Datos masivos en bibliotecas / Big data in libraries

JUAN VOUTSSÁS M.

Atribución-NoComercial-CompartirIgual 4.0 Internacional (CC BY-NC-SA 4.0)

Este es un resumen legible por humanos (y no un sustituto) de la licencia. Advertencia.

Usted es libre de:

Compartir — copiar y redistribuir el material en cualquier medio o formato

Adaptar — remezclar, transformar y construir a partir del material

La licenciante no puede revocar estas libertades en tanto usted siga los términos de la licencia

Bajo los siguientes términos:

Atribución — Usted debe dar crédito de manera adecuada, brindar un enlace a la licencia, e indicar si se han realizado cambios. Puede hacerlo en cualquier forma razonable, pero no de forma tal que sugiera que usted o su uso tienen el apoyo de la licenciante.

NoComercial — Usted no puede hacer uso del material con propósitos comerciales.

CompartirIgual — Si remezcla, transforma o crea a partir del material, debe distribuir su contribución bajo la la misma licencia del original.

Datos masivos en bibliotecas /
Big Data in Libraries

COLECCIÓN
SISTEMATIZACIÓN DE LA INFORMACIÓN DOCUMENTAL
Instituto de Investigaciones Bibliotecológicas y de la Información

Datos masivos en bibliotecas / Big Data in Libraries

Juan Voutssás Márquez

Universidad Nacional Autónoma de México
2022

Z678.93 Voutssás Márquez, Juan.
B54V68 Datos masivos en bibliotecas = Big Data in Libraries / Juan Voutssás Márquez. - México : UNAM. Instituto de Investigaciones Bibliotecológicas y de la Información, 2022.

418 p. – (Sistematización de la información documental)
ISBN: 978-607-30-6397-5

1. Big data. 2. Datos masivos. 3. Bibliotecas. I. Título.

Diseño de cubierta: Mario Ocampo Chávez

Primera edición: agosto 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: 978-607-30-6397-5

Publicación dictaminada
Impreso y hecho en México

Contenido

Contents

Introducción

Los hechos son ontológicos, la evidencia es epistemológica, los datos son retóricos.

DANIEL ROSENBERG,
2013. *Data before The Fact*

El concepto de *"big data"* (datos masivos) no es la evolución vertical de una cierta idea o tecnología a lo largo de las décadas, sino la conjunción simultánea de múltiples fenómenos, necesidades, tecnologías, teorías, herramientas y métodos relacionados con la información que al concurrir en un cierto punto se convierten en algo más complejo. Tiene múltiples antecedentes, facetas y componentes, y se puede analizar desde diversos enfoques; por lo mismo, para entender el concepto es necesario revisar los principales entre ellos para poder obtener una visión completa de su esencia.

Durante la segunda mitad del siglo XX muchos autores –como por ejemplo Derek de Solla Price, F. W. Lancaster o Alvin Toffler– mencionaron y estudiaron el fenómeno denominado "explosión de la información" (*information explosion).* Como es sabido, básicamente el término se refiere al enorme e incesante crecimiento de la información publicada, a los problemas que este volumen ha representado para su adecuado manejo, así como a los efectos sociales, técnicos, económicos, etcétera, derivados del fenómeno. De acuerdo con el *Diccionario Oxford* (s.d.), en el diario

"Constitución" de Lawton, Oklahoma, del 30 de noviembre de 1941 se utilizó por primera vez este término para referirse al enorme crecimiento de la información global. Para la década de los sesenta comenzó a generalizarse: en 1960 el diario *Oshkosh Daily Northwestern Newspaper* (1960, 19) del estado de Wisconsin, Estados Unidos, menciona ya el concepto en su acepción completa "[…] La explosión de información en la que solo el campo de la ciencia estima que la cantidad de material disponible se ha duplicado en los últimos 10 años". Durante la década el término aparecería cada vez más en documentos técnicos y académicos.

