CONTENIDOS DIGITALES:
CONVERGENCIA, CONECTIVIDAD, MODELOS Y NUEVAS CARACTERÍSTICAS

Ariel Alejandro Rodríguez García
Coordinador

Z666.7
C66

Contenidos digitales : convergencia, conectividad, modelos y nuevas características / Coordinador Ariel Alejandro Rodríguez García. - México : UNAM. Instituto de Investigaciones Bibliotecológicas y de la Información, 2022.

xvi, 292 p. - (Bibliotecología, información y sociedad)
ISBN: 978-607-30-6167-4

1. Metadatos - Modelos. 2. Datos vinculados. 3. Indización - Aspectos sociales. 4. Recuperación de información. I. Rodríguez García, Ariel Alejandro, coordinador. II. ser.

Diseño de portada: Nube Magenta

Primera edición: 19 de mayo de 2022

Instituto de Investigaciones Bibliotecológicas y de la Información
Circuito Interior s/n, Torre II de Humanidades,
pisos 11, 12 y 13, Ciudad Universitaria, C. P. 04510,
Alcaldía Coyoacán, Ciudad de México

ISBN: 978-607-30-6167-4

Publicación dictaminada.

Impreso y hecho en México.

Contenido

MODELOS

NUEVAS CARACTERÍSTICAS

Metadatos, ciencia de los datos y bibliotecas

JUAN VOUTSSÁS MÁRQUEZ
Instituto de Investigaciones Bibliotecológicas y de la Información, UNAM

INTRODUCCIÓN

En años recientes, el desarrollo de la ciencia de los datos ha aportado nuevas herramientas de análisis de información para la toma de decisiones en organizaciones, las bibliotecas entre ellas. El fenómeno denominado Big Data o datos masivos conforma una parte importante de esta ciencia. Las bibliotecas ya están desarrollando nuevas actividades y proyectos con esta metodología. En particular, el uso de metadatos en estos proyectos representa grandes oportunidades y retos. Es conveniente por tanto estudiar y comprender cómo se interrelacionan la ciencia de los datos, los datos masivos, los metadatos y las bibliotecas por ser un elemento de valor agregado tanto para la organización, como para el personal que se dedica a esas actividades.

CIENCIA DE LOS DATOS

La ciencia de los datos (*Data science)* ha tomado auge en los últimos años, pero como muchos otros conceptos tecnológicos, no es

del todo nueva. Peter Naur menciona en un texto desde 1968 "[…] la Datalogía, la ciencia de los datos y de los procesos de datos y su lugar en la educación". El autor definió: "[…] consiste en la ciencia del tratamiento de datos, una vez que se han establecido, mientras que la relación de los datos con lo que representan se delega a otros campos y ciencias" (Naur 1975, s. p.). Lógicamente, el término ha evolucionado en la actualidad y abarca aún más campos: el diseño, compilación y selección de datos, su gestión y fuentes, las técnicas para la minería de datos; los almacenes y repositorios de datos, el análisis y visualización de ellos, la Inteligencia Artificial y sistemas expertos; su preservación, políticas y gobernanza. La Ciencia de los Datos puede definirse actualmente como el estudio de los datos organizados para identificar aquellos que son importantes para la solución de problemas específicos de las organizaciones o de un cierto modelo de negocio; también se relaciona con el desarrollo de técnicas, herramientas y algoritmos que resuelven problemas a gran escala en las organizaciones.

Como toda ciencia, la de los datos está conformada por varios campos o subdisciplinas; entre las principales distinguimos la "gestión de datos" (*data management)*, y con ella una serie de principios, teorías, métodos, modalidades, herramientas y tecnologías para su tratamiento y aplicación, y cuyo propósito es colectar, validar, organizar, almacenar y utilizar datos de forma segura, eficiente y rentable para convertirlos en un recurso valioso dentro de una organización. La "ingeniería de datos" (*data engineering)*, que tiene que ver con la organización y recuperación de los datos, y se especializa en cuán intrínsecamente depurados y estructurados están los conjuntos de datos en una organización. El "análisis de datos" (*data analytics)* es la actividad de identificar cuáles variables de la organización pueden ser correlacionadas con ciertos datos para plantear preguntas correctas y eventualmente extraer soluciones.

Para funcionar eficazmente, la gestión de datos requiere de inicio de una "estrategia de datos", así como de métodos preestablecidos y fiables para su manejo y acceso: colecta, normalización, organización, almacenamiento, seguridad, etcétera, todo lo cual propiciará su correcto análisis. Una estrategia sólida de gestión de

datos es indispensable, ya que reunir grandes cantidades de datos sin un plan minucioso los convierte rápidamente en algo inútil y de difícil manejo. El verdadero valor de los datos no consiste simplemente en su existencia, sino en lo que se puede hacer con ellos: una verdadera capacidad de poder extraer información valiosa y útil de su conjunto, que además se traduzca en un beneficio tangible y cuantificable para la organización.

