el material complementario también podía verse en el OPAC.

Los usuarios tenían la posibilidad de usar diversas opciones de interfaces, influenciadas directamente por los cambios tecnológicos derivados de las microcomputadoras y el internet, tales como la búsqueda simple y la compleja. Babu y O'Brien resumieron algunas de estas características:

- Interfaz gráfica de usuario
- Disponibilidad de enlaces de hipertexto a través de registros bibliográficos
- Emularon la apariencia y funciones de los motores de búsqueda
- Disponibilidad de texto completo
- Interfaz para buscar toda la información electrónica

Quinta generación

Esta generación también se conoce como OPAC 2.0. Nahotko (2020) menciona a algunos autores como Katie Wilson, Tanja Merčun, Maja Žumer, Sridevi Jetty,

John Paul AK., P. K. Jain y Alan Hopkinson, quienes resumieron en diversos artículos o ponencias las características.

Al incorporar el CMS (Content Management System), el sistema permitió que los usuarios pudieran agregar etiquetas, puntuación y opiniones de los recursos de la biblioteca, funciones y servicios típicos de la Web 2.0.

Además, los usuarios podían personalizar la interfaz del catálogo, guardar resultados, ordenar y resellar los recursos, así como pagar multas.

En este sentido, el elemento más importante para el sistema bibliotecario es el usuario, la información que se necesita para el uso y mejora del OPAC.

Sexta generación

Para la sexta generación de OPAC, se considera el Descubrimiento de Escala Web (WSD, por sus siglas en inglés) como una de las características fundamentales.

El sistema de descubrimiento permitía administrar los recursos de la biblioteca de forma unificada, independientemente del formato y la ubicación del recurso, para este momento puede considerarse que la arquitectura ya no está centrada en el servicio.

Además, proponía la navegación por facetas, los resultados de búsqueda por relevancia y el acceso a los recursos electrónicos, a través de una sencilla interfaz y un índice integrado. De la misma forma, se simplificó más la búsqueda al utilizar el lenguaje natural, así como nuevas capacidades para la navegación por facetas.

Nathoko (2020) cita a los autores Yang y Hofmann, quienes resumieron las características como:

- Único punto de entrada para todos los recursos de la biblioteca
- Interfaz web de última generación
- Contenido enriquecido
- Navegación por facetas
- Cuadro de búsqueda de palabra clave simple que puede ir hacia búsqueda avanzada
- Ranking de relevancia
- Revisión ortográfica
- Recomendaciones a materiales relacionados
- Contribución del usuario (etiquetas, calificaciones, reseñas)
- Fuentes RSS
- Integración con sitios de redes sociales
- Enlaces persistentes

Los diseñadores de esos sistemas conocieron a los usuarios —sus objetivos en cuanto al uso de la información—, las características del autoservicio (en dirección a volver autosuficientes a los usuarios), e intentaron imitar la satisfacción y la interfaz de sitios como Google o Amazon.

Nathoko resalta que es entonces cuando el enfoque del OPAC cambia, y pasa a ser descubridor de información en lugar de recuperador de información.

Séptima generación

Ahora bien, la séptima generación de OPAC utilizó los principios del modelo Requerimientos Funcionales para Registros Bibliográficos (FRBR, por sus siglas en inglés, Functional Requirements for Bibliographic Records) y BIBFRAME (Bibliographic Framework), dado que las Reglas de Angloamericanas serían reemplazadas por las RDA (Resource Description and Access) y adaptadas a FRBF y BIBFRAME.

Ya que estas reglas siguen el enfoque centrado en las relaciones, los OPAC deberían tener la manera de representar mejor las relaciones bibliográficas, así como brindar mejor acceso y considerar más las interacciones del usuario con los metadatos.

Por un lado, las relaciones indican conexiones entre la obra y sus expresiones o manifestaciones y, por el otro, entre los trabajos relacionados. El modelo es útil para trabajos con relaciones extensas, muchas ediciones, traducciones y derivados, a diferencia del modelo previo, centrado únicamente en la manifestación.

Para esta época se intentaba evidenciar la jerarquía entre los elementos bibliográficos de los recursos de información y permitir una búsqueda mejorada, así como una fácil y avanzada navegación en un entorno web.

Octava generación

La generación de los OPAC basados en la nube (también conocidos como OPAC basados en datos abiertos vinculados) sería la octava generación (Nahotko, 2020).

Esta tecnología consideraría buenas prácticas y reglas para interconectar documentos legibles por máquina, conjuntos de datos utilizando URIs y metadatos RDF para mostrar, difundir y fusionar datos en ambientes web.

Esto quiere decir que los datos que integren estos OPAC podrían integrarse en otros recursos de la web. Se aplicaría no solo en registros de metadatos, sino también en lenguajes controlados, lo que incluye ontologías (Nahotko, 2020).

Los registros que antes estaban cerrados, que constituían entes en sí mismos, se convierten en subgéneros de metadatos autónomos, con significado propio y

combinado con relaciones que van más allá del OPAC o de las colecciones de la biblioteca.

Al eliminar la dicotomía entre los metadatos del registro y los lenguajes controlados, se ha permitido que los OPAC puedan involucrarse en un entorno web abierto, lo que hace que los metadatos no solo sean procesados por las computadoras, sino también interpretados por computadoras.

Estas funciones son especialmente útiles para integrar el catálogo de biblioteca y sus colecciones en entornos basados en internet, lo que los vuelve datos útiles para otro tipo de prácticas y metodologías.

1.2.2.2 Objetivos de los OPAC

Ya se han mencionado algunos de los propósitos de los OPAC, que los vuelven herramientas importantes para el trabajo en bibliotecas.

Además, como se dijo anteriormente en el apartado de definiciones, los objetivos que más se resaltan de los OPAC son:

- Describir los recursos de información, las colecciones y su localización de manera precisa
- Brindar información organizada y verídica a los usuarios
- Proporcionar el número deseado de resultados, lo más cercanos a los intereses del usuario
- Proporcionar dicha información dentro de las instalaciones de la biblioteca o fuera de ella

Adicionalmente, podemos agregar los sugeridos por Charmi Ammi Cutter, en 1891, en su texto *Ruler for a dictionary catalog*, para todo catálogo de biblioteca:

- Permitir que cualquier persona encuentre un libro o recurso de información si se conoce su autor, título o materia.
- Cumplir una función de recolección, al mostrar lo que la biblioteca posee de un determinado autor, materia o tipo de literatura. Proporcionar asistencia en la elección de un libro, ya sea por su edición, carácter literario o materia.

1.2.2.3 Características

Ulate (2020) menciona que algunas características generales de los OPAC son:

> Consulta vía web de las colecciones bibliográficas, según las siguientes especificaciones
> -Búsquedas simples
> -Búsquedas avanzadas utilizando incluso operadores booleanos
> -Estrategias de búsqueda con operadores booleanos, palabras clave, diccionario, etcétera
> -Mensajes de ayuda
> -Disposición de recuperación de datos
> -Acceso a otros catálogos
> -Búsquedas en todas las bibliotecas del sistema bibliotecario en cuestión o en una biblioteca particular
> -Búsquedas en un formato particular o todos los formatos (libros, revistas, CD-ROM)
> -Visualización en diferentes formatos de los registros bibliográficos (p. 20).

También podemos resaltar las siguientes:

- Búsquedas exhaustivas: los OPAC pueden navegar entre las colecciones de manera eficiente, los usuarios pueden buscar por palabras clave, títulos, autores, encabezamientos de materia o en combinación de varios de estos elementos.

- Filtros y clasificación avanzados: para refinar los resultados de las búsquedas acorde a lo que quieren y necesitan los usuarios.
- Información sobre la disponibilidad y ubicación de los recursos de información: en dónde se encuentran los recursos de manera física o digital
- Gestión de solicitudes y retenciones: permite a los usuarios administrar sus actividades, préstamo o reserva de manera autónoma.
- Gestión de cuentas de usuario: los OPAC ofrecen funciones de gestión de cuentas personalizadas que posibilitan a los usuarios acceder y controlar sus cuentas de biblioteca, por lo que los usuarios pueden ver su historial de préstamos, fechas de vencimiento, renovar materiales en línea. A este elemento podemos llamarlo “autoservicio”.
- Integración de recursos digitales: actualmente, los OPAC ya integran libros, revistas y otros recursos digitales que los usuarios puedan consultar en un solo sitio.
- Recomendaciones personalizadas: propias de la octava generación de OPAC y relacionado al tema central de este trabajo, algunos OPAC ya tienen el desarrollo suficiente para utilizar algoritmos basados en las preferencias del usuario.

Los OPAC son una herramienta fundamental para las bibliotecas, dado que son una ventana amigable que funciona en dos sentidos: permite a la biblioteca mostrar lo que organiza y resguarda, y deja a los usuarios explorar qué es lo que se alberga en esa instancia.

Como ya se ha mencionado, en los antecedentes que muestran la evolución de los OPAC, permiten, por un lado, comprender que han sido mejorados a partir de las necesidades de los usuarios, pero que también compiten con la eficacia y precisión de otros servicios.

Actualmente y con la posibilidad de que hasta el más simple electrodoméstico se pueda conectar a internet, ¿de qué manera responden los OPAC a las consultas de los usuarios? ¿Es útil comparado con las respuestas y sugerencias que brindan otras herramientas? ¿Es posible vincularlo con otras posibilidades?

Nathoko (2020) menciona que el desarrollo actual de los OPAC, en materia de metadatos, permite acercar a las bibliotecas a las soluciones de Big Data, pero ¿qué es y cómo puede potenciar los OPAC para mejorar los servicios de las bibliotecas?

La evolución de los OPAC ha permitido una gestión más eficiente de la información en bibliotecas y centros de documentación. Sin embargo, con el crecimiento exponencial de los datos y el avance de las tecnologías de inteligencia artificial, surgen nuevos desafíos y oportunidades en la organización, recuperación y análisis de la información. En este contexto, el uso de datos masivos y sistemas expertos ofrece soluciones avanzadas para mejorar la precisión en la búsqueda de información, la personalización de servicios y la toma de decisiones en entornos digitales. El siguiente capítulo explora cómo estas tecnologías están transformando la gestión del conocimiento y la automatización de procesos en distintos ámbitos.

Capítulo 2. Datos masivos y sistemas expertos

La era de la información presenta muchas respuestas, pero miles de preguntas sobre cómo se pueden aprovechar diversos tipos de datos, para la producción de conocimiento de valor.

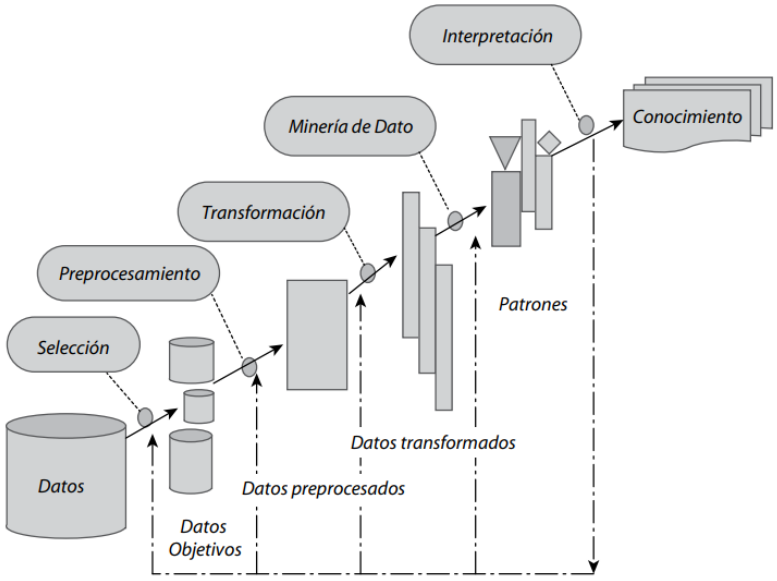
Y aunque los conceptos no son nuevos (dado que se han manejado desde la Guerra fría), sí lo es el paradigma y los usos; también lo son los modelos, las teorías, los enfoques y las herramientas que se utilizan para aprovechar los datos recabados.

El Descubrimiento del Conocimiento en las Bases de Datos (Knowdledge Discovery in Database), mencionado por Patetsky-Shapiro en 1991 (Flores Lagla et al., 2019 y Voutssás, 2022), es aquel que busca desarrollar y aprovechar las técnicas computacionales, así como las herramientas que sirven como apoyo a las organizaciones parala extracción de información útil del veloz crecimiento en volúmenes de datos.

Dentro de este campo se entiende que “el conocimiento es el producto final de un descubrimiento basado en datos” (Flores Lagla et al., 2019, p. 958). Es un proceso automático en el que se combinan descubrimiento y análisis (Timarán-Pereira et al. 2016).

Timarán-Pereira et al., mencionan que el proceso de KDD (por sus siglas en inglés, Knowledge Discovery in Database) es interactivo y se divide en las siguientes etapas (2016, p. 65):

Ilustración 4. Etapas del proceso de KDD



Fuente: Timarán-Pereira et al., 2016, p. 65.

1. Selección: en el momento que se ha identificó el conocimiento relevante y prioritario, y fueron definidas las metas del proceso a partir del enfoque del usuario final, se crea un conjunto de datos objetivo, “al hacer la selección de todo este conjunto o una muestra representativa de este, sobre el cual se realiza el proceso de descubrimiento. La selección de los datos varía de acuerdo con los objetivos del proyecto” (Maimon y Rokach, 2010, pp. 3-4).

“

2. Procesamiento o limpieza: en este momento del proceso se realiza un análisis de la calidad de los datos, “se aplican operaciones básicas como la remoción de datos ruidosos, se seleccionan estrategias para el manejo de datos desconocidos, datos nulos, datos duplicados y técnicas estadísticas para su reemplazo. Para esta etapa es de suma importancia la interacción con el usuario o analista” (Maimon y Rokach, 2010, pp. 3-4).

3. Transformación o reducción: en esta etapa se mejora la calidad de los datos con transformaciones que involucran la reducción de dimensiones y transformación de atributos; este paso puede ser crucial para el éxito de todo el proyecto KDD y generalmente es muy específico (Maimon y Rokach, 2010, pp. 3-4).

4. Minería de datos: en esta etapa se decidirá el algoritmo del proceso de minería de datos, el cual puede ser clasificación, regresión o agrupamiento; esta decisión depende principalmente de los objetivos del KDD y de las etapas anteriores. Dentro de la minería de datos hay dos objetivos principales: la predicción (supervisada) y la descripción (no supervisada y visualización) (Maimon y Rokach, 2010, pp. 3-4).

5. Interpretación o evaluación: en este punto se interpretan los patrones descubiertos y posiblemente se retorna a las anteriores etapas para posteriores interacciones. También se incluye la visualización de los patrones extraídos, la remoción de los patrones redundantes o irrelevantes y la traducción de los patrones útiles, en términos que sean entendibles para el usuario. De igual manera, se consolida el conocimiento descubierto para incorporarlo en otro sistema para facilitar la toma de decisiones (Timarán Pereira et al., 2016, p. 65).

A partir del procesamiento de los datos de manera adecuada, al ser analizados y visualizados, se produce una sugerencia muy clara sobre lo que se debe hacer para mejorar lo que se analizó y no únicamente los números, las estadísticas o los términos inconexos.

Existen diversas metodologías que analizan los datos de distintas maneras, pero que el propósito sigue siendo descubrir valor en ellos y aunque todas han recibido nombres distintos, este trabajo se centrará en dos que podrían ser de utilidad para las bibliotecas: Big data y minería de datos, así como su relación con los sistemas expertos.

2.1 Datos masivos o Big data

2.1.1 Conceptos

Cuando hablamos de Big data, o en español datos masivos o macrodatos, nos referimos a "un gran cúmulo de datos que son almacenados y gestionados para generar información relevante en el contexto de la organización, es decir, generar conocimiento para la acción y que sea aplicable para la toma de decisiones, el diseño de acciones o la elaboración de planes estratégicos" (García-Alsalina, 2017, p. 910).

De este concepto podemos centrarnos en el término "conocimiento para la acción" en otras palabras, un conjunto de datos que nos dan información precisa para que sea posible ejecutar propuestas, acorde a ello.

Islas Monroy (2021) nos da su interpretación de cómo funcionan los datos masivos bajo una metodología de análisis “la manera más sencilla de interpretar el Big data es reconocerla como una base de datos gigantesca que es actualizada diariamente y que se encuentra en expansión. Es decir, algo similar a una presa que acumula cientos de litros de agua y que es liberada cada cierto tiempo para evita que colapse. Bueno, lo mismo ocurre con esta enormidad de base de datos, todos los días es llenada de información y con cada análisis realizado es vaciada un poco” (p. 19).

Islas menciona que las metodologías de datos masivos almacenan tantos datos que se requiere vaciar cada cierto tiempo algunos de ellos, debido a que una de sus características es que en todo momento está llegando información actualizada que puede variar los resultados finales de todo el proceso.