En esa época se introdujo también el término "ciencia de la información", como sucesor de los términos "comportamiento informativo" (*information behavior*), "estudios de usuarios" (*user studies)*, y "documentación" (*documentation)*. Alvares y Araújo (2010, 200) mencionan que "[…] La Conferencia Internacional sobre Información Científica, celebrada en Washington en 1958, marcó la transformación de la 'documentación' hacia la 'ciencia de la información'". Tefko Saracevic también sitúa en los sesenta el cambio de denominaciones hacia "ciencia de la información" en su reseña histórica de esta disciplina (Saracevic 1992, 5-27*)*. Él mismo hizo una definición del término tan temprano como 1967:

> [la ciencia de la información] se relaciona con las propiedades, comportamiento y circulación de la información. Abarca el análisis de los sistemas, los aspectos mesológicos de la información y la comunicación, de los medios de información y del análisis lingüístico, de la organización de la información, de las relaciones hombre-sistema, etcétera (Rees y Saracevic 1967, 2).

En algún momento de la segunda mitad del siglo XX, con el advenimiento de la era digital, el concepto de "explosión de la información" fue evolucionando a "sobrecarga de información" (*information overload)*, básicamente para hacer referencia –además de la inmensa cantidad de información producida– a la situación que enfrenta un sistema, ya sea computacional, social, etcétera, cuando la cantidad de información excede su capacidad

de procesamiento.[1] Hacia finales del siglo, los conceptos de "explosión de la información" "sobrecarga de información" y "ciencia de la información" se fueron extendiendo también a los datos; posteriormente y como consecuencia, estos generarían sus propias especialidades y contextos.

Los datos siempre han sido la fuente primaria de la información; básicamente, los datos ensamblados en contexto se convierten en información con el fin ulterior de inferir conocimiento. No obstante, durante siglos el énfasis estuvo en la información, y los datos eran únicamente una materia prima en su proceso. Como productos, se utilizaron mayormente piezas de información "terminadas"; es decir, el resultado concluido del análisis y síntesis de ciertos datos por parte de una persona o grupo en forma de una publicación. Estos productos se manifestaban en forma de libros, artículos en revistas académicas, textos periodísticos, manuales, tesis, compendios, patentes, etcétera. En la segunda mitad del siglo XX el énfasis absoluto en la información se fue perdiendo conforme los datos ganaban preponderancia al hacerse evidente que estos tenían inherentemente un valor adicional agregado y que las organizaciones podían generarlos y/o colectarlos para la mejor toma de decisiones y no tan solo con fines de producir publicaciones.

Por supuesto, ese cambio de énfasis entre información y datos no fue instantáneo: se fue dando de manera gradual durante las últimas décadas del siglo pasado. No obstante, un punto de inflexión interesante al respecto puede observarse en una nota aparecida en el diario *Washington Post* de la Unión Americana en febrero de 1989. Ahí se lee:

> […] según una estimación, en las últimas tres décadas se ha producido más información nueva que en los cinco milenios anteriores. La cantidad total de conocimientos impresos se duplica cada ocho años [....] ¿Cuál es el resultado? La ansiedad por la información,

1 Este concepto de "sobrecarga de información" proviene de las ciencias sociales, y se atribuye a Bertram Gross en 1964 en su texto *The managing of organizations: The administrative struggles*.

> descrita como 'el agujero negro entre los datos y el conocimiento'. ¿La diferencia entre datos e información?: los datos son la materia prima y son pasivos; la información es activa y –en el mejor de los casos– esclarecedora. A medida que nos sumergimos en la sobreabundancia de los primeros, la segunda se hace cada vez más difícil de alcanzar (Streitfeld 1989).

Es conveniente subrayar cómo en la época de esta nota los datos todavía "son materia prima y son pasivos", y la información "es activa y esclarecedora". De ello se concluye que en ese año ya se consideraba que había una sobreabundancia de datos e información, pero aquellos todavía no alcanzaban los niveles de ponderación actuales.