La ciencia y los datos

La ciencia se basó por siglos en dos paradigmas fundamentales: la teoría y la experimentación. Gracias a las teorías de Von Neumann y las computadoras a mediados del siglo pasado, se integró como tercer paradigma el modelado y la simulación con estos equipos. A comienzos de este siglo, Jim Gray planteó que existía ya un cuarto paradigma para la ciencia contemporánea que complementaba a los tres anteriores: los datos. La ciencia se basaba ya profundamente en ellos. Por lo mismo, era necesaria toda una nueva generación de conceptos, metodologías, herramientas y expertos para tratarlos. Hey y colegas (2009) publicaron la primera antología sobre el tema, considerada la piedra angular de nueva visión de la ciencia basada en los datos. Este autor también señaló desde 2006 que las nuevas facetas de la ciencia: la e-Ciencia o ciencia electrónica, ciencia abierta, etcétera, estaban ya estableciendo una nueva relación entre la ciencia y la biblioteca derivada justamente de los datos. Por lo mismo, desde entonces y cada vez más, los datos de investigación se han convertido en un gran insumo que requiere administrarse y preservarse de manera correcta.

Los datos masivos o Big Data

Una parte muy especial de la Ciencia de los Datos son los datos masivos o Big Data. Estos no son la evolución lineal de un cierto concepto o tecnología a lo largo de las décadas, sino la conjunción simultánea de múltiples fenómenos, necesidades, tendencias, tecnologías,

teorías, herramientas y métodos relacionados con la información, que al concurrir en un punto desembocan en algo más complejo. Tienen múltiples antecedentes, facetas y componentes que es necesario integrar para entender el concepto. Éstos son:

- La "explosión de la información" de la segunda mitad del siglo xx, que llevó a la "sociedad de la información".
- La llegada y el desarrollo del procesamiento de información con computadoras, el cual desarrolló el álgebra lineal, la simulación matemática, las bases de datos, las teorías de "colas" y de "árboles", los modelos económicos, las variables estocásticas, la gestión del conocimiento, etcétera.
- El crecimiento y abaratamiento del almacenamiento de la información.
- El inusitado desarrollo de la red mundial y de las telecomunicaciones a partir de los noventa, y en especial de las redes sociales.
- El desarrollo del *Internet de las cosas* y *wearables*, lo cual produjo todavía más datos.
- El desarrollo de conceptos avanzados como "ciudades digitales", "gobierno digital", "trámites digitales", etcétera.
- El uso cada vez mayor de datos para la solución de problemas empresariales, de negocios, gubernamentales, sociales, educativos y otros.

Se considera como punto de partida de la acepción actual de datos masivos una nota publicada a principios del 2001 por Doug Laney, analista especializado en información del Grupo Gartner, quien estableció las características fundamentales de este concepto, mismas que han sido utilizadas ampliamente a lo largo del siglo: *volumen, velocidad y variedad* (Laney 2001). El sitio web principal, de Google define los "datos masivos" como

> […] conjuntos de datos extremadamente grandes que pueden ser analizados computacionalmente para revelar patrones, tendencias y asociaciones, especialmente en relación con el comportamiento

> y las interacciones humanas [...] el uso actual del término 'datos masivos' tiende a referirse al uso del análisis predictivo, del análisis de comportamiento del usuario, o ciertos métodos avanzados de análisis de datos que extraen valor de los datos, y rara vez a un tamaño en particular de conjuntos de datos.

Metadatos y datos masivos

La clave fundamental para el manejo exitoso de datos masivos son los metadatos. Es imposible pensar en explotar datos –de cualquier volumen– sin contar con adecuados metadatos; sin ellos los conjuntos de datos, en especial los masivos, se vuelven una masa amorfa con poca o nula utilidad. Acerca de ellos existen algunos hechos muy interesantes:

- La Corporación IDC estableció en 2014 que se agregaban metadatos de forma sistematizada únicamente al 3 por ciento de la inmensa cantidad de datos que se estaba produciendo en el mundo (IDC 2014).
- Los metadatos son importantes en todo tipo de estructura de información, pero se vuelven centrales en el campo de los datos masivos, ya que gracias a ellos puede conocerse todo acerca de los datos: qué son, quién los generó, cuándo, cómo, dónde y por qué se generaron.
- Algo muy importante es que los metadatos no solo informan de los datos; también consignan sus elementos asociados: transacciones, formularios, programas y recursos informáticos, dispositivos, historias y un sinnúmero más de elementos potencialmente útiles y de interés para una organización.
- En el ámbito de los datos masivos, los metadatos pueden llegar a ser tan meticulosos y completos, que se convierten en "metainformación"; es decir, son segmentos de datos tan completos que pueden considerarse como información en sí mismos.
- Debido a lo anterior, los metadatos pueden ser más valiosos que el contenido.

Dependiendo de la naturaleza del dato y del emisor, los datos pueden ser estructurados, semi-estructurados o no estructurados. Conforme van teniendo más estructura, en cada tipo de dato habrá más metadatos asociados, pero a menudo no son evidentes para todos; es necesario conocer a fondo la estructura y esencia de cada tipo de dato para poder extraer de ellos algo valioso. Por ello es importante entender, diseñar, colectar y manejar los metadatos. Por ejemplo, Schmarzo (2018) consignó que existe la sorprendente cantidad de hasta veinte metadatos asociados a cada mensaje de Twitter, sin contar el contenido en sí mismo. Pocos caen en cuenta de la gran cantidad de metadatos existentes en algo tan pequeño y simple como un tuit de 280 caracteres. Es conveniente resaltar aquí que para los que estudian este tipo de mensajes, el contenido en sí no tiene ningún valor estadístico, pero esos eventuales veinte metadatos representan una mina de oro para el análisis de este tipo de red social.