Aunque el concepto de *datos masivos* no es nuevo (surgió el siglo pasado), la diferencia fundamental con la primera vez que se hizo mención del término es el almacenamiento y la capacidad para procesar los datos. Cuando surgió el

término, no se contaba con el equipo de cómputo ni con los softwares con la capacidad de almacenamiento ni la velocidad para procesar en menor tiempo, un gran volumen de datos, que es una de las características más se mencionada del concepto en la actualidad.

Asimismo, con la tecnología actual, no solo es posible analizar datos estructurados, tal como lo menciona Tascón (2018):

> Aquellos otros que provienen de fuentes de información conocidas [que tienen ciertos metadatos y que], por tanto, son fáciles de medir y analizar a través de los sistemas tradicionales, empezamos a poder y querer manejar datos no estructurados: los que llegan de la web, de las cámaras de los móviles y videos, redes sociales, sensores de las ciudades y edificios… La variedad de su origen, además de la rapidez con la que se incrementa su volumen, son algunos de los factores que habían dificultado su análisis hasta ahora. El nuevo software y los nuevos modelos permiten la incorporación a los estudios tanto de un tipo como de otro. Los avances en análisis semántico también permiten estructurar mínimamente parte de los textos escritos por personas de forma automática (párr. 7).

Tascón identifica que, debido a la variedad de formatos en los que se pueden recopilar estos datos, se hizo necesaria la implementación de softwares más potentes que les dieran estructura.

Mauro, Greco y Grimaldi (2016) proponen una definición basada en los aspectos más generales de otras definiciones, al considerar las características de la información que se procesa, la tecnología que se requiere, el valor que agrega a las instancias cuando se analizan y se diseñan planes. “Big data es el activo de información caracterizado por los criterios de volumen, velocidad y variedad, a un nivel tan elevado que requiere de tecnología y métodos analíticos específicos para su transformación en valor” (p.122).

Todos estos agregados concluyen en el siguiente concepto: “Conjuntos de datos extremadamente grandes que pueden ser analizados computacionalmente para revelar patrones, tendencias y asociaciones, especialmente relacionados con el comportamiento humano y sus interacciones” (Léxico, 2014, como se cita en Voutssás, 2022).

Podemos considerar, entonces, que los datos masivos son la recopilación, el almacenamiento y el análisis de diversos tipos de datos, de manera automatizada y en tiempo real.

2.1.2 Antecedentes

El comienzo de la sobrecarga de información no inició en épocas recientes, sino que fue un problema que fue percibido a partir de un hecho concreto: el Censo de los Estados Unidos de América de 1880, el cual tardó ocho años en tabularse con la metodología y herramientas del momento (Palva, 2016, p. 19).

Al percatarse de la problemática, el empresario Herman Hollerith (Cruz, 2013) desarrolló la máquina tabuladora Hollerith de tarjetas perforadas, la cual tenía la capacidad de procesar la cantidad descomunal de información recabada durante el censo. Esta tardó alrededor de un año en tabular los datos. Esto provocó que Hollerith fuera un pionero en nuevas metodologías para el procesamiento de datos y su empresa, por lo que pasó a formar parte de lo que hoy en día conocemos como IBM (L. J. C., 1946)

Para 1932, la sobrecarga de información, derivada con el aumento desmesurado de la población en Estados Unidos y en los países industrializados, conllevó a la gran emisión de números de seguridad social y los registros con los datos de toda la población, así como al desarrollo de la investigación. Estos fueron los principales agentes que requerían, de manera inmediata, un registro de información preciso, organizado y eficaz.

Luego, en el año 1940 (Andalia, 2004) y derivado del flujo de información, las técnicas y las herramientas de la biblioteca comenzaron a ser insuficientes y se comenzó a cuestionar la necesidad de generar nuevas metodologías para el almacenamiento y catalogación de la información.

Un año después, en el periódico *Lawton Constitution*, se mencionó por primera vez “la explosión de la información”. Posteriormente, se desarrolló el término, en marzo de 1964, en un artículo que hacía referencia a la dificultad de gestionar de manera eficiente los volúmenes de información disponibles (Oxford English Dictionary, 2023)

Es por ello por lo que Fremont Rider, bibliotecario de la Universidad de Wesley (Connecticut, EUA), realizó un cálculo, en el cual estableció que las bibliotecas de las universidades de ese país duplicaban su tamaño cada 16 años (Wiegand, 2000) Por lo cual, si esa tasa de crecimiento se mantuviera, la biblioteca de Yale tendría, para el año 2040, un aproximado de 200 millones de volúmenes que ocuparían 9 656 km de estanterías, por lo que, se requerirían 6 mil catalogadores para registrar, clasificar y procesar esa cantidad de recursos.

Para 1966, los sistemas de computación centralizados se encontraban en pleno auge en el sector científico, pero también industrial. Muchas instituciones y organizaciones comenzaron un diseño, un desarrollo y una implementación de sistemas informáticos que les permitían automatizar los sistemas de inventario por completo.

Pero fue hasta 1970 (IBMc, 2024) que surge una estructura elemental para las transacciones de datos que se llevan a cabo actualmente. Edgar F. Codd, matemático “que trabajaba en IBM, publicó un artículo en el que explicaba la forma en la que se podía acceder a la información almacenada en bases de datos de gran tamaño, sin saber cómo estaba estructurada la información o

dónde residía dentro de la base de datos". Esa estructura fue denominada "Teoría de la base de datos relacional".

Para ese momento surgió la demanda de comunicación bidireccional; es decir, las instituciones y organizaciones proveen información y los usuarios, al consumirla, generan una serie de datos que les permiten a los proveedores afinar los contenidos a partir de una serie de indicadores.

Para 1976, los sistemas de Planificación de Necesidades de Material (MRP, por sus siglas en inglés) se diseñaron como herramienta cuyo objetivo era organizar y planificar su información. En ese momento, el uso de las computadoras estaba en auge. Y a partir de ello, Oracle presentó y comercializó el Lenguaje de Consulta Estructurada (SQL, por sus siglas en inglés, Structure Query Language).

Ante la necesidad de datos precisos, en 1985 (datahack, s.f.) "Barry Devlin y Paul Murphy definieron una arquitectura para los informes y análisis de negocio en IBM, lo que posteriormente se convirtió en la base de almacenamiento de datos. En el centro del almacenamiento de datos en general se encuentra la necesidad de almacenamiento homogéneo y una alta calidad de datos completos y exactos".

Como resultado de dichas herramientas, en 1992, Crystal Reports creó el primer informe de base de datos sencillo y compatible con Windows. Así, se aminoró la presión que ya existía sobre el saturado panorama de datos y se permitió que las organizaciones emplearan el análisis de datos de modo asequible.

Derivado de la explosión de la World Wide Web, de acuerdo con IBM (2022) surgió una problemática en la gestión de datos: la incertidumbre ante el altísimo costo que suponía almacenarla. Por esa razón, la volatilidad de la información era un suceso común, dado que los datos resultaban más difíciles de mantener y comenzaron a emerger las plataformas de almacenamiento de datos.

Los autores Usama junto con otros mencionan el término *descubrimiento de conocimiento* en bases de datos y minería de datos, así como otros términos relacionados, en un documento titulado *De la minería de datos al descubrimiento de conocimiento en las bases de datos*:

> Históricamente, la noción de encontrar patrones útiles en los datos ha recibido una variedad de nombres, como minería de datos, extracción de conocimiento, descubrimiento de información, recolección de información, arqueología de datos y procesamiento de patrones de datos [data mining, knowledge extraction, information discovery, information harvesting, data archeology, data pattern processing] (como se cita en Voutssás, 2022, p. 11).

La diversidad de nombres es derivada, principalmente, por enfoque: cada uno de estos procesos tenía uno distinto, aunque los pasos para ser llevados a cabo fueran similares. La cita continúa:

> En nuestra opinión, el Descubrimiento de Conocimiento en Bases de Datos–KDD [Knowledge Discovery in Databases] se refiere al proceso general de descubrir conocimiento útil a partir de datos, y la minería de datos se refiere a un paso particular en este proceso. La minería de datos es la aplicación de algoritmos específicos para extraer patrones de los datos [...] los pasos adicionales en el proceso de KDD, como la preparación de los datos, la selección de los mismos, la limpieza de ellos, la incorporación de conocimientos previos apropiados y la interpretación adecuada de los resultados de la minería, son esenciales para asegurar que se deriven conocimientos útiles de los datos *(*Voutssás, 2022, p.11).

Es decir, la minería de datos es únicamente un paso o una etapa dentro del proceso de Knowledge Discovery in Databases porque

> La aplicación ciega de métodos de minería de datos (criticada con razón como el dragado de datos o data-dredging en la literatura estadística) puede ser una actividad peligrosa, que conduce fácilmente al descubrimiento de patrones inválidos y sin sentido *(*Voutssás, 2022, p.18).

En este último apartado de la cita, se resalta que, si no se conduce con claridad el proceso, no será sencillo encontrar patrones de utilidad. Por ello, es de vital importancia entender qué se quiere saber.

Para 1997 (al igual que en 1880), el ritmo del crecimiento de los datos se había triplicado, y tanto los sistemas como las metodologías eran insuficientes para su correcta gestión.

> Puede que la cantidad de información ascienda a varios miles de petabytes y la producción de cinta y disco alcanzará ese nivel en el año 2000. Pero esto significa que, en unos años, podremos guardarlo todo porque no será necesario eliminar información y que, la mayoría de la información jamás será consultada por un ser humano (Lesk, s.f.).

A finales de los noventa, a pesar del rápido desarrollo de sistemas y softwares para el manejo de información, estos no eran de utilidad pues la extracción de datos conllevaba poseer determinados conocimientos informáticos y tiempo, para así realizar la búsqueda de acuerdo con estrategias previamente diseñadas por los departamentos informáticos. Además, los empleados dependían completamente de ellos para obtener acceso y resultados.

Durante el siglo pasado, la información como tal era lo que brindaba la posibilidad al conocimiento. Sin embargo, a partir de las décadas de los setenta y ochenta, algunos ingenieros y científicos comenzaron a dar mayor relevancia a los datos como materia prima, la cual puede ser analizada con diferentes herramientas, enfoques y metodologías. Es por eso por lo que han proliferado un sinnúmero de términos y diversas metodologías para su análisis.

En 2005, Roger Mougalas, de O'Reilly Media, acuñó el término *Big data* para describir conjuntos de datos tan grandes y complejos que resultan difíciles de manejar y analizar con las herramientas tradicionales de inteligencia empresarial. Esto ocurrió un año después de que surgiera el concepto de *Web 2.0* (EGOS BI, 2021, párr. 10).

En ese mismo año, “Yahoo! creó Hadoop, construido sobre MapReduce de Google. Su objetivo era indexar toda la World Wide Web y, hoy en día, muchas organizaciones utilizan Hadoop de código abierto para analizar grandes cantidades de datos” (EGOS BI, 2021, párr. 11).

Con el crecimiento de las redes sociales y el auge de la Web 2.0, la cantidad de datos generados diariamente se incrementa significativamente. Algunas empresas emergentes innovadoras empiezan a explorar este vasto volumen de información, mientras que los gobiernos también inician proyectos relacionados con Big data. En 2009, el gobierno de la India decidió recopilar escaneos de iris, huellas dactilares y fotografías de sus 1 200 millones de ciudadanos, con lo que creó así la base de datos biométrica más grande del mundo (EGOS BI, 2021, párr. 13).

En 2010, durante la conferencia Techonomy en Lake Tahoe, California, Eric Schmidt comentó que la humanidad produjo 5 exabytes de información desde el inicio de la civilización hasta 2003, y que esa misma cantidad ahora se genera cada dos días (EGOS BI, 2021, párr. 13).

Luego, en 2011, el informe McKinsey sobre Big data (EGOS BI, 2021) mencionaba “que la próxima frontera para la innovación, la competencia y la productividad establecía que tan solo en 2018 EUA enfrentarían una escasez de 140 000–190 000 científicos de datos, así como 1.5 millones de administradores de datos” (párr. 14).

De acuerdo con ORACLE (2021), el Cloud Computing ha potenciado las capacidades del Big data, con lo que ha permitido una escalabilidad flexible que facilita a los desarrolladores crear clústeres temporales para experimentar con subconjuntos de datos. Además, las bases de datos gráficas están ganando relevancia, ya que permiten visualizar grandes volúmenes de datos de manera que su análisis es ágil y completo (párr. 15).

2.1.3 Objetivos

Los modelos de datos masivos son utilizados en diversas instancias, industrias y organizaciones para identificar patrones y tendencias, así como para resolver preguntas, detectar necesidades y obtener información sobre los usuarios y clientes (Bello, 2022).

2.1.4 Características

Diversos autores, mencionan que tres características que representan los datos masivos son las 3V: volumen (muchos datos), variedad (de diversos tipos, con distintos formatos y estructuras, de distintas plataformas o software) y velocidad (para ser recopilados, ordenados, analizados y visualizados).

Aunado a estas 3V, García-Alsina (2017) menciona que al profundizar en el estudio de datos masivos, existen otras características como veracidad (que sean fiables y acorde a la realidad), valor (de los análisis y “la extracción de información crean conocimiento, fuente de innovación, competitividad y ocupación”), visualización (que los “datos se presenten visualmente de manera práctica, dinámica, interactiva y comprensible”), “verificación (incluye las vías para asegurar la integridad de los datos mediante mensajes de autenticación, firmas digitales, certificados de terceros”, etc.), variabilidad (“la velocidad con la que se producen los datos” y su obsolescencia) y viabilidad (que “cualquier proyecto que se defina para gestionar datos masivos debe tener en cuenta su

adaptación a las necesidades organizativas y las características de los datos") (pp. 11-12).

Además, Voutssás (2022) menciona que a pesar de que ahora existen concepciones que toman en consideración diversas características de lo que son los datos masivos, dependiendo el enfoque con el que se analice la metodología, algunos de los elementos que los caracterizan son los siguientes:

> 1. Los datos masivos consisten en el tratamiento y análisis de conjuntos de datos grandes, variados, complejos y dispares.
> 2. Son producidos a una velocidad rápida y provenientes de muy diversas fuentes.
> 3. Que los equipos, programas y procedimientos tradicionales de procesamiento de información: servidores, bases de datos, buscadores, etc., no son suficientes.
> 4. Se requieren métodos, equipos y programas mucho más poderosos, sofisticados y especializados para compilarlos, analizarlos y correlacionarlos.
> 5. Todo con el fin de extraer rápidamente patrones, tendencias y asociaciones de esos datos, principalmente, del comportamiento y de las interacciones humanas (p. 26).

2.1.5 Herramientas

Actualmente, existen diversas herramientas de software libre que permiten analizar datos al volumen y diversidad de Big data. Y que, a diferencia del software de licencia, permite flexibilidad y libertades que el código cerrado no. Debido a la crisis económica mundial que afecta directamente los presupuestos de las bibliotecas y otras instancias de índole académico y cultural, para este trabajo se presentarán algunos de ellas.

2.1.5.1 Apache Hadoop

Ilustración 5. Logotipo Hadoop

Fuente: Apache Hadoop, 2023.

Desarrollador	Apache Software Foundation
Sistema Operativo	Multiplataforma (Windows, Linux, OS X)
Licencia	Apache License 2.0
Lenguaje de programación	Java
Última versión	3.3.6
Sitio oficial	https://hadoop.apache.org/

Fuente: Apache Hadoop, 2023.

Hadoop es un *framework* de código abierto, puede utilizarse gratis. Permite procesar grandes volúmenes de datos en lote usando modelos de programación simples (Ramírez, 2022). De acuerdo con Pérez Marqués (como se cita en Flores y Villacís, 2017): “Es una infraestructura digital de desarrollo creada en código abierto bajo licencia Apache” (p. 36).

Algunas características de este software son las siguientes (Ramírez, 2022; Flores y Villacís, 2017):

- Hadoop gestiona una gran cantidad de datos en petabytes, lo que hace posible que el funcionamiento de sus aplicaciones sea el adecuado.
- Detecta los fallos y vuelve a ejecutar la instrucción, así que prueba con otros caminos de nodos, para que los resultados que se obtienen no sean inconsistentes
- Es escalable, por lo que puede pasar de operar en un solo servidor a hacerlo en múltiples.
- Es un sistema con un alto nivel de seguridad (p. 36).

2.1.5.2 Apache Spark

Ilustración 6. Logo Apache Spark

Fuente: Apache Spark, 2023.

Desarrollador	Foundation, UC Berkeley AMPLab, Databricks
Sistema Operativo	Multiplataforma (Microsoft Windows, OS X, Linux)
Licencia	Apache License 2.0
Lenguaje de programación	Scala, Java, Python, R
Última versión	2.4.5
Sitio oficial	https://spark.apache.org/

Fuente: Apache Spark, 2023.

Flores y Villacís (2017) mencionan que “Spark es una plataforma desarrollada para aumentar la velocidad y eficiencia de las aplicaciones de Big data, pues aceleran las consultas y el procesamiento de grandes volúmenes de datos. Esto optimiza la gestión de Big data al distribuir los procesos entre la memoria de varias máquinas” (p. 39).

Esta es una herramienta gratuita y *open source* que conecta numerosas computadoras y les permite el procesamiento de datos en paralelo. Funciona a través de un aprendizaje automático y otras tecnologías, lo que lo convierte en un sistema eficaz (Flores y Villacís, 2017).