Por otro lado, a fines de la primera década de este siglo se llegó también a la conclusión generalizada de que los datos recopilados a lo largo de una investigación académica debían ser guardados de una forma sistematizada después de concluido el proyecto para poder ser reutilizados posteriormente, ya que un cierto conjunto de datos compilados podía ser analizado de múltiples maneras por diversos grupos y podían extraerse así nuevos resultados de esos datos. Los datos no eran ya únicamente materia prima para producir información; eran un objeto y un producto en sí mismos, con un valor propio e intrínseco, y por eso requerían de un tratamiento específico.

Ese tratamiento cayó dentro de lo que ya se denominaba "gestión de datos" (*data management)*: una serie de teorías, principios, modalidades, métodos, herramientas, tecnologías, etcétera, para su manejo y uso. Surgieron además especialidades dentro de ella, tales como la "ingeniería de datos" (*data engineering)*, que trata la organización y recuperación de datos, y tiene que ver con cuán intrínsecamente limpios y estructurados están los datos dentro de un cierto conjunto de ellos. El "análisis de datos" (*data analytics*) es el trabajo de identificar cuáles variables de la organización pueden ser relacionadas con ciertos datos para el planteamiento de preguntas y la eventual obtención de soluciones a través de técnicas específicas. Todo ello fue conformando la "ciencia

de los datos" (*data science)*, la cual consiste en el estudio de esos datos organizados para identificar aquellos que son importantes en el contexto de un problema específico o un cierto modelo de negocio; también tiene que ver con el desarrollo de modelos y algoritmos que resuelven problemas a gran escala en las organizaciones. No debe confundirse la ciencia de la información con la ciencia de los datos. Si bien ambas se complementan y traslapan, no son sinónimos. Básicamente, la *ciencia de la información* es una ciencia interdisciplinaria que estudia las propiedades, comportamiento y flujos de la información; su cuerpo de conocimiento abarca la creación, colecta, organización, almacenamiento, recuperación, diseminación, utilización y preservación de la información registrada, como recursos documentales organizados. La *ciencia de los datos* es una ciencia interdisciplinaria que tiene que ver con el descubrimiento de conocimiento significativo y de información utilitaria a partir de datos. Además de los objetos de estudio, gran cantidad de autores coinciden en que otra diferencia se observa en las etapas de desarrollo: la *ciencia de la información* es una evolución vertical de la *documentación*, la cual como disciplina fue gradualmente cruzando nuevas fronteras de otras áreas del saber humano y agregando de ellas nuevo conocimiento hasta convertirse en lo que ahora es. La *ciencia de los datos* es la conjunción de múltiples disciplinas muy diferentes entre sí que evolucionaron paralelamente hasta que se unieron en algún punto dando origen a una totalmente nueva visión de los datos.

En los primeros años de este siglo, el investigador de la empresa Microsoft Jim Gray y sus colaboradores plantearon que la ciencia contemporánea llegaba a su cuarto paradigma. Se había aceptado por siglos que la ciencia se basaba en sus dos paradigmas fundamentales: la teoría y la experimentación. Con la llegada de las teorías de John von Neumann y las computadoras en la segunda mitad del siglo XX se había integrado como tercer paradigma la simulación y modelado con estos equipos. Gray estableció que llegaba ya un cuarto paradigma para la ciencia, que complementaba a los otros tres: los datos. La ciencia era ya una ciencia basada intensamente en ellos. Se requería, por tanto, una nueva

generación de conceptos, metodologías, herramientas y expertos para contender con ello. Hey y sus colegas (2009) editaron en ese año la primera antología al respecto, considerada la piedra angular de esta visión de la ciencia moderna en relación con los datos. Hey también fue uno de los primeros en señalar desde 2006 que las facetas emergentes de la ciencia de ese entonces –e-Ciencia o ciencia electrónica, ciencia abierta, etcétera– establecían una nueva relación entre la ciencia y la biblioteca debido precisamente a los datos (Hey y Hey 2006, 525-526); algo semejante apuntó también en ese año Carlson (2006). Las bibliotecas y los datos quedaban formalmente unidos desde entonces. Desde este particular enfoque de los datos emanados de investigaciones científicas, López-Yepes (2004, 59 y 411) los definió así:

> […] los datos constituyen la unidad de análisis básica de contenidos en este tipo de información, los cuales representan el testimonio o expresión mínima de un hecho mensurable […] son el punto de partida, los hechos y principios indiscutidos para una investigación científica. Es la experiencia sensible en el empirismo […] un conjunto de datos forman una información.