Metadatos, datos masivos y bibliotecas

Metadatos, datos masivos y taxonomías

Desde hace tiempo, los bibliotecarios han estado conscientes del gran valor de los metadatos en el mundo de la información y por lo mismo están familiarizados en forma general con su diseño, creación y uso. Por ello han creado y utilizado de tiempo atrás toda clase de taxonomías de la información.[1] En décadas recientes esto ha tomado dimensiones inéditas, y se ha ido sofisticando hasta llegar en la actualidad al nivel de complejas ontologías, pasando por toda una serie de niveles intermedios: 1) Lexicón-vocabulario con definiciones en lenguaje natural. 2) Taxonomía simple –dic-

1 Los sistemas de organización documental del Vaticano, el de Jacques-Charles Brunet, el de William Harris, el decimal de Dewey, y el de la Biblioteca del Congreso de Cutter datan del siglo XIX. La CDU de Otlet y La Fontaine data de 1905 y el primer Código de Catalogación Angloamericano con sus tablas y esquemas derivados existe desde 1908.

cionarios de datos, jerarquías–. 3) Tesauro –taxonomía con términos relacionados–. 4) Modelo relacional –uso de restricciones de tipos y relaciones entre entes–. 5) Teoría axiomática completa.

Las teorías y prácticas acumuladas acerca del uso de diversas taxonomías en la Bibliotecología y otras ciencias de la información han logrado que puedan ser utilizadas y explotadas cada vez más por sistemas informáticos. La creación y colecta de datos masivos para este propósito, su procesamiento informático y su análisis, en conjunto con nuevas metodologías, permiten crear nuevas taxonomías puntuales, las cuales se generalizan en la ciencia de la información, y son utilizadas en áreas específicas prácticas, como es el caso de la Bibliotecología o la Archivística.

Un ejemplo de ello son los modelos conceptuales subyacentes de las Resource Description and Access (RDA), el estándar de catalogación para la formulación de registros bibliográficos usado en bibliotecas, archivos, museos, etcétera. Como es sabido, consisten en un conjunto de directrices, instrucciones y elementos de datos para crear metadatos de recursos bibliotecarios y de patrimonio cultural correctamente formados de acuerdo con los modelos internacionales para aplicaciones de datos vinculados orientadas al usuario. Esos modelos conceptuales subyacentes de las RDA son: los Requisitos Funcionales para Registros Bibliográficos (FRBR), los Requisitos Funcionales para Datos de Autoridades (FRAD), los Requisitos Funcionales para Datos de Autoridades de Temas o (FRSAD), y la ontología press, avalados por IFLA y compatibles con el Modelo de Referencia de Bibliotecas – *Library Reference Model*.[2]

2 Para abundar en estos modelos conceptuales, véase la página de la Federación Internacional de Bibliotecas y Asociaciones (IFLA), IFLA*'s Bibliographic Conceptual Models*, https://www.IFLA.org/node/2016. PRESS es una ontología formal diseñada para representar información bibliográfica acerca de recursos de publicaciones seriadas (revistas, periódicos, etc.). Su propósito es proponer respuestas a antiguos problemas con la aplicación de la familia de modelos FRBR a esas publicaciones seriadas y recursos continuos.

Metadatos, datos masivos y catálogos

Los datos masivos no se encuentran en realidad en los catálogos de la biblioteca, pero sí en toda la información asociada a ellos. Los catálogos de las colecciones de las bibliotecas poseen intrínsecamente una inmensa cantidad de datos vinculados entre sí que conforman una red de datos masivos no perceptible a simple vista. Ahí se encuentran embebidos innumerables autores –personas y organizaciones–, eventos, lugares, editoriales, épocas, temas, fechas, citas, etcétera. Entre todos, conforman un inmenso entramado de interrelaciones ocultas y difíciles de extraer. No existen como la simple suma de todos los registros; solo existen en el conjunto y en el contexto, y con frecuencia se vuelven masivos. Además, los catálogos de bibliotecas generalmente están separados físicamente por cada tipo de material: catálogos de libros, revistas y sus tablas de contenido, tesis, audio, imágenes, verticales, etcétera, lo que hace todavía más difícil percibir y establecer esas interrelaciones entre los datos de diferentes catálogos.

Al respecto, existe desde 2002 un "Protocolo de la Iniciativa de Archivos Abiertos para la Recolección de Metadatos" (*Open Archives Initiative Protocol for Metadata Harvesting*, OAI-PMH). Esta iniciativa marcó las pautas para la colecta, análisis e interrelación de metadatos internamente en una biblioteca, entre conjuntos de ellos, de editores, proveedores, etcétera. Numerosos proyectos mucho más evolucionados se han derivado desde entonces basados en esta iniciativa.