Mientras que Ramírez (2022) dice “la característica más destacable de esta herramienta de Big data es su velocidad, siendo 100 veces más rápida que Hadoop. Spark analiza datos por lotes y también en tiempo real, y permite la creación de aplicaciones en diferentes lenguajes: Java, Python, R y Scala” (párr. 20).

Asimismo, Flores y Villacís mencionan lo siguiente:

> Empezó como un proyecto de investigación en la Universidad de California en Berkeley, en la cual se estaba desarrollando Hadoop y en donde observaron que MapReduce no era lo suficientemente eficiente para procesos de algoritmos iterativos o consultas interactivas; es así que lo redireccionaron a Spark para que este pueda dar soporte para persistencia en memoria y un eficiente sistema de tolerancia a fallos. Spark está escrito en Scala, el cual es más conciso, permitiendo más expresividad en menos líneas de código. No necesita Hadoop para ser ejecutado, sin embargo, es compatible tanto con Hadoop como con su estructura de datos, permitiendo la capacidad de procesar datos de cualquier fuente (2017, p. 41).

Por otro lado, acorde a Flores y Villacís (2017, p. 41) existen varios componentes que definen el ecosistema de Apache Spark, entre los cuales destacan:

- Spark Core Engine se encarga de la planificación, la distribución y la monitorización de aplicaciones ejecutadas en un clúster y Spark SQL, este último concede el soporte para las consultas interactivas.

- Spark Streaming es el elemento encargado de procesar y analizar datos caso en tiempo real. Por otro lado, MLlib es la librería de aprendizaje automático de Spark y otorga una gama amplia de algoritmos de alta calidad.

- GrapX se considera un motor para el análisis de grafos, con este el usuario es capaz de crear, transformar y obtener conclusiones sobre datos estructurados. SparkR es un paquete que integra Spark con el lenguaje estadístico R.

a

2.1.5.3 MongoDB

Ilustración 7. Logo Mongo DB

Fuente: Mongodb.com, 2023.

Desarrollador	MongoDB Inc.
Sistema Operativo	Multiplataforma
Licencia	GNU AGPL v3.0 (drivers: licencia Apache)
Lenguaje de programación	C++
Última versión	3.2.11
Sitio oficial	https://www.mongodb.com/es

Fuente: Mongodb.com, 2023.

MongoDB se trata de una base de datos NoSQL (base de datos no relacional) gratuita y optimizada "para trabajar con grupos de datos que varían con frecuencia, o que son semiestructurados. Es una base de datos distribuida en su núcleo por lo que la alta disponibilidad, escalabilidad y distribución ya se encuentran integradas" (Ramírez, 2022, párr. 17).

Es una aplicación que se utiliza para almacenar datos de aplicaciones y de sistemas de gestión de contenido, entre otras herramientas. Suele ser utilizada por diversas empresas, entre las que se encuentran Bosch y Telefónica.

"Organizaciones de todos los tamaños están usando MongoDB para crear nuevos tipos de aplicaciones, mejorar la experiencia del cliente, acelerar el tiempo de comercialización y reducir costes" (MongoDB, 2017, como se cita en Flores y Villacís, 2017, p. 42).

2.1.6 Ejemplos de aplicación

En una era donde muchas de las cosas que usamos pueden conectarse a internet (el refrigerador, la bocina, la televisión o la lavadora) y desarrollar su

propio concepto (el internet de las cosas o en inglés *internet of things*) es bastante común que las grandes corporaciones busquen recabar la mejor y la mayor cantidad de información sobre los usuarios.

Tenemos el caso de Shein, por ejemplo. Raído (2022) menciona que esta empresa china de producción y distribución de ropa económica (de la que poco se conoce sobre su CEO o quiénes son sus inversionistas) ha logrado superar al grupo textil Inditex (Zara, Pull and Bear, Stradivarius, etc.). Durante muchos años, Inditex se posicionó como la empresa textil de moda rápida que mejor utilizaba los datos: Basándose en las reacciones de los gurús de la moda, de las y los *influencers* o los propios consumidores a lo presentado en los desfiles de Nueva York o Milán, producía réplicas de alta costura para las masas.

Pero algo que Inditex lograba hacer en semanas, se estima que Shein lo hace en horas o días, pero no basándose en lo que se presenta en los desfiles, más bien producen lo que “leen” en redes sociales como TikTok o Instagram y, por supuesto, las tendencias que arroja Google Trends Finder.

Se sabe que en los equipos de trabajo detrás de empresa china, se combina el esfuerzo de diseñadores y analistas de datos de tendencias de moda, logrando así que se pueda diseñar, crear prototipos, y lanzar entre 1 000 y 1 500 piezas nuevas todos los días.

El ciclo continúa y se retroalimenta con la respuesta en datos del mercado que consulta sus productos en la aplicación, así como de las reseñas y los videos que muchas *influencers* realizan para promocionar las prendas de temporada. Ha revolucionado a tal punto el mercado de la moda rápida (o *fast fashion*) que ahora se considera “moda de tiempo real’’ (*real-time fashion*) (Raído, 2022).

Ahora bien, este es solo un ejemplo en un sector sumamente competitivo como el de los textiles. Sin embargo, ocurre también en otros sectores como el de

entretenimiento: Netflix hizo historia con la primera serie “House of cards” realizada a partir de los resultados de datos masivos de sus usuarios. En las ventas de productos en general, Amazon predice lo que deberías comprar a partir de tus últimas compras o productos vistos. En la música, Spotify puede sugerirte artistas, grupos o tipos de música a partir de tus búsquedas o tus últimas reproducciones.

Estos ejemplos nos llevan a la cuestión que atañe al presente trabajo: ¿qué pasa en las bibliotecas? ¿Es posible aplicar alguna metodología de análisis de datos? ¿Cuáles son las áreas de aplicación para este campo? ¿Existen proyectos que actualmente lo estén realizando?

2.1.6.1 Ejemplos de aplicación de Big data en bibliotecas

Una de las aplicaciones que ha encontrado el gobierno de Singapur para los datos masivos, junto con algunas empresas de tecnología, es el de analizar el comportamiento de búsqueda de los usuarios de la biblioteca nacional de ese país (TechNews, 2018).

Al aprovechar 20 millones de registros anónimos de servicios que han utilizado los usuarios de la biblioteca, se busca encontrar patrones y conocerlos más “sin estereotipos”[2] dijo Mr. Liu Feng-Yuan, director de la división de Servicios de Gobierno Digital (TechNews, 2018).

Ahora bien, Voutssás (2022) menciona algunos otros ejemplos de la aplicación de metodologías de datos masivos a bibliotecas en distintos aspectos. Por ejemplo, en los catálogos, se habla de Worldcat, que operado por la organización OCLC y con 450 millones de registros catalográficos en 500 idiomas de casi 18 mil bibliotecas del mundo, ha utilizado estructuras de entidad-relación para crear

[2] Traducción al español hecha por la autora.

datos enlazados de sus existencias, con lo que identifica las entidades en sus registros y asigna relaciones entre ellas.

Otro ejemplo que menciona el autor es el de la biblioteca digital HathiTrust, en la que la colección es tan grande que se requieren 24 horas para recorrer los 5 mil millones de páginas de sus 14 millones de libros digitalizados, además de que existe 14 mil computadoras trabajando simultáneamente (Plale, 2016, como se cita en Voutssás, 2022).

De igual manera, Voutssás menciona otros ejemplos, sin embargo, algunos otros autores como Xu et al., mencionan que los bibliotecarios no están seguros de cómo integrar datos masivos en la biblioteca y que se necesita una comprensión más profunda al respecto, por lo que en ese momento lo visualizaban más como una posibilidad que podría servir a los investigadores o usuarios comunes (como se cita en Garoufallou y Gaitanou, 2021).

La IFLA (International Federation of Library Associations and Institutions) organiza una conferencia anual World Library and Information Congress (WLIC) y en el marco de llevada a cabo en 2022, en Dublín, se presentaron tres propuestas más de posibles aplicaciones de datos masivos.

La Universidad Estatal de Texas presentó su ecosistema de repositorio de datos de investigación para la investigación académica abierta. Se habló sobre una infraestructura de investigación de datos en red, en línea, que permite compartir y archivar datos de investigación para trabajos académicos abiertos. El ecosistema permitió un ciclo de investigación académica de principio a fin, desde la búsqueda y recuperación de datos y contenidos hasta la recopilación, el análisis, la redacción y la publicación en línea (IFLA, 2024).

Los conjuntos de datos de investigación que se encuentran en el repositorio se encuentran dentro del rango de 1 GB. Al reconocer la necesidad futura de

almacenar conjuntos de datos más grandes, han comenzado los preparativos para la próxima compilación.

La Biblioteca del Congreso de Estados Unidos presentó un caso mediante el cual se recopilaron y analizaron datos físicos, químicos y ópticos para tomar decisiones informadas sobre descarte o preservación de 500 libros idénticos, publicados entre 1840 y 1940, que se encontraban en las colecciones de cinco grandes bibliotecas de investigación en diferentes partes de EUA (IFLA, 2024).

Para realizar lo anterior, se requería un enfoque de análisis de datos con el fin de llenar vacíos de conocimiento y realizar una evaluación objetiva de las colecciones y así identificar materiales que estaban en riesgo y diferenciar entre copias de libros de buena y mala calidad.

Además, se configuró una plataforma para almacenar datos de muestreo recopilados por varios instrumentos, como la resistencia a la tracción y la acidez. Estos datos se almacenan en CouchDB, un depósito de documentos JSON, junto con datos no científicos de cada libro.

Para analizar dichos datos, el equipo disponía de una herramienta de consulta para evaluar los puntos de datos discretos en tiempo real, además de una herramienta de comparación que aprovechaba el visor de código abierto del Marco Internacional de Interoperabilidad de Imágenes (IIIF) para evaluar la documentación fotográfica entre libros en primer plano.

Los principales factores que se debieron considerar y evaluar para las conservaciones, fueron el impacto del material, el medioambiente y el uso. A través de la investigación, el equipo descubrió que las propiedades inherentes del papel cuando se produjeron los libros eran el factor más crítico para predecir la condición.

El último caso presentado durante el WLIC 2022, en Dublín, fue el proyecto de la Junta Nacional de Bibliotecas de Singapur (NLB) titulado Recomendaciones de libros a través del aprendizaje automático (IFLA, 2024). En este utilizaron Amazon Personalize, un servicio de recomendación de aprendizaje automático basado en la nube, y propiciaron recomendaciones personalizadas basadas en su colección de libros impresos y libros electrónicos, con lo que integraron las tecnologías de la nube, Big data y los datos que ya tenían en su sitio web y aplicación móvil.

Así, la NLB cuenta con un almacén que contiene datos transaccionales. El equipo pudo utilizar la infraestructura para crear una vista unificada de los usuarios y así comprender los patrones. Después se centraron en entrenar el modelo para hacer mejores recomendaciones, por ejemplo, reducir las recomendaciones de títulos duplicados de diferentes formatos y las recomendaciones basadas en temas.

El contar con la disponibilidad de servicios de recomendación administrados por terceros, permitió a las bibliotecas aprovechar sus recursos financieros eficientemente y rediseñar la experiencia del usuario al interactuar con las recomendaciones.

2.2. Minería de datos

2.2.1 Conceptos

Según Medina y Gómez (2014) la minería de datos se puede describir como

> Uno de los procesos de extracción de información dentro de los grandes volúmenes de datos (como lo es datos masivos), su origen es matemático, a través de algunos teoremas de eficacia de algoritmos y ecuaciones, así como de la estadística y un aprendizaje automático de los datos, permiten demostrar la utilidad de las cosas en las diferentes áreas de conocimiento y del saber, antes de la puesta de su funcionamiento (p. 32).

Dentro de los procesos para aplicar una metodología de datos masivos, se encuentra la minería de datos, de la que se extrae un poco de todos los datos para analizarlos.

Voutssás (2022) cita a Han al decir que

> La minería de datos es un subcampo interdisciplinario de la informática y de la estadística […] utiliza métodos del aprendizaje de máquinas, la estadística y los sistemas de bases de datos […] consiste en el proceso de descubrir patrones en grandes conjuntos de datos con métodos inteligentes con el propósito general de extraer información de esos conjuntos de datos y transformarla en estructuras comprensibles para su uso posterior (p. 19).

Para aplicar la minería de datos no se requieren volúmenes tan grandes de estos y sus herramientas pueden ser menos robustas, como sí se requiere para datos masivos.

Menciono una última definición de la minería de datos: “Es un conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas en información estructurada para su explotación directa o para su análisis y conversión en conocimiento” (Flores Lagla, et al., 2019, como se cita en Marcano Anular, 2019, p. 968).

Esta definición es interesante porque resalta que la minería de datos sistematiza los datos en información estructurada, así resalta lo que se pretende analizar y, por ende, visualizar.

2.2.2 Objetivos

La minería de datos se concibe como una forma innovadora de obtener información comercial valiosa a partir del análisis de los datos contenidos en la

base de datos de una empresa. Dicha información funciona para una adecuada toma de decisiones empresariales (IBM, 2022).

Entonces, la minería de datos es, en esencia, un método con el que se puede utilizar la información ya existente en una empresa con la finalidad de mejorar procesos, potencializar el rendimiento de la inversión u optimizar el uso de recursos (IBM, 2022).

Asimismo, con la minería de datos es posible develar información exhaustiva con técnicas avanzadas de análisis y creación de modelos. A través de la minería de datos, también se pueden “hacer consultas mucho más complejas de sus datos que utilizando métodos de consulta convencionales. La información que la minería proporciona puede mejorar notablemente la calidad y fiabilidad de la toma de decisiones empresariales” (IBM, 2022, párr. 2).

Como ejemplo están los métodos convencionales que tienen la capacidad para informar cuál es el tipo de cuenta más rentable de entre las que ofrece una institución bancaria. En cambio, la minería de datos facilita la información para que un banco cree perfiles de los clientes que ya disponen de ese tipo de cuenta. Luego, esta institución bancaria podría utilizar la minería de datos para encontrar otros clientes que coincidan con ese perfil, y así poder emprender una campaña comercial dirigida específicamente a esos usuarios (IBM, 2022).

De ese modo, "las herramientas de minería de datos facilitan y automatizan el proceso de descubrir esta clase de información en bases de datos de gran tamaño” (IBM, 2022, párr. 6).

2.2.3 Características

Flores Lagla et al. (2019) mencionan que algunos de los aspectos a tomar en cuenta para la minería de datos son:

- Verificación: donde el sistema es limitado para verificar la hipótesis del usuario.

- Descubrimiento: donde el sistema encuentra de forma autónoma nuevos patrones.
    - Predicción: al descubrir los patrones, este se subdivide en dónde el sistema encuentra dichos patrones para predecir el comportamiento futuro de algunas entidades.
- Descripción: donde el sistema encuentra patrones con el fin de presentarlos a un usuario en una forma humana comprensible (p. 964).

De igual manera, Bharati y Ramageri (2010) mencionan que la minería de datos involucra tres etapas importantes, las cuales son:

> ➢ Exploración: en este paso, los datos se limpian y se transforman en otra forma, también se determinan las variables importantes y, luego, la naturaleza de los datos en función del problema.
> ➢ Identificación del patrón: en este paso se trata de identificar el patrón o algoritmo más apropiado para el proyecto.
> ➢ Implementación: en este punto se analizan los patrones para inferir los resultados (p. 301).

Hablando de manera general, existen dos tipos de minería de datos: “el primero orientado a la verificación y el segundo se dirige hacia al descubrimiento. Los métodos de descubrimiento son aquellos que identifican automáticamente patrones en los datos, por medio de la predicción y la descripción” (Maimon y Rokach, 2010, p. 1)

Por otro lado, “los métodos descriptivos se orientan a la interpretación de datos, y se centran en comprender cómo los datos se relacionan entre sí. Los métodos orientados a la predicción se enfocan principalmente en la utilización de algoritmos que permitan estructurar los datos de una forma específica para poder predecir comportamientos, tendencias, etc." (Maimon y Rokach, 2010, p.1).

2.2.4 Técnicas de aplicación

“La minería de datos implementa varios algoritmos o técnicas como clasificación, agrupación, regresión, redes neuronales, reglas de asociación, árboles de decisión, algoritmos genéticos, etc., para el descubrimiento de conocimiento a partir del análisis a bases de datos" (Bharati y Ramageri, 2010, p. 302).

Ahora, se describirán algunos algoritmos o técnicas que utiliza la minería de datos:

A) Clasificación: es la técnica de minería de datos más utilizada y permite clasificar un conjunto de datos de acuerdo con etiquetas o patrones asignados dentro de los datos, mediante árboles de decisión, redes neuronales, redes bayesianas, etc. (Bharati y Ramageri, 2010; Unidad de Información y Análisis Financiero, 2014).

> ➢ Redes neuronales: también se denominan aprendizaje conexionista, ya que crea conexiones entre los datos […] el algoritmo de red neuronal más popular es la retro-propagación (Bharati y Ramageri, 2010, p. 302)

“Las redes neuronales son mejores para identificar patrones o tendencias en los datos y son muy adecuadas para las necesidades de predicción o pronóstico” (Bharati y Ramageri, 2010, p. 302).