Otro de los grandes factores que aportaron al desarrollo de los datos fue el inusitado crecimiento de la red mundial y de las telecomunicaciones a partir de la década de los noventa, el cual generó un nuevo y creciente desarrollo de la información en su forma digital. Se crearon millones de piezas de información y de datos en esta modalidad, las cuales se sumaron a lo ya existente, lo cual multiplicó exponencialmente la cantidad de información acumulada.

Durante la última década, el enorme crecimiento de diversos sectores de la Internet ha ido creando un flujo inusitado de datos, en especial el crecimiento de las redes sociales. Originalmente, los datos básicos de las personas que se suscribían a ellas eran conjuntos simples con un valor económico, los cuales eran vendidos a empresas u organizaciones con fines de publicidad o mercadeo. No tardaron los dueños de estas redes en darse cuenta de que la interacción de millones de personas cada día produce enormes

flujos de datos, los cuales podían ser analizados para extraer de ellos nueva información adicional, la cual a su vez también tenía un nuevo valor económico y podía ponerse a la venta.[2] Aun cuando un usuario no proporcione explícitamente ningún dato acerca de él, el uso de las redes implica dejar una huella de datos, tales como su ubicación, preferencias, selecciones, compras o ventas, tiempo invertido, lugares visitados, el historial de búsqueda, etcétera. En la segunda década de este siglo se agregó a esto el fenómeno conocido como el "Internet de las Cosas" (*Internet of Things*, IdC o IoT), el cual consiste en que numerosos objetos cotidianos –más allá de computadoras y teléfonos inteligentes– se conectan a la Internet para intercambiar datos con otros dispositivos y/o sistemas de forma automática sin mediar instrucción expresa: paquetería, electrodomésticos, GPS, mercancía en un almacén, termostatos, semáforos, cámaras, etcétera. Del intercambio de datos puede derivarse o no una cierta acción; un capítulo especial de estos son los dispositivos personales (*wearables),* también interconectados a la Internet; es una categoría especial que se refiere a los dispositivos digitales de uso personal que colectan e intercambian datos dentro de la red: relojes de pulso, implantes médicos, pulseras rastreadoras de acondicionamiento físico o *fitness trackers*, localizadores de personas y/o mascotas, anteojos de realidad aumentada, audífonos, etcétera.

Todos estos fenómenos fueron agregando volumen a la información y a los datos. El enorme crecimiento de las redes sociales y el Internet de las Cosas agregaron cantidades inéditas de datos a lo ya existente que fueron aportando al concepto de *big data* (datos masivos, macrodatos, inteligencia de datos, grandes datos o datos a gran escala). Todos estos son términos que hacen referencia al concepto de "conjuntos de datos extremadamente grandes que pueden ser analizados computacionalmente para revelar patrones, tendencias y asociaciones, especialmente relacionadas con

2 Como ejemplo, el sitio InternetLiveStats consigna que en el año 2020 se enviaron más de 200 mil millones de tuits, y que Facebook tiene en este momento 2,500 millones de usuarios.

el comportamiento humano y sus interacciones" (Lexico 2014). A lo largo de este texto, se utilizará el término "datos masivos" para hacer referencia en español del concepto de *big data*. Como ha podido observarse, el concepto de los datos masivos no es la evolución vertical de una cierta idea a lo largo de las décadas, sino la conjunción de múltiples fenómenos, teorías, métodos, tecnologías, etcétera, que desembocaron en algo más complejo. Se abundará en ello más adelante.