Como ejemplos relevantes de estos esfuerzos, se distinguen el proyecto de la Biblioteca Británica y el de la Biblioteca del Congreso de Estados Unidos, las cuales ya han comenzado a extraer los vínculos entre los datos de sus respectivas colecciones –las cuales comprenden millones de registros– para a partir de ellos modelar las interrelaciones entre personas, eventos, lugares, etcétera, contenidos en sus catálogos. El proyecto inglés se denomina "Modelo de datos para libros de la Biblioteca Británica" (British Library 2020) y el de la Unión Americana "Servicio de Datos Vinculados de la Biblioteca del Congreso de los Estados Unidos"

(Library of Congress s.f.). Hallo y colegas (2015) describieron ampliamente la teoría detrás de este concepto. Algunos editores están haciendo algo semejante. Por ejemplo, Springer Nature está desarrollando una iniciativa denominada *SN SciGraph* (Springer Nature s.p.) bajo el concepto de "datos abiertos vinculados". Básicamente consiste en un "descubridor" en ciencias naturales que compila datos de las publicaciones en este campo de esa editorial, además de las de otras instituciones académicas asociadas. Su base de datos recopila ya información de artículos de investigación, investigadores, libros y capítulos, instituciones, conferencias, citas, etcétera, construyendo vínculos semánticos entre todos ellos. El proyecto aspira a compilar eventualmente a dos mil millones de ítems interrelacionados.

Existen otros proyectos de bibliotecas que ya están agregando grandes conjuntos de metadatos adicionales a sus catálogos, optimizando radicalmente sus buscadores. Como un ejemplo de ello, se agregan la tabla de contenido y el glosario cada libro, interrelacionando estos datos con el registro catalográfico original. Esto optimiza sobremanera la búsqueda, pues el buscador no dispone ya solamente de las palabras del autor, título o tema; cuenta además con muchas más palabras adicionales contenidas en el índice y glosario de cada libro. Estos modelos de datos interrelacionados permiten que el buscador informe al usuario que un cierto ítem buscado en uno de los catálogos, como autor, tema o título, aparece en otros catálogos, o que ese ítem es citado en otros textos: libros, artículos, tesis, etcétera.

El buscador puede así informar también que las personas que consultaron un cierto ítem también consultaron otros relacionados, apuntando hacia ellos: las posibilidades se vuelven así infinitas sin perder precisión. Con las correspondientes variantes y adecuaciones, esto puede hacerse en colecciones especializadas en todas las disciplinas, como literatura y teatro, química, matemáticas, etcétera, adaptándolo al contexto y características de cada una de ellas.

El punto central de todo ello es que hay conjuntos de metadatos adicionales agregables a los catálogos que pueden potenciarlos

enormemente. Pero obviamente esa extracción de datos, sus interrelaciones y su agregación no pueden ser elaborados por métodos manuales; ni siquiera con técnicas informáticas básicas; requieren de un tratamiento especial que cae ya en el campo de los datos masivos por su volumen, su velocidad de generación y sus complejas estructuras.

Estos son algunos ejemplos aplicados de las posibilidades de lo que puede lograrse en la práctica con el uso de datos masivos y metadatos para la optimización mayor de catálogos en bibliotecas. Para abundar en el tema de cómo trasformar metadatos en datos vinculados dentro de las bibliotecas, véase Schilling (2012).

Metadatos, metría y bibliotecas

Los datos masivos se encuentran también relacionados con la biblioteca en los estudios métricos de la información documental, en todas sus especialidades: bibliometría, archivometría, informetría, bibliotecometría, así como en otras asociadas: cienciometría, webmetría y altmetría. Todas tienen como común denominador la aplicación de modelos y métodos matemáticos y estadísticos a las actividades bibliotecaria, bibliográfica, archivística, las redes sociales, la investigación en ciencias y humanidades, su comunicación y divulgación, entre muchas otras. Son otro ejemplo clásico de la minería de datos aplicada.

Como es sabido, a partir de la extracción de datos de diversos ítems de literatura científica, estas disciplinas permiten obtener información significativa para diversos estudios acerca del desarrollo y comportamiento de alguna o algunas actividades científicas, académicas y editoriales con el fin de estudiarlas, planearlas, optimizarlas, etcétera. Las diversas variantes de los estudios métricos de la información permiten estudiar diferentes fuentes y modalidades de la misma. El punto central de esto consiste en que el común denominador de todas ellas es la extracción de grandes cantidades de datos para lograr su propósito. Al clasificarse, ordenarse y agruparse, eventualmente muchos de esos datos se convierten en metadatos como parte de los resultados.

Entre estas técnicas, se distingue la bibliominería, la cual es la combinación de técnicas de minería de datos, almacenamiento de datos y bibliometría, utilizadas específicamente para analizar colecciones y servicios bibliotecarios. Al igual que las anteriores, su base proviene de la extracción y el manejo de grandes cantidades de datos que eventualmente se convierten en metadatos.

Entre la minería de datos aplicada a los documentos, resalta la "minería de textos" o "análisis de textos". Con la utilización de la minería de datos, técnicas estadísticas, lingüísticas, de aprendizaje de máquina, de comprensión del lenguaje natural, recuperación de información y razonamiento basado en casos, este tipo de estudios ayuda a las organizaciones a obtener nuevos conocimientos extrayendo información significativa proveniente de grandes cantidades de textos documentales no estructurados disponibles en la Internet y en las intranets corporativas, utilizando elementos tan variados como el análisis lexicográfico y semántico, agrupamientos, categorizaciones y taxonomías; vínculos, relaciones y asociaciones entre entidades; análisis de sentimientos o minería de opiniones, frecuencia de palabras, etcétera. Por lo mismo, sus aplicaciones son muy variadas: identificación de textos, extracción de elementos de ellos, categorización y/o taxonomía de textos, extracción de conceptos, entidades, relaciones, eventos; traducción de textos, por citar algunas.