> ➢ Árboles de decisión: estos proveen de una herramienta de clasificación muy potente, que permite representar y analizar los datos para generar una predicción con base en el comportamiento del conjunto de datos (Bharati y Ramageri, 2010, p. 302).

> ➢ Redes Bayesianas: “las redes bayesianas proveen una forma compacta de representar el conocimiento y métodos flexibles de razonamiento basados en teorías probabilísticas (teorema de Bayes) capaces de predecir el valor de variables no observadas y explicar las observadas” (Bonilla Gordillo y Ojeda Schuldt, 2006, p. 7).

B) Agrupamiento (clustering): es un algoritmo que agrupa datos en categorías considerando sus similitudes y minimizando sus diferencias. Con el uso de esta técnica se pueden identificar zonas dispersas entre datos u objetos y patrones de distribución, así mismo las correlaciones entre los atributos de los datos (Bharati y Ramageri, 2010, p. 302)

C) Regresión: la técnica de regresión se basa en el aprendizaje supervisado y se utiliza para predecir un dato continuo o numérico. De igual manera, el algoritmo puede predecir ventas, ganancias, tasas hipotecarias, pies cuadrados, temperaturas, etc. (Gera y Goel, como se cita en Martínez Acevedo y Polo Bautista, 2021, p. 35).

Ilustración 8: Algoritmos de minería de datos

Minería de datos			
Verificación	***Descubrimiento***		
(Lenguaje Estructurado de Consultas) SQL	***Descripción***	***Predicción***	
Procesamiento Analítico en Línea (OLAP)	Visualización	*Clasificación*	*Regresión*
Análisis estadístico	Agrupamiento	Árboles de decisión	Árboles de regresión
	Reglas de asociación	AANNs	Árboles de modelos
	Descubrimiento de subgrupos (Subgroup Discovery)	Redes bayesianas	AANNs
	Redes de creencia	IBL	

Fuente: Escobar Terán, Alcivar y Puris, 2016, como se cita en Martínez Acevedo y Polo Bautista, 2021, p. 35.

Ahora bien, a partir de lo presentado se puede concluir que tanto datos masivos como minería de datos procesan una gran cantidad de datos para convertirlos en información y conocimiento de valor para las organizaciones.

Sin embargo, podemos concluir que datos masivos representa el procedimiento y la minería de datos, la herramienta con la que se lleva a cabo el proceso analítico. Ambos se complementan cuando se aplican sus técnicas de análisis.

2.2.5 Ejemplos de aplicación

Algunos bibliotecarios de bibliotecas universitarias han adquirido habilidades para el manejo de datos y comparten esos conocimientos con los docentes e investigadores de su institución, con el fin de potenciar sus proyectos, pero también para maximizar el aprovechamiento de la información recabada durante la investigación.

Por ejemplo, la biblioteca Bodleiana de la Universidad de Oxford (Reino Unido), la segunda biblioteca más importante de ese país ofrece servicios relacionados

a encontrar y usar datos, el acompañamiento de bibliotecarios especializados en esta línea y brinda capacitación a investigadores y estudiantes de esa universidad en el uso de datos.

Además, cuentan con una guía en su sitio web (Data Mining, 2022), en el que hacen un listado de las herramientas que pueden utilizarse para la minería de datos. Las herramientas enlistadas son de acceso abierto y en la guía se mencionan, además, recursos para aprender a utilizar esos programas.

Las bibliotecas de la Universidad de Harvard cuentan con un programa permanente denominado Research Data Management Program (Programa de Gestión de Datos de Investigación) (Harvard Library, 2022a). La misión del programa es proveer servicios y recursos alrededor de la gestión de datos y apoyar a los proyectos multidisciplinarios de los investigadores y estudiantes de la universidad.

Ofrecen capacitación, curaduría de datos, gestión de datos, servicios de asesorías, pautas o lineamientos y tecnología. También cuentan con diversos diplomados, talleres y cursos constantes, además de su propio repositorio Harvard Dataverse, que ofrece “una amplia variedad de conjuntos de datos para respaldar su investigación. Ofrece una búsqueda avanzada y minería de texto en más de 2 000 universos de datos, 75 mil conjuntos de datos y más de 350 mil archivos que representan instituciones, grupos e individuos en Harvard y más allá” (Harvard Library, 2022b).

2.3 Sistemas expertos

Hemos hablado ya de dos metodologías de análisis de datos, pero debe existir alguna forma de procesarlos una vez obtenidos y estudiados, para que se puedan tomar decisiones de manera eficaz y útil.

Los sistemas expertos son una respuesta ante esta pregunta como subconjunto de la inteligencia artificial (Rossini, 2000, como se cita en Badaró, Ibañez y Agüero, 2013, p. 351), y podrían aportar mucho a la bibliotecología en una era donde no basta con solo conocer el contexto a través de los datos, sino de qué manera se brindan soluciones de acuerdo con las necesidades de los usuarios.

2.3.1 Conceptos

Badaró (2013) explica que, en realidad, el nombre completo es “Sistema experto basado en conocimiento y que es un sistema que emplea conocimiento humano capturado en una computadora, para resolver problemas que normalmente requieren de expertos humanos” (p. 351).

Calva (2023) menciona que es un producto de la investigación en inteligencia artificial que emula el conocimiento de un experto humano en la solución de problemas (p. 8).

Podría decirse que es un sistema informático capaz de almacenar la experiencia, el conocimiento, las habilidades propias de una o más personas especializadas en un área particular del conocimiento humano, y de resolver problemas específicos de esa área mediante la deducción lógica de conclusiones. Los sistemas expertos no son sistemas informáticos convencionales, ya que no están conformados por dos partes diferenciadas, datos o instrucciones (Davara, 1994, como se cita en Calva, 2023, p. 8).

Voutssás (2022) define los sistemas expertos como programas informáticos que trabajan principios y métodos de la inteligencia artificial con los que resuelven problemas dentro de un campo especializado que usualmente requeriría de la experiencia de personal experto. Además, menciona que “incorporan los conocimientos técnicos acumulados por las personas expertas en un tema y se diseñan para funcionar lo más parecido a ellas” (p. 174). Explica, además, que cuentan con una base de conocimientos de hechos y relaciones representados

en forma de datos y tienen la capacidad de hacer inferencias basadas en ellos (p. 174).

2.3.2 Antecedentes

Para comenzar a hablar de los antecedentes, propongo una cita que Badaró, Ibañez y Agüero (2013) retoman de Turban (1995):

> Los sistemas expertos fueron desarrollados por la comunidad de inteligencia artificial [en lo subsecuente IA] a mediados de los años 60. En este periodo de investigación de IA se creía que algunas reglas de razonamiento sumadas a las poderosas computadoras podían producir un experto o rendimiento superhumano. Un intento en esta dirección fue el General Purpose Problem Solver (GPS, solucionador de problemas de propósito general) (p. 352).

El General purpose Problem Solver (Newell, 1958, como se cita en Badaró, Ibañez y Agüero, 2013, p. 352) o GPS fue un precursor de los sistemas expertos (SE en lo subsecuente).

Al igual que otros programas similares, el GPS no cumplió con las expectativas de sus creadores, pero dejo importantes beneficios. Con el desarrollo de DENDRAL, Badaró, Ibañez y Agüero (2013) menciona que se llegaron a algunas conclusiones (p. 353):

> -La complejidad de los problemas requiere una cantidad considerable de conocimiento sobre el área del problema.
> -Los solucionadores de problemas generales eran muy débiles para ser utilizados como base para construir SE de alto rendimiento.
> -Los expertos humanos son buenos solo cuando actúan en un dominio muy acotado.
> -Los SE necesitan ser actualizados constantemente con nueva información.

Algunos sistemas expertos, que fueron clave para el desarrollo de este campo, son los citados a continuación (Badaró, Ibañez y Agüero, 2013, p. 353):

DENDRAL: “primer SE en ser utilizado para propósitos reales, al margen de la investigación computacional. Durante aproximadamente 10 años, el sistema tuvo cierto éxito entre químicos y biólogos, ya que facilitaba enormemente la inferencia de estructuras moleculares” (Turban, 1995, como se cita en Badaró, Ibañez y Agüero, 2013, p. 353).

MYCIN:

> Es un SE para la realización de diagnósticos, iniciado por Ed Feigenbaum y posteriormente desarrollado por E. Shortliffe. Su función es la de aconsejar a los médicos en la investigación y determinación de diagnósticos en el campo de las enfermedades infecciosas de la sangre (Nebendahl, 1991, como se cita en Badaró, Ibañez y Agüero, 2013, p. 353).

CADUCEUS:

> Fue un SE médico programado para realizar diagnósticos en medicina interna. Fue completado a mediados de la década de 1980. Si bien el inicio de su desarrollo se remonta a la década de 1970, fue programado por Harry Pople, de la Universidad de Pittsburgh y tomó como punto de partida una serie de entrevistas de Pople al doctor Jack Meyers. Pretendía mejorar el MYCIN, sistema focalizado sobre las bacterias infecciosas de la sangre (Nebendahl, 1991, como se cita en Badaró, Ibañez y Agüero, 2013, p. 353).

XCON:

> El programa R1 (luego llamado XCON, por Configurador Experto) era un sistema de producción basado en reglas escrito en OPS5 por John P. McDermott de CMU (1978) con el propósito de asistir los pedidos de los sistemas de computadores VAX de DEC (Digital Equipment Corporation) al seleccionar los componentes del sistema de acuerdo con los requerimientos del cliente. El desarrollo de XCON siguió a dos fracasos de escribir un sistema

experto para esta tarea en FORTRAN y BASIC (Nebendahl, 1991, como se cita en Badaró, Ibañez y Agüero, 2013, p. 354).

2.3.3 Objetivos

Uno de los objetivos que tiene un sistema experto “es ayudar a encontrar la solución óptima a un problema concreto sin tener que recurrir a un experto humano en la materia’’ (DIGIXEM360, 2023).

Asimismo, el sistema experto puede llevar a cabo una consulta o acción, “incluso con datos incompletos. Trabaja con tipos de datos cualitativos más que cuantitativos y utiliza la llamada lógica difusa, es decir, un razonamiento “aproximado” que conduce a resultados altamente probables’’ (DIGIXEM360, 2023).

Podemos decir que el objetivo de estos desarrollos informáticos es el de solucionar problemas a través de un conjunto de datos, información variada, “análisis de protocolos y procedimientos escritos, la descripción verbal de tareas realizadas por una persona, los cuestionarios, encuestas, entrevistas, observación de procesos y simulación’’ (Voutssás, 2023).

2.3.4 Características

Badaró, Ibañez y Agüero (2013) mencionan que “los sistemas expertos están compuestos por dos partes principales: el ambiente de desarrollo y el ambiente de consulta. El ambiente de desarrollo es utilizado por el constructor para crear los componentes e introducir conocimiento en la base de conocimiento” (p. 354).

Por otro lado, los siguientes son los componentes básicos de un SE (Badaró, 2013, p. 354):

- Base de conocimiento

Contiene el conocimiento necesario para comprender, formular y resolver problemas. Incluye dos elementos básicos: heurística especial y reglas que dirigen el uso del conocimiento para resolver problemas específicos en un dominio particular.

- Base de hechos

Es una memoria de trabajo que contiene los hechos sobre un problema y alberga los datos propios correspondientes a los problemas que se desean tratar.

- Motor de inferencia

Es el cerebro del SE, también conocido como estructura de control o interpretador de reglas. Este componente es esencialmente un programa de computadora que provee metodologías para razonamiento de información en la base de conocimiento. Este componente provee direcciones sobre cómo usar el conocimiento del sistema para armar la agenda que organiza y controla los pasos para resolver el problema cuando se realiza una consulta.

El SE tiene tres elementos principales (Turban, 1995, como se cita en Badaró, Ibañez y Agüero, 2013):

> (1) Intérprete: ejecuta la agenda seleccionada;
> (2) programador: mantiene el control sobre la agenda;
> (3) control de consistencia: intenta mantener una representación consistente de las soluciones encontradas (p. 355).

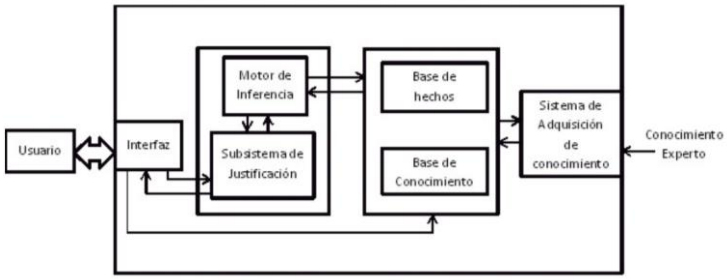
Subsistema de justificación

Se encarga de explicar el comportamiento del SE al encontrar una solución. Permite al usuario hacer preguntas al sistema para poder

> entender las líneas de razonamiento que este siguió. Resulta especialmente beneficioso para usuarios no expertos que buscan aprender a realizar algún tipo de tarea (Turban, 1995, como se cita en Badaró, Ibañez y Agüero, 2013, p. 355).

Estos elementos pueden identificarse en el siguiente esquema:

Ilustración 9 Estructura de un sistema experto



Fuente: Badaró, Ibañez y Agüero, 2013, p. 355.

3.2. Tipos de sistemas

De acuerdo con Badaró, Ibañez y Agüero (2013), existen diversos tipos de sistemas expertos, algunos de ellos son:

- Basados en reglas

“Los sistemas basados en reglas trabajan mediante la aplicación de reglas, comparación de resultados y aplicación de las nuevas reglas basadas en situación modificada” (p. 355).

Asimismo, se usan a partir de “inferencia lógica dirigida. Inician con una evidencia inicial en una determinada situación y se dirigen hacia la obtención de una solución. También pueden trabajar con una hipótesis sobre las posibles soluciones y volviendo hacia atrás para hallar evidencias ya existentes (o una deducción de una evidencia existente) que apoya una hipótesis en particular” (p. 355).

“Los sistemas que incluyen múltiples tipos de conocimiento a veces se conocen como sistemas híbridos, o etiquetados, después de un determinado tipo de representación del conocimiento, por ejemplo, basado en casos” (O’Leary, 2008, como se cita en Badaró, Ibañez y Agüero, 2013, p. 356).

- Basados en casos

El razonamiento basado en casos es el proceso de solucionar nuevos problemas apoyándose en las soluciones de anteriores. Badaró, Ibañez y Agüero (2013, p. 356) mencionan como ejemplo el caso de un mecánico de automóviles que repara un motor porque recordó que otro auto presentaba los mismos problemas. A ese tipo de razonamiento se le llama basado en casos.

El razonamiento basado en casos es una manera de razonar haciendo analogías y se ha sostenido que todo razonamiento es basado en casos porque está fundamentado en la experiencia previa

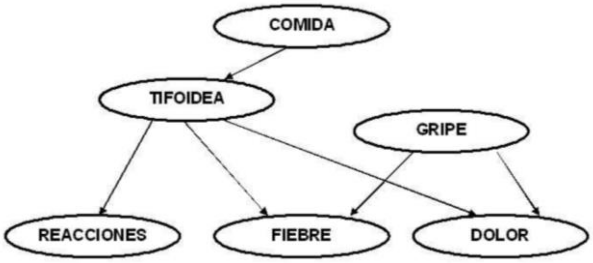
- Basados en redes bayesianas

Una red bayesiana, red de bayes, red de creencia, modelo bayesiano o modelo probabilístico en un gráfico acíclico dirigido, es un modelo gráfico probabilístico (un tipo de modelo estático) que representa un conjunto de variables aleatorias y sus dependencias condicionales a través de un gráfico acíclico dirigido (DAG, por sus siglas en inglés). Por ejemplo, una red bayesiana puede representar las relaciones probabilísticas entre enfermedades y síntomas. Dados los síntomas, la red puede ser usada para computar las probabilidades de la presencia de varias enfermedades.

Por ejemplo, Sucar (s.f.) explica un ejemplo hipotético “de una red bayesiana en el que los nodos representan enfermedades, síntomas y factores que causan enfermedades y la topología o la estructura de la red (figura 9), brinda información sobre las dependencias probabilísticas entre las variables. También

representa las independencias condicionales de una variable o conjunto de variables dadas a otras variables.''

Ilustración 10 Ejemplo de una red bayesiana



Fuente: Sucar, s.f., p.3.

- **Sistemas expertos difusos**

Acorde con Sucar (s.f.), los sistemas expertos difusos usan el método de lógica difusa para desarrollarse, esta lógica trabaja con incertidumbre. Para que se lleve a cabo la técnica, se emplea el modelo matemático de conjuntos difusos; esta simula el proceso del razonamiento normal humano, lo que permite a la computadora ser menos precisa y más lógica que las computadoras convencionales. Este proceso se usa así porque la toma de decisiones no es siempre una cuestión de blanco y negro, verdadero o falso; a veces involucra áreas grises y el término "quizás" (Holland, 1992, como se cita en Badaró, Ibañez y Agüero, 2013, p. 356).