Como muchas otras tecnologías, los datos masivos fueron presentados y tratados desde su inicio y por años como la gran panacea del manejo de datos, y por lo mismo crearon en personas y organizaciones demasiadas expectativas. En palabras de Jakob Nielsen, "[…] los dos errores clásicos que se comenten al predecir el futuro de un cambio tecnológico son sobrestimar su impacto a corto plazo y subestimar su impacto a largo plazo". Actualmente, la euforia de los datos masivos comienza a descender y a tomar sus dimensiones y perspectivas reales, y por eso éste es buen momento para analizarlo. La principal razón de ello es que –al margen de todas las exageraciones al respecto– el manejo de datos masivos sí representa una herramienta nueva y válida de análisis de información para ayuda de la toma de decisiones en las organizaciones, entre ellas las bibliotecas y los archivos, y por lo mismo es conveniente que el personal de esas organizaciones dedicado a la gestión de la información se introduzca al conocimiento de esas técnicas, herramientas y procedimientos. Este conocimiento y habilidades no pueden dejarse solo a cargo del personal de informática: es indispensable que los bibliotecarios y archivistas también lo manejen, ya que es un elemento de valor agregado tanto para la organización donde laboran como para el personal que se dedica a la gestión de información. Todo indica que los datos masivos pueden potenciar grandes innovaciones en la totalidad del ambiente de los servicios de información –obviamente esto incluye a las bibliotecas– para crear y proporcionar servicios inéditos y personalizados, a la vez que se optimizan costos.

Además, paralelamente se vienen desarrollando otros fenómenos relacionados a la información que tienen que ver muy de cerca

con los datos masivos, como el concepto de "ciudad inteligente" o "ciudad digital", el cual, básicamente consiste en

> […] el espacio virtual de interacción entre todos los actores que participan en la vida de una ciudad –ciudadanos, empresas, administraciones, visitantes, etcétera– utilizando como soporte los medios electrónicos y las TIC, ofreciendo a dichos actores acceso a un medio de relación y comunicación innovador, a través del canal que elijan, en cualquier momento y lugar. El objetivo principal es la mejora de la relación y los servicios entre los actores que interactúan en la ciudad, tanto en los servicios existentes como en los futuros, potenciando un desarrollo sostenible económico y social de la ciudad (Manual de Ciudades Digitales 2012).

Por supuesto, no puede concebirse un concepto de "ciudad digital" sin considerar las bibliotecas y servicios de información propios al efecto.

La afamada revista estadounidense *Forbes*, especialista en finanzas y economía, afirmó que en 2015 existían en el mundo menos de 20,000 especialistas en datos y su análisis, y que para 2020 se requerirían 2.7 millones de ellos (Columbus 2017). De acuerdo con este dato, existe una brecha enorme entre la oferta y la demanda de personal especializado en el manejo de datos.

Por tanto, conviene que bibliotecarios y archivistas comiencen a manejar los principios y teorías de los datos masivos, los modelos de datos estructurados, no estructurados o semi-estructurados, así como algunas de las herramientas y tecnologías de *hardware* y *software* existentes al efecto para su análisis. Se abundará en ello más adelante.

Antecedentes

Se gastan millones de dólares anuales en la recopilación de datos... Aquellos que pueden permitirse máquinas de procesamiento de tarjetas perforadas pueden procesar más datos que nunca: ¡3,000 registros por hora!

WILLARD BRINTON, 1914

Si bien el concepto actual de *datos masivos* data de principios de este siglo, la idea subyacente del análisis de datos para solucionar problemas complejos proviene de mucho más atrás. En un sentido estricto, cualquier esfuerzo de aplicar la ciencia más allá de la teoría para comprender y resolver problemas complejos es un antecedente de la ciencia de los datos y el análisis de los mismos. Algunos autores remontan estos esfuerzos hasta el siglo XVII con los trabajos acerca del "problema de los puntos" tratado por Christiaan Huygens y Blaise Pascal, los cuales plantearon las bases de solución de problemas con métodos estadísticos primigenios (Bell 1947). Se menciona también al respecto la investigación de Charles Babbage sobre el costo del transporte y clasificación del correo, la cual dio lugar al *Penny Post* o tarifa universal postal de Inglaterra en 1840 (Staff 1993). En el mundo de las empresas, Richard Devens menciona en 1865 por primera vez el término "inteligencia de negocios" (*business intelligence)* para describir cómo el banquero Sir Henry Furnese obtuvo beneficios al recibir y actuar con información acerca de su entorno, antes que sus competidores (Devens 1865, 210). Más allá de esos primeros intentos,

con el advenimiento de la Segunda Guerra Mundial los países en conflicto comenzaron a crear grupos multidisciplinarios de científicos para resolver los problemas que se presentaban a través del análisis integral de datos. Se crearon así las primeras teorías, principios, metodologías y grupos de lo que se empezó a denominar "Investigación de Operaciones" (Operations Research o IO). Una de las herramientas que el mismo conflicto propició fue la computadora electrónica, la cual impulsó el análisis cada vez mayor de datos en tiempos menores.