Como puede observarse, los metadatos son una entidad siempre existente y estrechamente relacionada con todos los estudios de metría y de minería de datos en las bibliotecas, y por lo mismo es indispensable entenderlos y saberlos manejar. No es posible hacer metría sin datos, e igualmente no es posible hacerlo bien sin metadatos.

Las bibliotecas como grandes repositorios de datos

A lo largo de este siglo, las comunidades académicas y de información cayeron en cuenta de que los datos recopilados durante las investigaciones científicas, periodísticas, sociales, etcétera, tenían un valor agregado después de concluidos sus proyectos, pues podían

ser reutilizados posteriormente por otras personas, ya que obviamente un cierto conjunto de datos compilados puede ser analizado desde múltiples enfoques y metodologías por grupos diferentes, y podían extraerse así nuevos resultados de esos datos. A partir de este enfoque, los datos no son ya únicamente materia prima para producir información, sino un objeto de información en sí mismos, con un valor propio e intrínseco; derivado de ello, requieren de un tratamiento específico.

Además, los organismos gubernamentales de financiación académica han comenzado a requerir cada vez más que tanto los resultados como los datos de las investigaciones se hagan públicos de forma abierta; esto ha sido resumido en el proyecto Open Data. Pero hacer públicos grandes conjuntos de datos requiere de método y normalización. Súbitamente, las comunidades de investigación se vieron obligadas a comenzar a gestionar sus datos de una manera sistematizada y estandarizada para estar en posibilidad de preservarlos y hacerlos accesibles.

Las instituciones de investigación –en especial aquellas en universidades– se vieron ante la necesidad de comenzar a crear repositorios de datos de sus proyectos. Los investigadores se encontraron sin suficiente tiempo, habilidades y recursos para manejar de esta forma sus datos durante sus proyectos, con el problema adicional de encontrar depósitos apropiados para sus datos. En consecuencia, muchas instituciones acudieron a sus bibliotecas para asesoría y para que ellas comenzaran a alojar esos conjuntos de datos, y así se integraron los repositorios de datos a las colecciones usuales de las bibliotecas. Ello representaba un reto inédito. Diversas organizaciones bibliotecarias comenzaron a esbozar estos nuevos retos, como la Asociación de Bibliotecas Universitarias y de Investigación de Estados Unidos, Association of College and Research Libraries (ACRL), la cual es una subdivisión de la American Library Association (ALA) (Tenopir *et al.* 2012 y 2015), y también la Liga de Bibliotecas Europeas de Investigación" (LIBER) (Tenopir *et al.* 2017).

La IFLA ha realizado también estudios detallados de los temas relacionados con el uso de repositorios de datos en bibliotecas; en su revista, en el último fascículo del año 2016 y primero del 2017

(IFLA *Journal* 2016 y 2017), compiló alrededor de 20 textos y reflexiones al respecto, dividiéndolos en cuatro grandes temas: 1) las necesidades de los investigadores, 2) las habilidades requeridas de los bibliotecarios, 3) los posibles servicios a ofrecer y 4) la alfabetización en datos. A partir de estos estudios preliminares, la IFLA creó una iniciativa al respecto llamada Proyecto del Curador de Datos (Data Curator Project) (IFLA s.f.), cuyo objetivo principal fue determinar las funciones y responsabilidades de los profesionales bibliotecarios que ya trabajaban en ello en diversos países. El estudio se centró además en la terminología utilizada para describir las prácticas emergentes y las nuevas funciones profesionales. Witt y Horstmann (2016) establecieron que las actividades centrales requeridas a los bibliotecarios al respecto son: 1) ayudar a los investigadores a entender y resolver las necesidades a lo largo del ciclo de vida de los datos de las investigaciones; 2) asesorar en la construcción de planes de gestión de datos y metadatos; 3) diseñar soluciones de publicación y conservación de datos; 4) crear guías y tutoriales web para capacitar a investigadores y usuarios; 5) alojar y mantener repositorios en sus acervos. Como puede verse, el punto dos de esta lista hace especial énfasis en la gestión de metadatos por parte de los bibliotecarios en los repositorios.

Todas estas nuevas necesidades, conceptos y propuestas crearon una nueva especialidad en el campo de la información denominada "Gestión de Datos de Investigación" o Research Data Management (RDM). Whyte y Teds (2011, s. p.) la definen como "[…] la organización de los datos, desde su entrada en el ciclo de investigación hasta la difusión y el archivado de los resultados valiosos".

En general, la RDM tiene que ver con los aspectos relativos a los datos de investigación: su ciclo de vida, sus colecciones; colecta, depuración, coherencia y normalización; sus formatos, metadatos, los repositorios y servicios de consulta de datos, anonimización y seguridad de datos, su preservación, las habilidades y funciones requeridos para su gestor, alfabetización en datos para investigadores y hasta como citarlos. Pinfield y colegas (2014) establecieron siete grandes campos de desarrollo o "impulsores" para el estudio de

la Gestión de Datos de Investigación: almacenamiento, seguridad, preservación, cumplimiento de políticas y leyes, calidad, difusión y compromiso.