2.3.5 Ejemplos de aplicación

Badaró, Ibañez y Agüero (2013) mencionan que "los sistemas expertos se han venido aplicando con éxito en múltiples campos: medicina, geología, química, ingeniería, etc., para realizar tareas muy diversas, como interpretación, predicción, diagnóstico, diseño, planificación, instrucción, control" (p. 360). A continuación, se mencionan algunos ejemplos como: auditoría, contabilidad de costos y de gestión, contabilidad financiera, análisis de estados financieros, planificación financiera e industria de los servicios financieros.

Sin embargo, se sabe que también puede ser aplicado a bibliotecas, dado que como menciona Voutssás, mucho del conocimiento de los bibliotecarios acerca de la gestión y explotación de información es aquel generado por la experiencia, interiorizado por diferentes procesos.

Asemi, Ko y Nowkarizi (2021, p. 2) mencionan ejemplos de la aplicación de sistemas expertos aplicados a bibliotecas:

1. Indexación basada en el conocimiento
2. Procesamiento y resúmenes del lenguaje natural
3. Servicios de referencia
4. Catalogación
5. Recuperación de información en línea
6. Utilizar interfaces inteligentes en sistemas de almacenamiento y recuperación de información en línea
7. Análisis y representación de las necesidades de información, incluidos diferentes servicios, como clasificación, indexación y resúmenes
8. Desarrollo de colecciones
9. Hipertexto e hipermedia

Njuku (2022, p. 5) menciona que en la Universidad Northwestern desarrollaron varios sistemas expertos para catalogación. Por ejemplo, DELICAT (Data Enhancement of Library Catalogues u Optimización de los Datos de Catálogos de Bibliotecas) es capaz de detectar automáticamente errores en los catálogos de la biblioteca y notificar a los bibliotecarios.

IESCA (The Interactive Electronic Serials Cataloging Aides o asistente interactivo de catalogación de publicaciones electrónicas seriadas) es un tutorial que ayuda a las bibliotecas a catalogar documentos electrónicos. Guía a los usuarios a través del proceso e incluye la mayoría de las reglas y estándares aplicables para la catalogación (Njuku, 2022, p. 5).

Njuku (2022, p. 6) también menciona que la clasificación es una función difícil de realizar mediante sistemas expertos, dado que no existen reglas estrictas para capturar las temáticas de los contenidos (y que, además, pueden ser ambiguas). En 1986, Paul Burton realizó una investigación exploratoria en la Universidad de Strathclyde, en el Reino Unido, con el objetivo de evaluar los méritos de diversas formas de representación del conocimiento y evaluar la idoneidad de los sistemas expertos en clasificación.

Dicha investigación dio como resultado un prototipo de sistema experto que fue capaz de sugerir un número de clasificación del sistema Dewey, con base en la información proporcionada por el usuario. Después de la investigación, OCLC desarrolló un asistente de catalogación y esto se probó en la Universidad Carnegie Mellon, con la que se reclasificó la colección de Matemáticas e Informática.

La biblioteca de la Universidad Tecnológica de Indonesia basó su catálogo digital en sistemas expertos, al utilizar el lenguaje de programación PHP y las bases de datos en MYSQ. Además, desarrolló el método de encadenamiento hacia adelante y de encadenamiento hacia atrás como motor de inferencia (Njuku, 2022, p. 2). Al tener su biblioteca basado en SE, ha agilizado el servicio y facilitado la búsqueda entre las colecciones de libros, revistas, periódicos a los usuarios, sin tener que acudir físicamente al campus.

Asimismo, Voutssás (2023, p. 68) menciona otras posibles aplicaciones de los SE en bibliotecas:

- Muchas bibliotecas estudian el lenguaje natural de los usuarios, con el fin de enseñar a las computadoras a entender ese lenguaje y empatarlo con los lenguajes controlados que se utilizan en bibliotecología. Por lo que se podría utilizar en: la extracción de información de textos, recuperación de

información en bases de datos y catálogos, traducción automática, reconocimiento y síntesis del habla, etc.

- La mayoría de las bibliotecas conserva información de las búsquedas previas de los usuarios para, de ese modo, personalizar la página de cada uno de ellos, al recordar lo que han buscado estableciendo patrones. Así, el sistema posteriormente puede hacer sugerencias como “las personas que consultaron este texto también consultaron estos otros” o “este autor se relaciona con este otro” o “este tema se relaciona con este otro”.
- Ciertas bibliotecas sacan datos de las redes sociales de sus usuarios, las cuales están conectadas a los servicios de la biblioteca, para recibir sugerencias de adquisición de obras, detectar “temas de tendencia”, contar “me gusta”, verificar eficacia y dar seguimiento de sus servicios.

Capítulo 3. Metodología para la aplicación de Big data en bibliotecas académicas

En los capítulos anteriores se ha mencionado que el OPAC de una biblioteca académica es una parte integral de su infraestructura de información, ya que proporciona a los usuarios un medio para buscar y acceder a los recursos bibliográficos disponibles.

Aunque el OPAC tradicionalmente se ha centrado en proporcionar acceso a los catálogos de la biblioteca y a los recursos físicos y electrónicos, también puede integrarse en un modelo de Big data de varias maneras:

- **Recopilación de datos de OPAC:** el OPAC genera una gran cantidad de datos sobre las búsquedas de los usuarios, los elementos consultados, las consultas realizadas, las interacciones con los recursos, etc. Estos datos pueden ser recopilados y almacenados como parte de un modelo de Big data para su posterior análisis.
- **Análisis de uso y comportamiento del usuario:** los datos recopilados del OPAC pueden ser analizados para comprender mejor el comportamiento y las necesidades de los usuarios. Esto puede incluir patrones de búsqueda, tendencias de uso de recursos, preferencias de formato, áreas temáticas de interés, etcétera.
- **Personalización de servicios:** el análisis de datos del OPAC puede utilizarse para personalizar los servicios y las recomendaciones para los usuarios. Por ejemplo, los sistemas de recomendación basados en el análisis de datos pueden sugerir recursos relevantes o servicios adicionales apoyados en los patrones de búsqueda y el historial de uso de un usuario específico.
- **Optimización de la colección:** los datos del OPAC pueden proporcionar información valiosa para la gestión de la colección de la biblioteca. El análisis de datos puede ayudar a identificar áreas de interés para la

adquisición de nuevos materiales, la eliminación de materiales obsoletos o poco utilizados, y la optimización de la disponibilidad de recursos.

- **Mejora de la experiencia del usuario:** la comprensión del comportamiento del usuario a través de los datos del OPAC puede ayudar a mejorar la experiencia del usuario al diseñar interfaces más intuitivas, proporcionar mejores recomendaciones de búsqueda, optimizar los tiempos de respuesta y mejorar la accesibilidad de los recursos.

El OPAC de una biblioteca académica puede ser una fuente importante de datos que puede integrarse y aprovecharse en un modelo de Big Data para mejorar la gestión de la biblioteca, personalizar los servicios para los usuarios y optimizar la experiencia del usuario en general.

Pero para poder aprovechar al máximo lo que ofrece el OPAC, se requiere un modelo que nos permita definir los datos, de dónde se tomarán esos datos y cómo serán procesados.

3.1 Metodología

Como ya se ha abordado en otros apartados, en la era digital, las bibliotecas académicas enfrentan el desafío de gestionar vastas cantidades de información de manera eficiente. Es por esto por lo que un modelo de *Big Data* para bibliotecas académicas se presenta como una solución esencial para optimizar la gestión y el análisis de datos. Este modelo permite integrar diversas fuentes de información, como catálogos de libros, bases de datos académicas y patrones de uso de los usuarios, ofreciendo una visión holística y profunda de los recursos y servicios bibliotecarios.

La implementación de *Big Data* en las bibliotecas académicas no solo mejora la administración interna, sino que también enriquece la experiencia de los usuarios al proporcionarles acceso más rápido y personalizado a la información que necesitan.

La idea de construir un modelo de *Big Data* para bibliotecas académicas surge de la necesidad de adaptarse a un entorno académico cada vez más interconectado y centrado en los datos. Con el incremento exponencial de publicaciones y recursos electrónicos, las bibliotecas se ven limitadas en su capacidad para procesar y organizar la información utilizando los métodos tradicionales. Inspirados en los avances tecnológicos en otras áreas, los bibliotecarios comenzaron a explorar cómo las técnicas de *Big Data* podrían aplicarse para superar estas limitaciones. El resultado es un modelo que utiliza análisis avanzados para detectar tendencias, mejorar la toma de decisiones y personalizar los servicios bibliotecarios.

La utilidad de un modelo de *Big Data* en bibliotecas académicas es múltiple. En primer lugar, permite una gestión más eficiente de los recursos, al optimizar la adquisición y el mantenimiento de colecciones. Además, facilita la identificación de patrones de uso y preferencias de los usuarios, lo que ayuda a personalizar los servicios y mejorar la satisfacción del usuario.

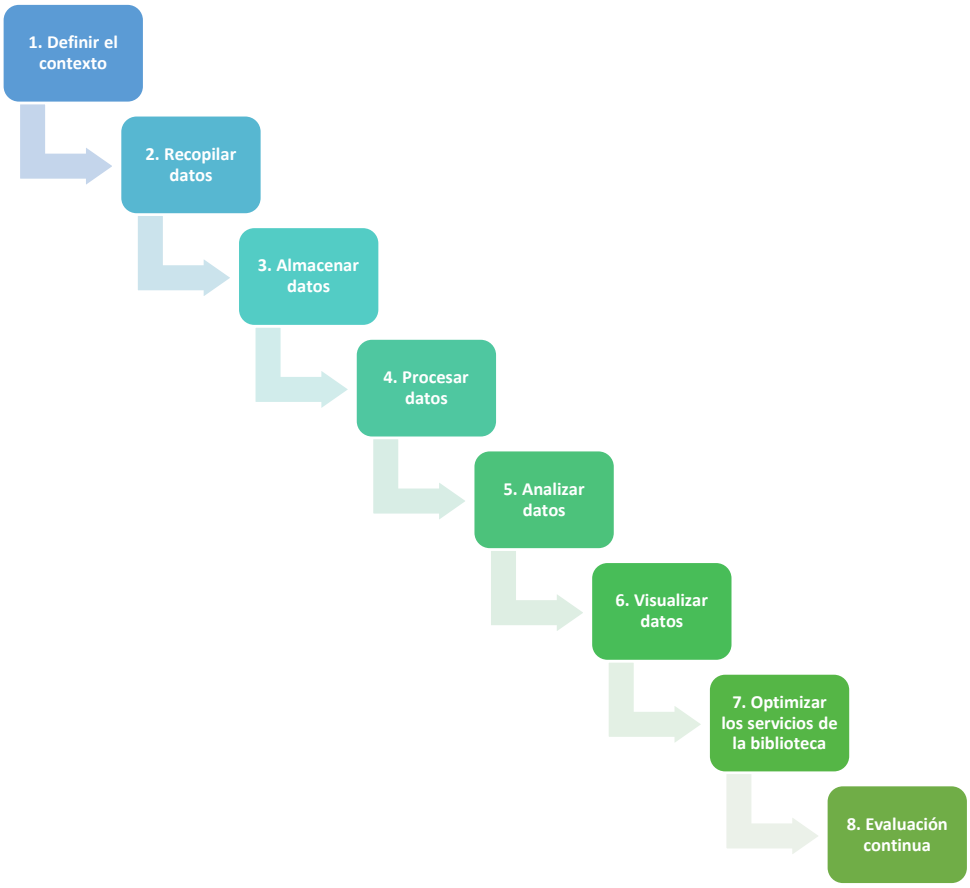
Finalmente, este modelo también apoya la investigación académica, pues ofrece herramientas avanzadas para el análisis de datos bibliográficos y facilita la colaboración interdisciplinaria. En resumen, la propuesta de un modelo de *Big Data* para bibliotecas académicas no solo es relevante, sino que es crucial para mantener la competitividad y relevancia de estas instituciones en el siglo XXI.

Ahora bien, es importante precisar que el modelo de *Big Data* propuesto en este trabajo se concibe como un modelo conceptual, ya que busca estructurar y describir los elementos clave involucrados en la recuperación y análisis de datos provenientes del OPAC en bibliotecas académicas. No se trata de un modelo teórico en sentido estricto, pues no pretende desarrollar una nueva teoría, ni de un modelo sistemático, ya que no presenta un esquema formal de implementación con procesos detallados. Sin embargo, también puede entenderse como un proceso, dado que establece una serie de pasos lógicos para la extracción y análisis de datos bibliográficos, los cuales pueden aplicarse

en distintas bibliotecas académicas con adaptaciones según su infraestructura y necesidades.

A continuación, se presenta una propuesta de modelo que puede ser utilizada como punto de partida para aplicar metodologías de datos masivos a una biblioteca académica.

Ilustración 11 Modelo de Big data para bibliotecas académicas



Fuente: Elaboración propia.

1. **Definir el contexto**

Para llevar a cabo cualquier actividad, es necesario definir los qué, cómo y para qué. Como se mencionó al inicio de este apartado, se puede aplicar metodologías de datos masivos a distintas áreas de la biblioteca partiendo del catálogo público, pero definir cuál es el aspecto que se quiere estudiar es vital para trazar un plan.

Algunas cuestiones que pueden ayudar como guía son:

- ¿Cuál es el problema que se pretende resolver?
- Plantear objetivos
- Desarrollar preguntas concretas que se quieran resolver

2. **Recopilar datos**

La biblioteca recopila datos a través del catálogo público y cómo los usuarios lo consultan. Sin embargo, como se mencionó en el apartado 1.2.1 Sistemas Integrales de Automatización de Bibliotecas, en la figura 1, en la que podemos ver la estructura de los SIAB en cuánto a las bases de datos y cómo se relacionan, es importante definir, a partir del paso 1 “definir el contexto”, qué datos se utilizarán para la consulta, dado que la información se almacena en distintas bases de datos.

¿Qué datos se utilizarán? De los datos proporcionados del OPAC, pueden ser los registros de préstamos, los temas consultados, los registros de uso de recursos electrónicos, el tiempo de permanencia en el OPAC, entre otros.

Por otro lado, hay datos que también podrían utilizarse fuera del OPAC, pero dentro de los que recaba la biblioteca: las encuestas de satisfacción de usuarios, datos de acceso a la biblioteca física, etcétera.

Una vez que se sabe qué datos se utilizarán, se recopilarán los datos que correspondan a esas preguntas.

3. **Almacenar datos**

Los datos recopilados se almacenan en un repositorio centralizado. Dependiendo de la escala y los requisitos de rendimiento, este podría ser un

sistema de almacenamiento de datos tradicional o una solución de Big data como Hadoop, almacenamiento en la nube o una combinación de ambos.

También deberá decidirse el formato en el que se almacenarán los datos, esto dependerá de su volumen y de la naturaleza (por ejemplo, si se están obteniendo en tiempo real).

Algunos ejemplos de formatos pueden ser los siguientes:

- **csv (Comma-Separated Values):** los archivos CSV son simples y fáciles de usar, lo que los hace populares para el intercambio de datos entre aplicaciones y sistemas. Sin embargo, pueden no ser la opción más eficiente para grandes volúmenes de datos debido a su estructura basada en texto plano y su falta de compresión (Microsoft, s.f.). Son utilizados comúnmente con Excel u programas de hojas de cálculo.

- **JSON (JavaScript Object Notation):** es un formato de datos liviano y fácil de leer que es ampliamente utilizado en aplicaciones web y servicios API. Aunque no es tan eficiente en términos de almacenamiento como los formatos de columna, sigue siendo una opción popular para el intercambio de datos entre sistemas heterogéneos (IBM, 2024a).

- **Apache Avro:** es un formato de serialización de datos que proporciona un esquema compacto y eficiente para el almacenamiento y la transmisión de datos. Es compatible con la evolución de esquemas y se utiliza en entornos donde la flexibilidad y la interoperabilidad son importantes (IBM, 2024b). Se vincula su uso principalmente para Apache Hadoop.

- **Apache Parquet**: es un formato de archivo de columna abierto y eficiente en términos de almacenamiento, diseñado para procesamiento de datos

en clústeres. Es especialmente adecuado para conjuntos de datos con un gran número de columnas (Datos.gob.es, 2021).

4. **Procesar datos**

Los datos se procesan para su limpieza, transformación y enriquecimiento. Por ejemplo, en la limpieza se incluye lo siguiente.

- Anonimización de los datos. Se elimina cualquier dato personal que pueda identificar a los usuarios, por ejemplo. Se implementan medidas de seguridad para proteger la privacidad y confidencialidad de los datos de los usuarios y garantizar el cumplimiento de las regulaciones de protección de datos.
- Eliminación de datos irrelevantes, incompletos o duplicados.
- Eliminación de las búsquedas realizadas por usuarios que no son el objetivo. Por ejemplo, las búsquedas que realizan los bibliotecarios a modo de “prueba”.
- Corrección de errores.

Para la transformación, se incluye la normalización de datos y para el enriquecimiento, la asignación de metadatos y la agregación de datos de múltiples fuentes.