La "Investigación de Operaciones" se definió básicamente como una disciplina que aplicaba métodos analíticos avanzados para encontrar soluciones óptimas a problemas complejos que involucren toma de decisiones, empleando técnicas matemáticas como el modelado y optimización matemáticos, así como el análisis estadístico. Por esta razón, la IO se traslapa con otras disciplinas, metodologías y herramientas como el álgebra lineal, la simulación matemática, la ingeniería industrial, la teoría de colas, las cadenas de Markov, la teoría de "árboles", los modelos económicos, la gestión del conocimiento, las variables estocásticas, etcétera. Por su naturaleza de estar basada en el análisis de datos, se entrelazó fuertemente y desde un principio con las ciencias de la computación, como la informática, la ingeniería de *software,* los sistemas expertos, el diseño de algoritmos, la estructura de datos, etcétera.

Terminado el conflicto, el conocimiento creado por la IO se trasladó rápidamente hacia la industria, el comercio, la economía, la meteorología, las finanzas, etcétera. Sus principios y técnicas se usaron para resolver problemas complejos de la vida real que involucraban grandes cantidades de variables en diversos sectores productivos o de gobierno. El enorme e incesante desarrollo de las capacidades de los equipos y programas computacionales en la segunda mitad del pasado siglo propició que los problemas planteados y la cantidad de datos manejados crecieran exponencialmente. El nombre de "Investigación de Operaciones" fue desapareciendo en la segunda mitad del siglo XX conforme se desarrollaban nuevos enfoques, teorías, herramientas, etcétera, utilizadas para el análisis de datos dando paso a nuevas especialidades con otras

denominaciones; no obstante, la IO sentó las bases de esos nuevos modelos.

Durante los años sesenta, se crearon los principios y teorías básicos que hoy forman la ciencia de los datos, la gestión de información, la gestión de datos, etcétera. En 1962 John Tukey escribió:

> Durante mucho tiempo pensé que yo era un estadístico, interesado en las inferencias de lo particular a lo general. Pero a medida que he visto evolucionar las estadísticas matemáticas, he tenido motivos para preguntarme y dudar [...] He llegado a sentir que mi interés central está en el análisis de datos [...] El análisis de datos, y las partes de la estadística que se adhieren a él deben [...] asumir las características de la ciencia en lugar de las de las matemáticas [...] el análisis de datos es intrínsecamente una ciencia empírica [...] ¿Cuán vital y cuán importante [...] es el surgimiento de la computadora electrónica con programas almacenados previamente? (Tukey 1962, 2-4).[3]

Conforme la década avanzó, la respuesta se fue dando, al tiempo que se desarrollaban las computadoras y programas y se fueron sentando los cimientos para el manejo de bases de datos. A John Tukey se le atribuye haber expresado las primeras ideas que dieron origen a la ciencia de los datos.

Se creó así durante la década de los sesenta el concepto de "base de datos". Si bien el diccionario de inglés *Oxford* en línea cita un informe de 1962 de la "System Development Corporation" de California como la primera utilización del término "base de datos" (*database)* en su sentido técnico específico, varios testimonios de pioneros al respecto mencionan que a fines de los sesenta el desarrollo y concepto eran todavía muy incipientes. Entre otros, Mike Blasgen menciona que: "[…] alrededor de 1968 en San José, California, ya trabajaban en bases de datos. Pero no se llamaba así