Como puede observarse de lo anterior, la "gestión de datos" –especialmente los de investigación– es una tarea multidisciplinaria, pero obviamente los bibliotecarios deben estar entre los profesionales que los manejan. Esto requiere de nuevos conocimientos, capacitación y entrenamiento, pero ciertamente ellos cuentan con bases profesionales para esta tarea.

Inteligencia artificial y sistemas expertos

Otro campo de acción del análisis de datos en la biblioteca que tiene estrecha relación con el uso de metadatos es el de la Inteligencia Artificial (IA). Este concepto se atribuye a Arthur Samuel (Science Direct, s. p.) en 1959, pionero en juegos de computadora y en la IA. Él lo definió básicamente como "la capacidad de las computadoras de aprender sin necesidad de una programación explícita". En general, este campo tiene muchas áreas de acción, y el imaginario popular tiende a asociarlo con robots dentro de la biblioteca. Siendo esta una aplicación válida, no es la más utilizada en ellas. Existen otros subcampos de la IA que son más utilizados en la actualidad en este tipo de instituciones. Entre ellos destacan el denominado "aprendizaje de máquina" –*machine learning*– también llamado "aprendizaje automático". Consiste en diseñar y programar un cierto sistema específico para que sea susceptible de ser enseñado, entrenado o preparado para realizar diversas acciones opcionales sin la intervención humana directa. Estos sistemas específicos reciben datos que pueden interpretar y extraer patrones significativos de ellos; dependiendo de esos datos y sus interpretaciones, un cierto sistema responderá en una u otra forma. Obviamente, entre más datos existan, la respuesta será más precisa. El aprendizaje de máquina es similar a la minería de datos en el sentido en que ambos son técnicas para explorar grandes conjuntos de datos con el fin de descubrir patrones y correlacio-

nes. La diferencia principal se encuentra en que el aprendizaje de máquina se extiende hasta la predicción de patrones y no tan solo a su descubrimiento.

El aprendizaje de máquina es usado hoy en día no tan solo en bibliotecas, sino en toda la industria relacionada con Bibliotecas y Servicios de Información –*Library and Information Services* o LIS– para muy diversos propósitos: indización, catalogación, clasificación, recuperación de información en línea, elaboración de resúmenes, servicios de referencia, tablas de contenido, traducción, etcétera.

Otra de las aplicaciones prácticas comunes de la IA en las bibliotecas son los "Sistemas Expertos". De hecho, ésta es la aplicación práctica de IA en bibliotecas más antiguas, pues su uso y documentación data de los ochenta. Son programas informáticos que utilizan principios y métodos de la Inteligencia Artificial para resolver problemas en un campo especializado, los cuales por lo general requerirían de la experiencia de personal experto. Incorporan los conocimientos acumulados por personas avezadas en un tema y se diseñan para funcionar lo más parecido a ellas. Generalmente, contienen una *base de conocimientos* de hechos y relaciones representados en forma de datos y metadatos, y tienen capacidad de hacer inferencias basadas en ellos. Quienes diseñan y construyen estos sistemas utilizan diversas técnicas para la construcción de esa base de conocimientos, tales como la descripción verbal de tareas realizadas por una persona, el análisis de protocolos y procedimientos escritos, cuestionarios, entrevistas y encuestas, el descubrimiento y la documentación del conocimiento tácito dentro de la organización, así como la observación de procesos y su simulación. Algunos ejemplos clásicos de los Sistemas Expertos en las bibliotecas son los "Descubridores" de información, y el uso de "lenguaje natural" por parte de los usuarios en sus búsquedas.

Análisis del aprendizaje

Continuando con las aplicaciones de IA en bibliotecas –relacionadas con datos y metadatos–, se encuentra un rubro que ha despertado

gran interés en años recientes: el denominado "Análisis del aprendizaje" (*Learning Analytics)*. Si bien este tema es de interés en general de las instituciones académicas, por su cercanía con las bibliotecas es desarrollado con frecuencia dentro de ellas (Baepler y Murdoch 2010, 3). Este concepto se define como "[…] la medición, recopilación, análisis e información de datos sobre los alumnos y sus contextos, con el fin de comprender y optimizar el aprendizaje y los entornos en los que se produce" (First International Conference on Learning Analytics 2011, s.p.).

El propósito de este tipo específico de análisis de datos consiste en:

- Predecir; por ejemplo, detectar alumnos "en riesgo" en lo relativo a deserción o fracaso escolar; o lo contrario, identificar estudiantes con potencial o habilidades por encima del promedio.
- Personalizar y adaptar para dotar a los alumnos de métodos, herramientas y canales de aprendizaje y hasta de materiales de evaluación personalizados.
- Retroalimentar para evaluar el interés y la satisfacción de cursos, materiales educativos, servicios de información, técnicas y modalidades de instrucción, etcétera.
- Asesorar para dotar a los docentes de información pertinente y oportuna para tutelar y apoyar a los alumnos.
- Visualizar la información, básicamente en la modalidad de "tableros de aprendizaje" –*learning dashboards*, que proporcionen datos generales del aprendizaje a través de herramientas de visualización de datos.