5. **Analizar datos**

Se aplican técnicas de análisis de Big data para extraer información útil de los datos, como el modelado, la identificación de patrones y análisis descriptivos. Por ejemplo, patrones de uso de la biblioteca, análisis de tendencias de préstamos, análisis de popularidad de recursos, análisis de patrones de búsqueda, términos más utilizados al momento de la búsqueda, etcétera.

También pueden aplicarse técnicas avanzadas como el análisis predictivo para predecir tendencias futuras en la demanda de recursos o servicios.

6. **Visualizar datos**

Los resultados del análisis se presentan de manera visualmente atractiva y fácil de entender a través de cuadros de mando interactivos, gráficos, tablas y otros formatos visuales. Esto permite a los administradores de la biblioteca y a los bibliotecarios tomar decisiones informadas basadas en los datos.

Algunos ejemplos de software que puede ser utilizado para la visualización de los datos, son Metabase, Google Analytics, Google Data Studio, Tableau, Power BI, Plotly, D3.js, QlikView y Qlik Sense.

La elección del software de visualización dependerá del volumen de los datos, de la variedad de los datos y de qué se pretende resaltar en los resultados obtenidos.

7. **Optimizar los servicios de la biblioteca o toma de decisiones**

A partir de las perspectivas obtenidas a través del análisis de datos, se utilizarán para optimizar los servicios y recursos de la biblioteca.

Esto puede incluir la asignación eficiente de recursos, la personalización de servicios según las necesidades de los usuarios, la identificación de áreas de mejora en la oferta de servicios, etcétera.

8. **Evaluar el proceso y los resultados**

El modelo se somete a una evaluación continua para identificar áreas de mejora y adaptarse a medida que cambian las necesidades y tecnologías de la biblioteca y los usuarios.

3.3 Ejemplo de aplicación

Con el fin de evaluar el modelo propuesto previamente, se realizó un pequeño ejercicio utilizando datos de la biblioteca académica Gregorio Torres Quintero de la Universidad Pedagógica Nacional. Para ello, se solicitó autorización a la

directora de biblioteca de ese momento, la maestra Rosenda Ruiz Figueroa. Este ejercicio se realizó en marzo de 2023.

1. **Definir el contexto**

La biblioteca Gregorio Torres Quintero es una biblioteca académica en Ciudad de México, especializada en temas relacionados con la educación: psicología educativa, pedagogía, sociología de la educación, educación básica, etc., que da servicio a investigadores, docentes y estudiantes de licenciatura, maestría y doctorado.

Esta biblioteca tiene poco tiempo (apenas más de dos años) desde que migró a Koha, después de un complejo proceso de cambio.

- ¿Cuál es el problema que se pretende resolver? Conocer a los usuarios de la biblioteca, a partir de sus interacciones con el catálogo
- Los objetivos son búsquedas de los usuarios, términos empleados, cuántos usuarios utilizan el OPAC.
- Preguntas
    - ¿Cómo buscan los usuarios en el OPAC?
    - ¿Cuáles son los términos que utilizan?

2. **Recopilar datos**

Para este ejercicio, el Departamento de Informática de la Biblioteca Gregorio Torres Quintero sugirió realizar un clon del servidor que almacenaba los datos del SIAB Koha, con el fin de mantener los datos del servidor principal protegidos.

A partir de este clon, se exploraron las bases de datos para localizar aquellas que tenían datos referentes al OPAC.

3. **Almacenar datos**

Una vez identificados los datos, se utilizó el formato CSV para almacenarlos.

4. **Procesar datos**

En el caso de los datos obtenidos de la biblioteca Gregorio Torres Quintero, los datos en formato CSV se procesaron para su limpieza, transformación y enriquecimiento.

5. **Analizar datos**

Para este ejercicio, se seleccionaron como herramientas Metabase, Google Analytics y Orex.

Metabase, 2023, es una herramienta que se vincula con las bases de datos y ofrece al instante, las visualizaciones de ellos. En su página web se menciona que su interfaz es tan intuitiva que es una herramienta para todos, aunque no seas programador o desarrollador de sistemas. Algunas de sus herramientas son de acceso abierto y las funciones más avanzadas son de pago.

Te permite generar tus propios filtros para personalizar tus consultas, ofrece más de 15 opciones de gráficos o elementos visuales que permiten una mejor visualización. En caso de que conozcas SQL, te permite generar plantillas en ese lenguaje. En su sitio web ofrecen tutoriales o instructivos para aprender a utilizar la herramienta.

Por otro lado, Google Analytics es una herramienta gratuita, de análisis de datos de sitios web que pertenece a la empresa Google. Analytics permite visualizar datos obtenidos de una página web estática y arroja datos en tiempo real.

Para poder usar Google Analytics se requiere conectar con el servicio, utilizando los datos solicitados. Asimismo, se trabaja desde la nube por lo que no se requiere la instalación de ningún software en una computadora.

Por último, tenemos Orex Analytics. En su sitio web, lo definen como

> Una herramienta de recomendación lectora mediante inteligencia artificial y una herramienta de Big data para bibliotecas, que permite analizar en tiempo real todo lo que pasa en su biblioteca, ofreciendo potentes herramientas de análisis y detección de tendencias, así como la posibilidad de añadir al sistema de gestión la experiencia de su personal, creando un valor diferencial no disponible en los grandes buscadores (Orex, 2023, párr. 1).

Este desarrollo funciona en dos vertientes; por un lado, ofrecen Orex Analytics Big data (OA-B)**,** herramienta que "recolecta e indexa en tiempo real información de múltiples fuentes tanto internas como externas y las pone a disposición mediante múltiples paneles especializados con indicadores, gráficos e informes totalmente interactivos" (Orex, 2023, párr. 3).

La segunda es Orex analytics Recomendación (OA-R), inteligencia artificial para recomendación lectora, que sugiere otros recursos de información que podrían ser "relevantes a los usuarios, mediante el análisis del patrón de lectura de usuarios complementado con las recomendaciones del personal de la biblioteca y de verificaciones de los usuarios" (Orex, 2023, párr. 4). La herramienta tiene un costo y se desconoce cómo es su línea de código.

Con el fin de comparar estos tres softwares, se presenta a continuación la tabla 2. Con cada "sí" se sumará un punto que nos permitirá evaluar cada uno.

Tabla 3 Comparativo de características entre Google Analytics, Metabase y Orex

Características	Metabase	Google Analytics	Orex analytics (Orex, 2022)
1. Idioma (español)	Sí	Sí	Sí
2. Documentación para la instalación y uso	Sí	Sí	No
3. Requisitos del sistema (que pueda ser utilizado en diferentes sistemas operativos)	Sí	Sí	No
4. Interfaz amigable	Sí	Sí	Sí

5. Conocimientos técnicos para su vinculación con las bases de datos	Sí	Sí	No
6. Mayor número de funcionalidades, por ejemplo, gráficas, contador de consultas, etcétera	Sí	Sí	Sí
7. Visualización de datos en tiempo real	Sí	Sí	Sí
8. Volumen	Sí	Sí	Sí
9. Variedad	Sí	No	Sí
10. Velocidad	Sí	Sí	Sí
11. Costo asequible	Sí	Sí	No
Total:	**11**	**10**	**7**

Fuente: elaboración propia.

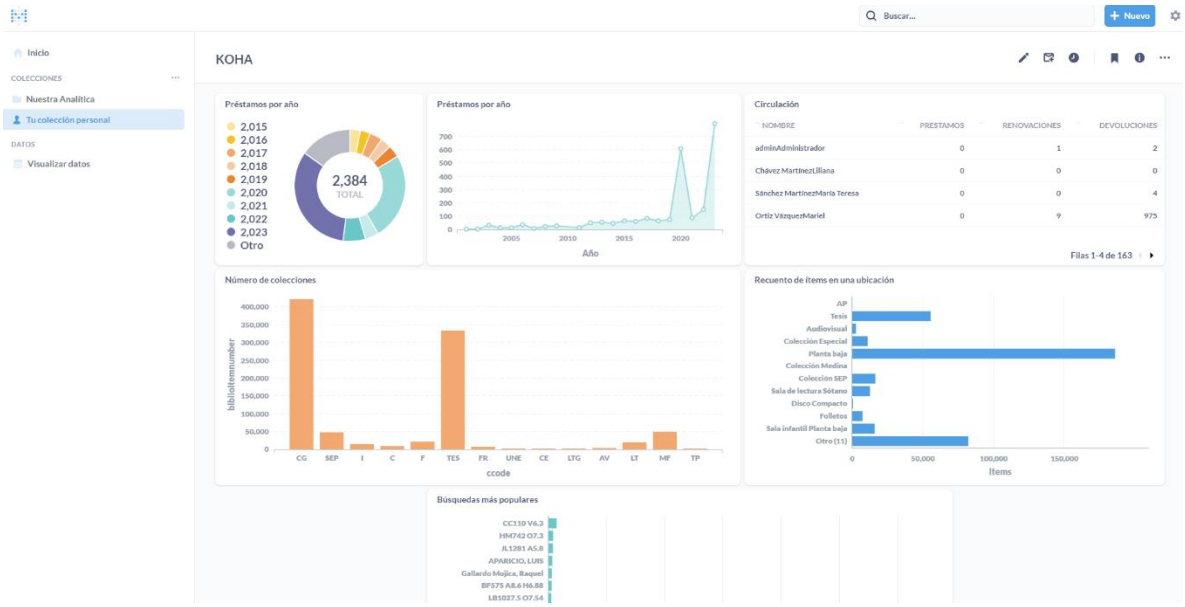
A partir de la comparación de estas 11 características, se puede observar que Metabase obtuvo el mayor puntaje, seguido por Google Analytics y dejando al último el software Orex.

Es importante resaltar que, para el análisis de los datos, puede usarse Metabase y Google Analytics, con el fin de obtener más claridad y profundidad sobre los datos obtenidos considerando que cada herramienta evalúa aspectos distintos pero la consigna sigue siendo la misma; es importante tener claridad sobre lo que se le va a preguntar a estos sistemas.

6. **Visualizar datos**

Para la visualización de los datos obtenidos a partir de este ejercicio, se utilizaron los mismos softwares elegidos en el paso 5, Metabase y Google Analytics. Comenzaremos con las visualizaciones obtenidas de Metabase (ilustración 12):

Ilustración 12 Datos generales que muestra Metabase



Fuente: elaboración propia con apoyo de Luis Roberto Polo Bautista y Eduardo Martínez García, equipo de informática de la Biblioteca Gregorio Torres Quintero de la UPN.

La consulta se realizó el día 29 de marzo de 2023. Al vincular los datos de Koha con Metabase, de inmediato nos ofrece las siguientes gráficas en las que podemos visualizar los préstamos por año con diferentes esquemas, el número de colecciones, las búsquedas más populares, algunos datos de circulación y el recuento de ítems en una ubicación.

Estas son algunas de las opciones que nos brinda el software, pero es importante aclarar que se pueden obtener muchísimas más visualizaciones si se programan consultas en SQL o si desde Metabase, se selecciona la carpeta y los datos de Koha que se quieran utilizar y con qué gráfico.

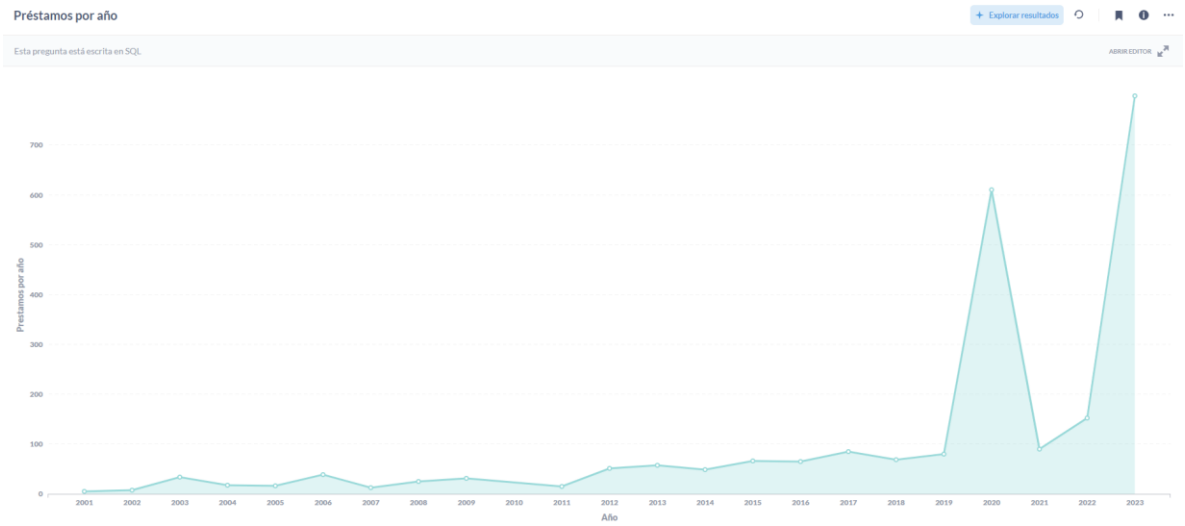
Para conocer los préstamos por año, se puede realizar la consulta mediante SQL con la siguiente fórmula:

select count(*) as 'Prestamos por año', year (issues.issuedate) as 'Año' from ((issues left join borrowers on issues.borrowernumber =

borrowers.borrowernumber) left join categories on borrowers.categorycode = categories.categorycode)

where issues.borrowernumber <> 0
group by año

Ilustración 13 Préstamos por año



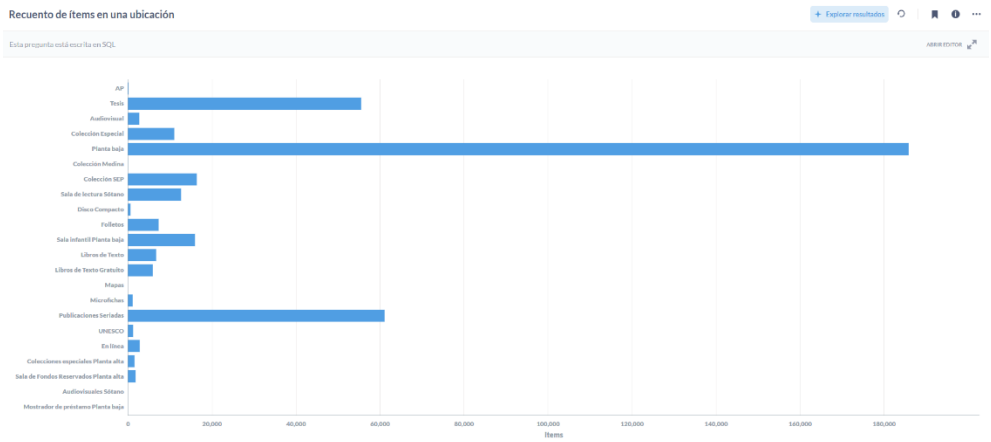
Fuente: elaboración propia con apoyo de Luis Roberto Polo Bautista y Eduardo Martínez García, equipo de informática de la Biblioteca Gregorio Torres Quintero de la UPN.

Podemos visualizar que durante este año se han prestado más recursos de información que los años previos, dado que, aunque tuvo su último repunte en el 2020, comenzó la pandemia haciendo que estos disminuyeran.

Otra consulta que se realizó es la del recuento de ítems en una ubicación, con la fórmula en SQL:

select v.lib as loc, count(i.itemnumber) as items
from authorised_values v
left join items i ON (i.location=v.authorised_value)
where v.category='LOC'
group by v.id

Ilustración 14 Recuento de ítems en una ubicación

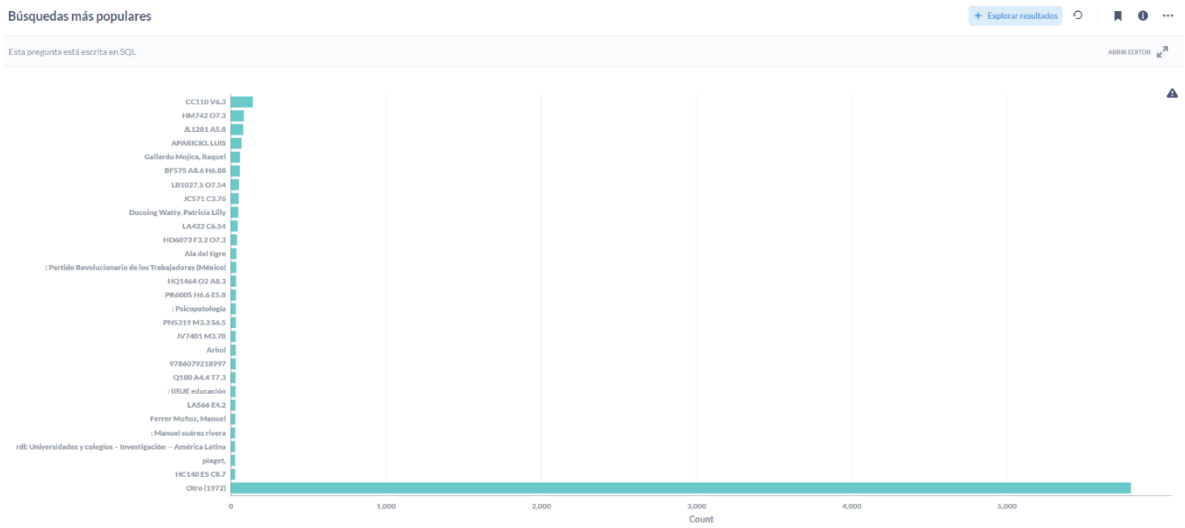


Fuente: elaboración propia con apoyo de Luis Roberto Polo Bautista y Eduardo Martínez García, equipo de informática de la Biblioteca Gregorio Torres Quintero de la UPN.