---

3 A John Tukey se le atribuye también haber acuñado dos términos básicos de la computación: "bit", como contracción de "binary digit" o "dígito binario", y "software", como la contraparte intangible del "hardware" o "ferretería".

entonces; se llamaba 'administración de datos' [data management] o 'sistemas de archivo' [file systems]" (McJones 1995, 7). Del estudio de los más relevantes textos al respecto, se desprende que el despegue del concepto se dio en realidad en la primera mitad de los setenta, y su comercialización y uso se dio en la segunda mitad de esa década. El "Modelo Relacional de Bases de Datos" se perfeccionó en 1973 en el Laboratorio de Investigación de IBM en San José, California, el cual era denominado entonces "System R", mismo que daría origen al famoso SQL (*Structured Query Language* o "Lenguaje de Consulta Estructurado"). Oracle, todavía con el nombre de "Relational Software, Inc." comenzó a comercializar el concepto en 1979.

Edgar Codd publicó en 1970 un texto que sentó los fundamentos para el desarrollo del concepto de bases de datos y se convirtió en todo un hito en el manejo de los mismos. Hasta ese entonces, los datos eran manejados en las computadoras con estructuras al libre albedrío de los programadores. El autor estableció ahí por primera vez que

> […] Hay que evitar que los futuros usuarios de grandes bancos de datos tengan que saber cómo están organizados los datos en la computadora; esto es, la representación interna. Un servicio de preguntas y respuestas que proporcione dicha información no es una solución satisfactoria. Las actividades de los usuarios y la mayoría de los programas de aplicación no debiesen verse afectados cuando se modifica la representación interna de los datos" (1970, 377).

Nótese como el autor usa el término "bancos de datos" y no "bases de datos", concepto que precisamente explicaría ahí. Previo a este punto, los datos y sus programas de explotación estaban intrínsecamente entrelazados. La importancia capital de todo lo anterior consiste en que a partir del desarrollo de las bases de datos, al agregar a ellos ciertas estructuras estarían ya separados de sus programas, con lo que se inicia una nueva etapa en su manejo y explotación con ayuda de computadoras, lo cual da un enorme impulso a su gestión.

El desarrollo de las bases de datos puede dividirse en tres grandes épocas basadas en el modelo o estructura de datos: bases de datos de navegación o jerárquicas, SQL o relacionales y post-relacionales. Surgieron así variadas teorías de las bases de datos jerárquicas y los primeros sistemas informáticos para su manejo, conocidos como *Database Management Systems* (DBMS), por parte de diversos fabricantes. Estos evolucionarían hacia las bases de datos relacionales, las orientadas a objetos, distribuidas, etcétera, con innumerables productos especializados a la fecha.

En 1974, Peter Naur publicó su obra *Concise Survey of Computer Methods,* la cual es un tratado de los métodos computacionales de procesamiento de datos de ese entonces. Él definió ahí: "[…] los datos son una representación formalizada de hechos o ideas capaz de ser comunicada o manipulada por cierto tipo de proceso". Algo muy relevante de esta obra es que el autor menciona en su prefacio que en un congreso en 1968 se presentó un plan de acción titulado "Datalogía, la ciencia de los datos y de los procesos de datos y su lugar en la educación", y que en el texto de esa obra el término "ciencia de los datos" fue utilizado libremente. El autor ofrece además una definición de "ciencia de los datos": "[…] 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 1974). Ésta es la referencia más antigua encontrada al respecto de ciencia de los datos. En la actualidad, abarca todavía más campos: desde la colecta y selección de datos, su gestión y preservación, las fuentes de datos masivos, la aplicación de técnicas para la minería de datos, los "almacenes de datos", la detección de tendencias en redes sociales, la interacción entre hombre y computadora, el análisis y visualización de datos y la evaluación de la calidad de datos e información, hasta las políticas de información.