Existe una variante adicional del Análisis del Aprendizaje, denominada "Análisis Académico" (*Academic Analysis)*, el cual también utiliza la ciencia de los datos y la IA. Se define como "[...] un área que combina datos institucionales, análisis estadísticos y modelos predictivos para crear inteligencia sobre la cual los estudiantes, instructores o administradores pueden influir y cambiar el comportamiento académico" (Baepler y Murdoch 2010, 3). Este

campo pretende establecer hasta dónde hay una correlación entre los estudiantes que consumen más material de biblioteca y su éxito académico. Young (2017) obtuvo durante años datos minuciosos acerca de cómo el uso de la biblioteca es comparable a otras métricas de éxito académico. Como resultado de ello, se demostró cuantitativamente que en realidad, un mayor uso de los recursos de la biblioteca incide significativamente en ese éxito. A partir de estos estudios, varias universidades ya han hecho cambios, como trasladar el departamento de asesoría estudiantil y el laboratorio de escritura hacia la biblioteca; estos cambios se concibieron tanto para atraer más alumnos a la biblioteca, como para hacer que la asesoría fuera más aceptable y eficiente para los estudiantes.

Conclusiones

Los campos y aplicaciones anteriormente mencionados no son una lista exhaustiva. Son simplemente una muestra representativa de todo lo que se puede hacer y se está haciendo hoy en día con los datos –en especial los masivos– en las bibliotecas. Los datos se usan hoy en día en innumerables negocios y organizaciones de todo tipo de sectores económicos, industriales y académicos. Dentro de este último sector, se encuentran las bibliotecas, las cuales también son susceptibles de beneficiarse de este desarrollo. Las organizaciones bibliotecarias multiinstitucionales a lo largo de todo el mundo: IFLA, ALA, ARL, JISC, etcétera, han tomado conciencia de la importancia de los datos dentro de su entorno y por eso han creado grupos de interés y estudio acerca del tema y han llegado a la conclusión de que es indispensable que las bibliotecas formen parte proactiva de este fenómeno.

Como ha podido observarse, los datos masivos se usan ya en bibliotecas en:

- Diversas taxonomías de la información: ontologías, tesauros, etcétera.
- Análisis, extracción y agregación de grandes conjuntos de datos adicionales a sus catálogos, lo que optimiza radicalmente sus buscadores.

- Estudios métricos de la información documental en todas sus especialidades: bibliometría, archivometría, informetría, bibliotecometría, así como en otras asociadas: cienciometría, webmetría, altmetría. Tienen como común denominador la aplicación de modelos y métodos matemáticos y estadísticos a las actividades bibliotecaria, bibliográfica, archivística, las redes sociales, la investigación en ciencias y humanidades, su comunicación y divulgación, etcétera.
- La minería de textos o análisis de textos para identificación de textos, extracción de elementos de ellos, categorización y/o taxonomía de textos, extracción de conceptos, entidades, relaciones, eventos; traducción de textos, etcétera.
- El análisis de datos con técnicas de IA, como el "aprendizaje de máquina". En él se diseñan y programan sistemas específicos para que sean susceptibles de ser enseñados, entrenados o preparados para realizar diversas acciones opcionales sin la intervención humana directa.
- La industria relacionada con Bibliotecas y Servicios de Información (*Library and Information Services* o LIS) para muy diversos propósitos: indización, catalogación, clasificación, recuperación de información en línea, elaboración de resúmenes, servicios de referencia, tablas de contenido, etcétera.
- Sistemas expertos, que tratan problemas como la indización basada en el conocimiento, el procesamiento de lenguaje natural, la catalogación, la consulta recuperación de información, el conocimiento tácito en la biblioteca, etcétera.
- El diseño, implementación y mantenimiento de repositorios de datos de investigación.

Como se ha constatado a lo largo de este texto, el punto central es que:

- No es posible una buena gestión de datos –en especial los masivos– sin el diseño, la inclusión y el manejo de adecuados, y suficientes metadatos: ésta es la clave del éxito en esa gestión.

- La gestión de datos –en especial los masivos– requiere de técnicas, herramientas y profesionales adecuados al respecto.
- Existe un enorme déficit de expertos en datos y sus metadatos a nivel mundial.
- Es una tarea multidisciplinaria, pero sin duda los bibliotecarios tienen la formación y los antecedentes profesionales adecuados para ello.

La gestión de datos, los datos masivos y sus correspondientes metadatos representan a la vez un reto y una oportunidad para la biblioteca y su personal profesional: por un lado, implica que ese personal bibliotecario debe adquirir nuevos conocimientos, capacitación, habilidades y entrenamiento para su correcto manejo. Por otro, representa nuevas oportunidades de actividad profesional altamente calificada para el personal bibliotecario. Además, ofrece la oportunidad de reposicionar a la biblioteca dentro de las responsabilidades y los quehaceres contemporáneos de su comunidad. Estas actividades han ido creciendo en la última década de forma notable y requieren de estructura organizacional y personal calificado para realizarlas adecuadamente.

Referencias bibliográficas

Baepler, Paul y Cynthia Murdoch. 2010. "Academic Analytics and Data Mining in Higher Education". En: *International Journal for the Scholarship of Teaching and Learning,* vol. 4, núm. 2 doi:10.20429/ijsotl.2010.040217.