Podemos ver que las colecciones o ubicaciones con mayor número de ítems son las de la planta baja (en la que se encuentra la colección general), publicaciones seriadas y por último la de tesis, seguida por la colección SEP, la colección de la sala infantil y los recursos de la sala de lectura del sótano.

Asimismo, tenemos la consulta sobre las búsquedas que más se realizan en el OPAC, realizada en SQL.

Ilustración 15 Búsquedas más populares



Fuente: elaboración propia con apoyo de Luis Roberto Polo Bautista y Eduardo Martínez García, equipo de informática de la Biblioteca Gregorio Torres Quintero de la UPN.

Algo que llama mucho la atención es que, de primera vista, se puede observar que los usuarios primero buscan por signatura topográfica. Después buscan por autores, pertenecientes a la bibliografía básica de los programas de estudios y, posteriormente, algunos temas en general.

Una posible explicación a esto es que, dado que se encuentran en marcha diversos proyectos de organización con el equipo de catalogadores, ellos realizan búsquedas en el OPAC con la signatura topográfica, por lo que estos datos no son del todo fiables en cuanto a que provengan de las comunidades objetivo.

7. **Optimizar los servicios de la biblioteca o toma de decisiones**

A partir de los resultados obtenidos de esta consulta, se realizó un comunicado con la directora de la biblioteca de ese momento.

8. **Evaluar el proceso y los resultados**

Para este ejemplo, faltó ser mucho más precisos con las preguntas, así como incorporar las que fueron surgiendo.

3.4 Discusión

A partir del ejercicio anteriormente mencionado, se pudieron observar algunas cuestiones a considerar:

En cuanto a los pasos que integran el modelo, se pudo constatar que son flexibles y suficientes para su aplicación. Esto es relevante porque al ser un modelo moldeable, según las necesidades de la institución y los objetivos de observación, permite su aplicación en diversos contextos y situaciones. La adaptabilidad de dicho modelo es crucial para que diferentes bibliotecas puedan beneficiarse de sus principios y métodos.

Es de vital importancia comprender el contexto en el que se aplicará la metodología de Big data: el tipo de biblioteca, los usuarios y los objetivos específicos de observación. Sin una planificación clara desde el inicio, es fácil perderse en cada una de las etapas del proceso. Entonces, definir estos elementos contextuales es esencial para asegurar la coherencia y eficacia del análisis de Big data.

Por otro lado, el conocimiento técnico de los bibliotecarios sobre los sistemas informáticos no siempre es suficiente para localizar los datos necesarios para la consulta (paso 1 del modelo), analizar los datos (paso 5) y visualizarlos (paso 6). Por lo tanto, es imprescindible contar con el apoyo del personal de sistemas de la institución, y de ese modo llevar a cabo estas tareas técnicas de manera efectiva.

En cuanto a la dificultad de los demás pasos del modelo, con el tiempo, los bibliotecarios pueden adquirir las habilidades necesarias para realizarlos. Un ejemplo de ello es cómo aquellos dedicados a la bibliometría pueden crear redes bibliométricas, lo que demuestra que, con la capacitación adecuada, es posible superar las barreras iniciales.

En el caso del ejemplo mostrado en este apartado, se utilizó un volumen de datos pequeño. Si se hubieran empleado datos de un periodo más largo o de otras fuentes, como el sitio web, los formularios de satisfacción de usuarios, etc., se habrían obtenido consultas más precisas y fundamentadas. La falta de datos limitó la capacidad para realizar análisis más exhaustivos y concluyentes.

Por ejemplo, hizo falta realizar un análisis de uso de recursos para comprender cómo los usuarios interactúan con los libros, las revistas, las bases de datos y los recursos en línea disponibles. Identificar los materiales más y menos utilizados podría haber permitido tanto desarrollar campañas de promoción para los recursos subutilizados como evaluar su impacto en consultas posteriores.

Además, no fue posible visualizar cuáles son los temas más populares entre profesores y estudiantes debido a la insuficiencia de datos. Este tipo de análisis es crucial para ajustar las colecciones y los servicios a las necesidades actuales de los usuarios.

Un aspecto interesante para considerar es que Koha suele recomendar libros basándose en la proximidad en la estantería. Sería muy beneficioso, en futuros estudios, utilizar Big data para integrar un sistema experto que pueda recomendar libros de manera personalizada, basándose en el historial de búsqueda y préstamo de los usuarios.

Otro uso interesante que no se pudo implementar, debido a la falta de datos, fue el análisis de tendencias educativas emergentes a partir de los temas de investigación populares, cambios en el currículo de los programas o nuevas metodologías. Este tipo de análisis podría ser muy útil para la biblioteca, permitiéndole anticipar las necesidades futuras de los usuarios y garantizar que las colecciones y servicios estén alineados con las tendencias actuales.

Un aspecto que podría explorarse en futuros estudios es la evaluación del impacto educativo de los servicios y recursos de la biblioteca. Por ejemplo, mediante el seguimiento del rendimiento académico de los estudiantes que utilizan los recursos de la biblioteca o participan en programas de apoyo académico. Esto proporcionaría información valiosa sobre la efectividad de los servicios de la biblioteca en el apoyo al éxito académico; además, justificaría el uso de los recursos humanos y materiales.

Dado que en este trabajo se utilizaron principalmente los datos generados por el OPAC, se descartaron otras posibilidades como los formularios de registro de las salas de estudio, los registros de entrada a la biblioteca, el número de consultas a los referencistas (tanto presencial como en línea) y el uso de la hemeroteca o la sala infantil. Considerar estos datos adicionales podría optimizar la gestión del inventario, los horarios de los bibliotecarios y la planificación de los espacios. Comprender mejor los patrones de uso de los recursos disponibles posibilitaría una administración más eficiente y aseguraría que los servicios estén disponibles cuando y donde se necesiten.

3.5 Recomendaciones para la aplicación de datos masivos en bibliotecas académicas

El avance de las tecnologías de *Big data* representa una oportunidad significativa para que las bibliotecas académicas modernicen sus servicios, optimicen la gestión de la información y mejoren la experiencia de sus usuarios. Sin embargo, la implementación de estas tecnologías en el contexto mexicano enfrenta desafíos específicos que deben ser abordados para lograr una transición efectiva. A continuación, se presentan los principales elementos que las bibliotecas académicas en México deben considerar para adaptar sus servicios e infraestructura a las exigencias de los datos masivos.

1. Infraestructura tecnológica

Para procesar y analizar grandes volúmenes de datos, las bibliotecas requieren una infraestructura tecnológica robusta. Esto implica contar con servidores capaces de manejar almacenamiento masivo, bases de datos escalables y herramientas avanzadas de análisis de datos. Además, la adopción de soluciones en la nube puede representar una alternativa viable para instituciones con recursos limitados, permitiendo el acceso a tecnologías de datos masivos sin la necesidad de grandes inversiones en hardware.

2. Capacitación del personal bibliotecario

La implementación de procesos de datos masivos requiere personal capacitado en el manejo y análisis de datos. Es fundamental que los bibliotecólogos y otros profesionales de la información adquieran competencias en estadística, ciencia de datos, visualización de información y herramientas analíticas. Esto puede lograrse mediante programas de formación continua, certificaciones especializadas y colaboración con universidades y centros de investigación en el área de tecnologías de la información.

3. Políticas y normativas de gestión de datos

El manejo de datos masivos en bibliotecas académicas debe regirse por normativas que garanticen la seguridad, privacidad y uso ético de la información. Es necesario establecer políticas de recolección, almacenamiento y procesamiento de datos que cumplan con las regulaciones nacionales e internacionales sobre protección de datos. Asimismo, la transparencia en el uso de la información es clave para generar confianza entre los usuarios y asegurar que los datos se empleen con fines académicos y de mejora de servicios.

4. Colaboración interinstitucional

Dado que muchas bibliotecas académicas enfrentan limitaciones presupuestarias y tecnológicas, la colaboración interinstitucional se vuelve esencial. La creación de redes de bibliotecas que compartan infraestructura, conocimientos y estrategias de implementación de datos masivos puede facilitar el acceso a estas tecnologías. Además, establecer alianzas con universidades,

empresas tecnológicas y expertos en ciencia de datos puede fortalecer la capacidad de las bibliotecas para innovar y optimizar sus procesos.

5. Aplicaciones y beneficios de *Big data* en bibliotecas académicas

Finalmente, la adaptación de las bibliotecas a metodologías de datos masivos, no solo responde a una necesidad tecnológica, sino que también ofrece beneficios tangibles para la gestión de información y la experiencia del usuario. Entre sus aplicaciones más relevantes se encuentran la personalización de servicios bibliotecarios, la optimización de catálogos y sistemas de búsqueda, la automatización de procesos administrativos y la generación de análisis predictivos para mejorar la toma de decisiones dentro de las instituciones académicas.

Para la aplicación de metodologías de datos masivos en bibliotecas académicas mexicanas requiere un enfoque estratégico que contemple infraestructura, capacitación, normativas, colaboración y aplicaciones específicas. Si bien los desafíos son significativos, la adopción de estas tecnologías puede transformar el papel de las bibliotecas en el ecosistema académico, permitiéndoles ofrecer servicios más eficientes y adaptados a las necesidades de los usuarios en la era digital.

Conclusiones

Este trabajo surge con la inquietud de aplicar diferentes metodologías a las bibliotecas a partir del desarrollo de las herramientas de los últimos años, de dar un nuevo impulso a los esfuerzos que ya se han estado realizando.

Los datos masivos (Big data) son aquellos grandes volúmenes de datos que requieren tecnologías avanzadas para su análisis. La minería de datos es el proceso de descubrir patrones y conocimiento a partir de grandes conjuntos de datos. Estas definiciones permiten establecer un marco teórico sólido para la aplicación de Big data en bibliotecas académicas.

Las metodologías de datos masivos se han utilizado en áreas distintas dentro del ámbito humano y las bibliotecas no deberían ser una excepción. Existen diversas metodologías para el manejo y análisis de datos masivos, tales como el procesamiento en paralelo, algoritmos de aprendizaje automático y técnicas de análisis predictivo. Estos son cruciales para manejar los grandes volúmenes de datos generados por los sistemas de bibliotecas.

Las propuestas que consideran la vinculación de los datos masivos con la toma de decisiones benefician a las bibliotecas en diversos sentidos: permiten la planeación y propuesta de servicios y desarrollos de acuerdo con lo que las comunidades de usuarios necesitan, con la precisión de saber si es lo que requieren.

La mayoría de las bibliotecas, aun las académicas, no cuentan con la infraestructura o el soporte técnico para realizar experiencias de datos masivos, pero sí podrían realizar ejercicios de minería de datos para proponer proyectos con sustento cuantitativo.

En el contexto de las bibliotecas académicas, las preguntas pueden ser resueltas utilizando tanto Big data como minería de datos. Mientras que Big data se

focaliza en la gestión y el procesamiento de grandes volúmenes de datos, la minería de datos se centra en extraer conocimiento útil de esos datos. Es posible extraer y analizar los datos recopilados por un software de biblioteca utilizando estas metodologías. Los sistemas de gestión bibliotecaria, como el OPAC, generan datos valiosos que pueden ser procesados para obtener perspectivas relevantes.

La aplicación de metodologías de análisis de datos permite identificar patrones en el uso del OPAC, tales como las temáticas más buscadas, las tendencias en el uso de recursos y las preferencias de los usuarios. Estos patrones son esenciales para tomar decisiones informadas sobre la gestión bibliotecaria.
Los resultados del análisis de datos permiten implementar cambios en la catalogación y en las respuestas del OPAC. Es posible personalizar las recomendaciones de libros y mejorar la precisión de las búsquedas basándose en el historial de uso, los patrones de comportamiento de los usuarios y lo disponible de las colecciones.

Los bibliotecarios no contamos con los conocimientos para realizar estas propuestas en solitario; se debe colaborar con personas con conocimientos en ingeniería o sistemas. Y se requiere que las escuelas formadoras de profesionales o las asociaciones en México ofrezcan, tal vez de manera optativa, la posibilidad de adquirir conocimientos vinculados a lenguajes de programación, manejo de datos, etc.

Ahora bien, sobre el objetivo de este trabajo que es “Diseñar un modelo para extracción de datos provenientes del catálogo para analizarlos y, a través de ello, realizar una propuesta de mejora para el OPAC”. Se ha explicado, por un lado, cuáles son los pasos del modelo sugerido en el apartado 3.1, así como en el apartado 3.4 se ha mostrado cuáles son los datos que tendrían que utilizarse del OPAC, así como de otras fuentes, para aprovecharlos en la toma de decisiones.

Sobre la hipótesis planteada “Si se aplica un modelo de recuperación y análisis de datos provenientes del OPAC, se pueden utilizar para analizar el comportamiento de la comunidad y mejorar los servicios del catálogo”, se intentó aplicar un modelo de Big data, pero a partir del ejercicio realizado en el capítulo 3, es evidente que, en el caso de la mayoría de las bibliotecas académicas de México, no podrá aplicarse datos masivos debido a que la cantidad de datos recabados son limitados. Más bien, se tendría que aplicar minería de datos.

El modelo propuesto en este trabajo es de utilidad en minería de datos también. La tecnología de los sistemas expertos y los enfoques actuales sobre el uso de los datos permiten que la biblioteca pueda proporcionar mejores servicios, recursos de acuerdo con lo que realmente necesita su comunidad y la posibilidad de mejora constante.

La integración de Big data y minería de datos en el ámbito de las bibliotecas académicas ofrece un enorme potencial para mejorar la eficiencia y la calidad de los servicios bibliotecarios. Las metodologías de análisis de datos no solo permiten una gestión más efectiva de los recursos, sino que también proporcionan herramientas para anticipar las necesidades de los usuarios y adaptar los servicios de manera proactiva. La implementación de estas tecnologías puede transformar la manera en que las bibliotecas académicas operan, optimizando su impacto educativo y su capacidad de respuesta a las demandas cambiantes del entorno académico.

Investigaciones futuras

A partir de los resultados obtenidos en este estudio, se identifican diversas oportunidades para futuras investigaciones en el ámbito del *Big Data* y la minería de datos aplicados a bibliotecas académicas. Algunas de las líneas de investigación que podrían explorarse son:

1. **Evaluación del impacto del *Big Data* en la toma de decisiones bibliotecarias**

   Un estudio que analice cómo la integración de *Big Data* influye en la planificación estratégica de las bibliotecas, la optimización de recursos y la mejora en la toma de decisiones basadas en datos.

2. **Modelos de implementación de *Big Data* en bibliotecas académicas mexicanas**

   Un análisis de casos específicos que documente las estrategias, retos y beneficios de bibliotecas que han comenzado a adoptar herramientas de *Big Data*, con el fin de generar modelos replicables para otras instituciones.

3. **Análisis del uso de inteligencia artificial y sistemas expertos en bibliotecas**

   Investigar cómo la combinación de *Big Data* con sistemas expertos e inteligencia artificial puede optimizar la gestión del conocimiento, mejorar los sistemas de recomendación y automatizar procesos dentro de las bibliotecas académicas.

4. **Percepción y preparación del personal bibliotecario ante el uso de *Big Data***

   Un estudio cualitativo que explore la actitud, conocimientos y necesidades de formación del personal bibliotecario en torno a la adopción de herramientas de *Big Data* y análisis de información.

5. **Desafíos éticos y normativos del uso de *Big Data* en bibliotecas académicas**

   Analizar las implicaciones éticas y legales en el manejo de datos masivos en bibliotecas, incluyendo privacidad de usuarios, gestión de datos sensibles y cumplimiento de regulaciones nacionales e internacionales.

Estas líneas de investigación pueden contribuir significativamente al desarrollo del campo bibliotecológico en la era digital. Profundizar en estos temas permitirá

generar estrategias más efectivas para la implementación de *Big Data* en bibliotecas académicas mexicanas, impulsando su modernización y adaptación a las nuevas exigencias tecnológicas.