British Library. 2020. British Library Data Model. Disponible el 15 de marzo de 2022 en British Library Data Model – Books. https://www.bl.uk/bibliographic/pdfs/bldatamodelbook.pdf.

First International Conference on Learning Analytics and Knowledge (Proceedings of LAK 11 (2011). Nueva York: As-

sociation for Computing Machinery. https://dl.acm.org/doi/proceedings/10.1145/2090116.

Google. Entrada por Big Data. Disponible en abril de 2021 en http://www.google.com.

Hallo, Maria; Luján-Mora, Sergio; Maté, Alejandro y Juan Trujillo. 2015. "Current state of Linked Data in digital libraries". En: *Journal of Information Science,* vol. 42, núm. 2. DOI: 10.1177/0165551515594729.

Hey, Tony; Tansley, Stewart y Kristin Tolle (Eds.). 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, Wa.: Microsoft Research. https://digital.library.unt.edu/ark:/67531/metadc31516/m2/1/high_res_d/4th_paradigm_book_complete_lr.pdf.

IDC Corp. 2014. *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things.* https://www.emc.com/leadership/digital-universe/2014iview/index.htm.

IFLA. *Journal,* vol. 42, núm. 4 (2016), https://www.IFLA.org/publications/node/1691.

————— . *Journal*, vol. 43, núm. 1 (2017).

————— . "The Data Curator Project". Disponible el 16 de marzo de 2020 en https://www.IFLA.org/library-theory-and-research/projects.

Laney, Doug. 2001. "3D Data Management: Controlling Data Volume, Velocity, and Variety". En: *Application Delivery Strategies*, File 949. Meta Group, 6 de febrero de 2001. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Library of Congress. Linked Data Service. Disponible el 15 de marzo de 2022 en https://id.loc.gov/.

Naur, Peter. 1975. *Concise Survey of Computer Methods*. Studentlitteratur: Lund, Suecia. Citado por: Press, Gil. 2013. "A Very Short History of Data Science". En: *Revista Forbes* https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#64a5f94455cf.

Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). 2015. https://www.openarchives.org/OAI/openarchives-protocol.html

Pinfield, Stephen; Cox, Andrew y Jen Smith. 2014. "Research Data Management and Libraries: Relationships, Activities, Drivers and Influences". En: *PLOS ONE*, vol. 9, núm. 12 https://doi.org/10.1371/journal.pone.0114734.

Schilling, Virginia. 2012. *Transforming Library Metadata into Linked Library Data*, Chicago: American Library Association, September 25, 2012. http://www.ALA.org/alcts/resources/org/cat/research/linked-data.

Schmarzo, Bill. 2018. "Importance of Metadata in a Big Data World". En: *Data Science Central Blog*. Entrada del 23 de julio de 2018. https://www.datasciencecentral.com/profiles/blogs/importance-of-metadata-in-a-big-data-world.

Science Direct. Entrada por: Machine Learning. Disponible en abril de 2021 en https://www.sciencedirect.com/topics/psychology/machine-learning.

Springer Nature. SN SciGraph. A Linked Open Data Platform for the Scholarly Domain. Disponible el 14 de marzo de 2022 en https://www.springernature.com/gp/researchers/scigraph.

Tenopir, Carol; Birch Ben y Suzie Allard. 2012. *Academic libraries and research data services: Current practices and plans for the future*. ACRL http://www.ALA.org/ACRL/sites/ALA.org.ACRL/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf.

Tenopir, Carol *et al*. 2015. "Research Data Services in Academic Libraries: Data Intensive Roles for the Future?" En: *Journal of eScience Librarianship*, vol. 4, núm. 2. https://escholarship.umassmed.edu/jeslib/vol4/iss2/4/.

———————. 2017. "Research Data Services in European Academic Research Libraries". En: LIBER *Quarterly*, vol. 27, núm. 1: 23-44. DOI: http://doi.org/10.18352/lq.10180

Whyte, Angus y Jonathan Tedds. 2011. *Making the case for research data management. dcc* Briefing Papers. Edinburgo: Digital Curation Centre. https://www.dcc.ac.uk/guidance/briefing-papers/making-case-RDM.

Witt, Michael y Wolfram Horstmann. 2016. "International approaches to research data services in libraries". En*:* IFLA *Journal*, vol. 42, núm. 4, pp. 251–252. DOI: 10.1177/0340035216678726

Young, Jeffrey. 2017. "Libraries Look to Big Data to Measure Their Worth - and Better Help Students". En: *Digital Learning in Higher Education*. Entrada del 17 de Noviembre de 2017. Disponible en https://www.edsurge.com/news/2017-11-17-libraries-look-to-big-data-to-measure-their-worth-and-better-help-stents.

Contenidos digitales: convergencia, conectividad, modelos y nuevas características. Instituto de Investigaciones Bibliotecológicas y de la Información/UNAM. La edición consta de 100 ejemplares. Coordinación editorial Anabel Olivares Chávez; corrección especializada, Valeria Guzmán González; revisión de pruebas Valeria Guzmán González y Carlos Ceballos Sosa; formación editorial, Nube Magenta. Fue impreso en papel cultural de 90 gr. en los talleres de Litográfica Ingramex, Centeno 162-1, Col. Granjas Esmeralda, Alcaldía Iztapalapa, CDMX, C. P. 09810. Se terminó de imprimir en junio de 2022.