Referencias

AECL. (2013). *Historia de la biblioteca*. https://www.comprensionlectora.es/index.php/2013-11-27-16-50-54/2013-11-27-17-15-39/historia

ALA. (2013). LibGuides Definition of a Library: General Definition. https://libguides.ala.org/library-definition

ALA. (2016). *Special Libraries*. Education & Careers. https://www.ala.org/educationcareers/libcareers/type/special

Andalia, R. (2004). *De la piedra al web : análisis de la evolución histórica y del estado actual de la actividad bibliológico-informacional.* 12. https://www.researchgate.net/publication/28802175_De_la_piedra_al_web_analisis_de_la_evolucion_historica_y_del_estado_actual_de_la_actividad_bibliologico-informacional

Apache Hadoop. (2023). Apache.org. https://hadoop.apache.org/

Apache Spark. (2023). *What is Apache Spark™?* https://spark.apache.org/

Arriola Navarrete, O., y Montes de Oca, E. (2014). Sistemas Integrales de Automatización de Bibliotecas: una descripción suscinta. *Bibliotecas y Archivos* *1*(3), 47-76. http://eprints.rclis.org/24259/1/Art%C3%ADculo%20SIAB%20publicada.pdf

Asemi, A., Ko, A. y Nowkarizi, M. (2021). Intelligent libraries: a review on expert systems, artificial intelligence, and robot. *Library Hi Tech*, *39*(2), 412-434. https://doi-org.pbidi.unam.mx:2443/10.1108/LHT-02-2020-0038

Ashikuzzaman, M. D. (2018, 7 de febrero). *Online Public Access Catalogue (OPAC)*. Library & Information Science Education Network; Library & Information Science Education Network. https://www.lisedunetwork.com/online-public-access-catalogue-opac/

Badaró, S., Ibañez, L. y Agüero, M. (2013). Sistemas Expertos: Fundamentos, Metodologías y Aplicaciones. *Ciencia y Tecnología*, *13*, 349-364https://www.palermo.edu/ingenieria/pdf2014/13/CyT_13_24.pdf

Bello, E. (2022, 9 de marzo). *Big data: qué es, para qué sirve y por qué es importante*. Thinking for Innovation. https://www.iebschool.com/blog/valor-

big-data/#:~:text=El%20Big%20Data%20es%20usado,obtener%20informaci%C3%B3n%20sobre%20los%20clientes

Biblioteca. (consultado el 20 de julio de 2023). En *Wikipedia*. https://es.wikipedia.org/wiki/Biblioteca

Bharati, M. y Ramageri, Bharati. (2010). Data mining techniques and applications. Indian Journal of Computer Science and Engineering. *Indian Journal of Computer Science and Engineering*, *1*(4), 301-305. https://www.researchgate.net/publication/49616224_Data_mining_techniques_and_applications/fulltext/57a9affe08ae659d18249ecd/Data-mining-techniques-and-applications.pdf

Bravo, E. (2022). *Propuesta para la automatización de la Biblioteca Pública Municipal no. 22 "Dr. Alfonso G. Alarcón" de la ciudad de Acapulco, Gro* [tesis de licenciatura. Universidad Nacional Autónoma de México]. Repositorio. http://132.248.9.195/ptd2022/marzo/0823650/Index.html

Bonilla Gordillo, A. F. y Ojeda Schuldt, M. A. (2006). *Implementación de Minería de datos basada en redes bayesianas para la toma de decisiones en los registros académicos* [tesis de ingeniería, Escuela Superior Politécnica del Litoral]. Repositorio. https://www.dspace.espol.edu.ec/bitstream/123456789/3088/1/D-83914.pdf

Bronsoiler Frid, C. (1986). *La enseñanza de la automatización en la currícula de bibliotecología* [tesis de maestría, Universidad Nacional Autónoma de México].

Buonocore, D. (1976). *Diccionario de Bibliotecología: Términos relativos a la bibliotecología, bibliografía, bibliofilia, biblioteconomía, archivología, documentología, tipografía y materias afines*. Marymar.

Calva Díaz, J. E. (2023). *Sistema experto híbrido basado en balances macroscópicos de materia y energía para procesos metalúrgicos* [tesis de ingeniería, Universidad Nacional Autónoma de México]. Repositorio.

https://ru.dgb.unam.mx/bitstream/20.500.14330/TES01000839347/3/0839347.pdf

Comunidad Baratz. (2020, 21 de mayo). *Las distintas clasificaciones y tipologías de bibliotecas según UNESCO, INE, IFLA y ALA*. https://www.comunidadbaratz.com/blog/las-distintas-clasificaciones-y-tipologias-de-bibliotecas-segun-unesco-ine-ifla-y-ala/

Datahack (s.f.) Historia del Big Data. Años 80 https://www.datahack.es/historia-big-data-anos-80/

Data mining. (2022). Ox.ac.uk. https://www.bodleian.ox.ac.uk/collections-and-resources/data/accessing-and-using-data/data-analysis/data-mining

DGIRE-UNAM. (2007). *Catalogación descriptiva*. https://www.dgire.unam.mx/contenido/bibliotecas/texto/37.html

Datos.gob.es. (2021, 10 de diciembre). *¿Por qué deberías de usar ficheros Parquet si procesas muchos datos?* https://datos.gob.es/es/blog/por-que-deberias-de-usar-ficheros-parquet-si-procesas-muchos-datos

DIGIXEM360. (2023). *Sistemas expertos: qué son, cómo funcionan y para qué sirven*. https://www.innovaciondigital360.com/i-a/sistemas-expertos-que-son-su-clasificacion-como-funcionan-y-para-que-se-utilizan/

Enciso, B. (1993). *La Biblioteca: bibliosistemática e información*. El Colegio de México. https://www.cervantesvirtual.com/obra/la-biblioteca-bibliosistematica-e-informacion--0/

EGOS BI. (2021, 9 de febrero). *La historia del Big data: sus orígenes y evolución*. https://www.egosbi.com/historia-del-big-data/

Eserada, R. E. y Okolo, S. E. (2019). Use of Online Public Access Catalogue (OPAC) in Selected University Libraries in South- South Nigeria. *Library Philosophy and Practice (e-journal)*, 2586. https://digitalcommons.unl.edu/libphilprac/2586

Flores Lagla, G., Cadena Moreano, J., Quinatoa Arequipa, E. y Villa Quishpe, M. (2019). Minería de datos como herramienta estratégica. *Revista Científica Mundo de la Investigación y el Conocimiento*, *3*(1), 955-970. http://www.recimundo.com/index.php/es/article/view/400

Flores, P. y Villacís, A. (2017). *Análisis comparativo de las herramientas de Big data en la Facultad de Ingeniería de la Pontificia Universidad Católica del Ecuador* [tesis de ingeniería, Universidad Católica del Ecuador]. Repositorio. https://repositorio.puce.edu.ec/server/api/core/bitstreams/d9977d1a-f7e3-4748-ac67-47f80fe0cac6/content

García-Alsina, M. (2017). *Big data, gestión y explotación de grandes volúmenes de datos.* UOC. https://play.google.com/books/reader?id=SFgtEAAAQBAJ&pg=GBS.PT1&hl=es_419

Garduño, R. (2004). La sociedad de la información en México frente al uso de internet. *Revista Digital Universitaria*, (*5*)8, 2-13. http://www.revista.unam.mx/vol.5/num8/art50/sep_art50.pdf

Garoufallou, E. y Gaitanou, P. (2021). Big Data: Opportunities and Challenges in Libraries, a Systematic Literature Review. *College & Research Libraries*, *82*(3), 410. https://crl.acrl.org/index.php/crl/article/view/24918/32769

Garza Mercado, A. (1984). *Función y forma de la biblioteca universitaria: elementos de planeación administrativa para el diseño arquitectónico*. El Colegio de México.

Gómez, J. A. (2002). *Gestión de bibliotecas: Texto-Guía de las asignaturas de "Biblioteconomía general y especializada"*. Universidad de Murcia. http://eprints.rclis.org/10372/1/Gestion_de_Bibliotecas_Gomez-Hernandez_2002.pdf

González, E. (2020). *Modelo de diseño y formación de uso de una biblioteca infantil digital como apoyo a la educación preescolar mexicana* [tesis de licenciatura, Universidad Nacional Autónoma de México]. Repositorio. http://132.248.9.195/ptd2020/noviembre/0804848/Index.html

Harvard Library. (2022a). *About*. https://hlrdm.library.harvard.edu/about

Harvard Library. (2022b). *Harvard Dataverse*. https://library.harvard.edu/services-tools/harvard-dataverse

Herrera Morrilla, J. L. (2010). *Las nuevas tecnologías y las bibliotecas: una síntesis sobre su evolución y repercusiones*. http://www.aldee.org/cas/content/publicaciones/upload/jorna04.pdf

IBM Documentation. (2022, 28 de noviembre). Data mining goals. https://www.ibm.com/docs/es/db2/11.1?topic=overview-data-mining-goals

IBM. (2024a). *Formato JSON (JavaScript Object Notation)*. https://www.ibm.com/docs/es/baw/20.x?topic=formats-javascript-object-notation-json-format

IBM. (2024b). *¿Qué es Avro?* https://www.ibm.com/mx-es/topics/avro

IBM. (2024c). Edgar Codd. https://www.ibm.com/history/edgar-codd

IFLA. (2024) *Why Big Data Matters: Perspectives from the Libraries*. https://www.ifla.org/es/news/why-big-data-matters-perspectives-from-the-libraries/

Islas Monroy, E. (2021). *El análisis de Big data como recurso de investigación en la Sociología* [tesis de maestría, Universidad Nacional Autónoma de México]. Repositorio. http://132.248.9.195/ptd2021/agosto/0814212/Index.html

Jetty, S. A. K., J. Jain, P. K., y Hopkinson, A. (2011). *OPAC 2.0 Towards the next generation of online library catalogues*. https://core.ac.uk/download/pdf/17301580.pdf

L. J. C. (1946). *The Application of Commercial Calculating Machines to Scientific Computing*. Mathematical Tables and Other Aids to Computation, 2(16), 149–159. https://doi.org/10.2307/2002577

Lesk, M. (s.f.) How Much Information Is There In the World? https://www.lesk.com/mlesk/ksg97/ksg.html

Lyons, M. (2011). Libros: dos mil años de historia ilustrada. Lunwerg.

Maimon, O. y Rokach, Li (eds.). (2010). *Data Mining and Knowledge Discovery Handbook*. Springer. doi: 10.1007/978-0-387-09823-4

Mauro, A. de, Greco, M. y Grimaldi, M. (2016). *A formal definition of Big Data based on its essential features*. ResearchGate; Emerald. https://www.researchgate.net/publication/299379163_A_formal_definition_of_Big_Data_based_on_its_essential_features

Martínez Acevedo, K. y Polo Bautista, L. (2021). *Aplicación de la bibliominería medotológica en la elaboración de una ontología como sistema de representación del conocimiento de la enfermedad del tifo en México, 1904-1977* [tesis de licenciatura, Escuela Nacional de Biblioteconomía y archivonomía]. Repositorio. http://eprints.rclis.org/43272/1/Aplicaci%C3%B3n%20de%20la%20bibliominer%C3%ADa%20metodol%C3%B3gica%20en%20la%20elaboraci%C3%B3n%20de%20una%20ontolog%C3%ADa%20como%20sistema%20de%20representaci%C3%B3n%20del%20conocimiento%20la%20enfermedad%20del%20tifo%20en%20M%C3%A9xico%2C%201904-1977.pdf

Medina Rojas, F. y Gómez, C. (2014). Funcionalidades de la minería de datos. *Ingeniería y Región*, *12*(2), 31. https://doi.org/10.25054/22161325.728

Microsoft. (s.f). *Crear o editar archivos .csv para importarlos a Outlook*. https://support.microsoft.com/es-es/office/crear-o-editar-archivos-csv-para-importarlos-a-outlook-4518d70d-8fe9-46ad-94fa-1494247193c7

Nahotko, M. (2020). OPAC development as the genre transition process, PART 1: OPAC generations historical development. *Annals of Library and Information Studies*, (*67*)2, 107-117.

Njuku, S. (2022). Application Of Expert Systems in Library and Information Services. *Nigerian Academic Forum*, *29*(1), 1-10. https://www.globalacademicgroup.com/journals/the%20nigerian%20academic%20forum/V29N1P13_Forum_2022.pdf

ORACLE. (2021). *¿Qué es el big data?* https://www.oracle.com/mx/big-data/what-is-big-data/

Orex. (2023). *Orex Analytics*. https://orex.es/productos/orex-analytics/#:~:text=Orex%20Analytics%20es%20una%20herramienta,la%20posibilidad%20de%20a%C3%B1adir%20al

Orera, L. (s.f.). *Reflexiones sobre el concepto de biblioteca*. Consultado el 28 de mayo de 2022. http://148.202.167.116:8080/xmlui/bitstream/handle/123456789/3621/Reflexiones_sobre_concepto_biblioteca.pdf?sequence=1&isAllowed=y

Ornelas, T. (2011). *Automatización de la Biblioteca "Dr. Gonzalo Aguirre Beltrán" del Instituto de Antropología de la Universidad Veracruzana* [tesis de licenciatura, Escuela Nacional de Biblioteconomía y Archivonomía]. Repositorio. http://www.bibliotecaenba.sep.gob.mx/tesis/Biblioteconomia2011/043460.pdf

Oxford English Dictionary. (2023). Oed.com. https://doi.org/10.1093//OED//8252450177

Palva Muñoz, P. A. (2016). *Desarrollo de una arquitectura Big data para registros mercantiles*. Universidad Central de Venezuela. http://saber.ucv.ve/bitstream/10872/14696/1/Tesis_Paiva_Final.pdf

Quijano Solís, A. (2007). Aceptación de tecnologías de información y cambio organizacional: propuesta metodológica para su planeación en una biblioteca académica [tesis de doctorado, Universidad Nacional Autónoma de México].

Raído, G. (2022, 13 de julio). *Cómo SHEIN orquesta todas sus operaciones con el análisis de datos*. Deyde DataCentric. https://www.datacentric.es/blog/insight/como-shein-orquesta-todas-sus-operaciones-con-el-analisis-de-datos/

Ramírez, L. (2022, 19 de abril). *Las 10 mejores herramientas de Big Data 2023*. Thinking for Innovation. https://www.iebschool.com/blog/mejores-herramientas-big-data/

Real Academia Española. (2016). *Diccionario de la lengua española.* [Consultado el 24 de julio de 2017]. http://www.rae.es/

Romero Jacome, I. (2005). *Tendencias de los catálogos de acceso público en línea (OPAC's),* [tesis de licenciatura, Universidad Nacional Autónoma de México]. Repositorio. https://repositorio.unam.mx/contenidos/131219

Saorín Pérez, T. (2002). *Modelo conceptual para la automatización de bibliotecas en el contexto digital,* [tesis de maestría, Universidad de Murcia].

Sawa, M. (2007). *El libro de las bibliotecas: historia de las bibliotecas, del camello a la computadora*. Celta; Amecamecan.

Serrano Barrera, E. y Aguilar Estrada, S. (2002). Modernización de procesos y servicios de las bibliotecas públicas del estado de Colima. En *Memoria del segundo congreso nacional de bibliotecas públicas. Guadalajara Jalisco del 23 al 25 de septiembre de 2002* (pp. 189-193). Consejo Nacional para la Cultura y las Artes-Dirección General de Publicaciones; Gobierno de Jalisco-Secretaría de Cultura.

Sucar, L. E. (s.f.). *Chapter 1: redes bayesianas*. INAOE https://ccc.inaoep.mx/~esucar/Clases-mgp/caprb.pdf

Tascón, M. (2018, 15 de febrero). *Pasado, presente y futuro*. Telos. https://telos.fundaciontelefonica.com/archivo/numero095/pasado-presente-y-futuro/

TechNews. (2018, 29 de abril). *Serving citizens better with Big Data*. https://medium.com/technewssg/serving-citizens-better-with-big-data-a8a0153fb583

Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, S. J., Hidalgo-Troya, A. y Alvarado Pérez, J. C. (2016). El proceso de descubrimiento de conocimiento en bases de datos. En *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional* (pp. 63-86). Ediciones Universidad Cooperativa de Colombia. http://dx.doi.org/10.16925/9789587600490

Ulate Montero, J. (2020). Propuesta de un Sistema Integrado de Gestión Bibliotecaria para el Sistema de Bibliotecas Municipales de la Municipalidad de San José. *Bibliotecas*, *38*(1), 1-27. https://doi.org/10.15359/rb.38-1.1

Universidad Complutense de Madrid. (s.f.) *Ficha o registro catalográfico*. *Quid est liber: proyecto de innovación para la docencia en libro antiguo y patrimonio bibliográfico*. Www.ucm.es. https://www.ucm.es/quidestliber/ficha-o-registro-catalografico

Varela-Prado, C. y Baiget, T. (2012). El futuro de las bibliotecas académicas: incertidumbres, oportunidades y retos. *Investigación Bibliotecológica: Archivonomía, Bibliotecología e información*, *26*(56), 115-135. https://doi.org/10.22201/iibi.0187358xp.2012.56.33175

Voutssás Márquez, J. (2019). *Los inicios de la automatización de bibliotecas en México*. UNAM, Instituto de Investigaciones Bibliotecológicas y de la Información. https://iibi.unam.mx/voutssasmt/documentos/inicios_automatizacion.pdf

Voutssás Márquez, J. (2022). *Datos masivos en bibliotecas*. UNAM, Instituto de Investigaciones Bibliotecológicas y de la Información. https://ru.iibi.unam.mx/jspui/bitstream/IIBI_UNAM/448/1/datos_masivos.pdf

Wiegand, W. A. (2000). American Library History Literature, 1947-1997: Theoretical Perspectives? Libraries & Culture, 35(1), 4–34. http://www.jstor.org/stable/25548795