La "gestión de datos" (*data management*) comenzó también desde las décadas de los sesenta y setenta. En un principio, el concepto abarcaba exclusivamente teorías, herramientas y métodos para el manejo de datos con sistemas informáticos. Donald Knuth, uno de los más grandes desarrolladores de este tema, afirmaba

entonces que la clave para la adecuada gestión de datos –en ésa, su acepción original– consistía fundamentalmente en algoritmos susceptibles de programarse y técnicas matemáticas formales sistematizadas. Él escribió toda una serie de libros acerca de esos dos elementos, considerados hoy obras fundamentales de las herramientas para el tratamiento profundo de datos con ayuda de computadoras.[4] El concepto fue evolucionando gradualmente –como los demás conceptos asociados– hacia otros más amplios y multidisciplinarios, hasta tomar su forma actual. Hoy en día, la *gestión de datos* es una teoría y varias prácticas administrativas que consisten en colectar, validar, organizar, almacenar y utilizar datos de forma segura, eficiente y rentable para convertirlos en un recurso valioso dentro de una organización. El objetivo de la gestión de datos es ayudar a las personas y organizaciones a la toma de decisiones que maximicen el beneficio para ellas. Para ser eficaz, la gestión de datos requiere de una "estrategia de datos", así como 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 concierto los convierte en poco tiempo en algo inútil, además de difícil de manejar; su verdadero valor no depende de su simple existencia dentro de una organización, sino de lo que se puede hacer con ellos. Como puede verse, la gestión de datos pasó de ser un término exclusivamente informático a toda una estructura conceptual compleja y multidisciplinaria.

En años recientes, se agregó a la lista de conceptos la "gobernanza de datos" (*data governance)*. El Instituto de Gobernanza de Datos la define como "[...] un sistema de derechos de decisión y rendición de cuentas para los procesos relacionados con la información, ejecutados de acuerdo con modelos acordados que describen quién puede tomar cuáles medidas con qué información,

4 "The Art of Computing Programming" es una serie de doce libros publicados por Knuth sobre el tema a partir de 1970.

cuándo, en qué circunstancias, y con cuáles métodos" (Defining Data Governance s.f.). Se abundará en ello más adelante.

En 1989, antes de la red mundial, Roy Davies apuntó ya algunos de los primeros aspectos de la extracción de conocimiento a partir de los catálogos de biblioteca (Davies 1989). En ese mismo año, Edward Feigenbaum, pionero de la Inteligencia Artificial, considera a las bibliotecas de ese entonces "[…] almacenes de objetos pasivos, donde los libros y revistas se asientan en estantes en espera de que alguna persona use su inteligencia para encontrarlos, interpretarlos, y hacer que eventualmente divulguen el conocimiento que tienen guardado", y visualiza una "biblioteca del futuro" donde los libros interactuarían y colaborarían con el usuario a través de un sistema informático inteligente que sería capaz de interactuar con varios usuarios simultáneamente (Feigenbaum 1989, 122). A fines de la década de los ochenta, comienzan a surgir los conceptos de "extracción del conocimiento", "descubrimiento del conocimiento" y más específicamente "descubrimiento del conocimiento en bases de datos", los cuales marcarían una nueva etapa en el procesamiento de información, pero que incluirían ya fuertes componentes del procesamiento de datos en equipos y programas de cómputo, técnicas y metodologías, algoritmos, etcétera. Estos conceptos y sus técnicas derivadas contribuyeron fuertemente a consolidar los datos para llegar a obtener su valor y peso actuales.

En 1996, Usama Fayyad, Gregory Piatetsky-Shapiro y Padhraic Smyth publicaron una obra que sería un parteaguas al respecto: *De la minería de datos al descubrimiento de conocimiento en las bases de datos* (*From Data Mining to Knowledge Discovery in Databases)*. Allí establecieron*:*

> […] 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]*... 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. 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 (Fayyad, Piatetsky-Shapiro y Smyth 1996, 39).

Al año siguiente, en 1997, ellos editarían la revista *Data Mining and Knowledge Discovery*. El interés acerca del tema se incrementaría rápidamente: en los números del verano de 1999 de la revista *Library Trends* y de noviembre de ese año de *Communications of the ACM* el tema central fue precisamente el descubrimiento de conocimiento en bases de datos bibliográficas (Qin y Norton 1999).

Con respecto al término "minería de datos" o *data mining*, se considera una de las etapas del mencionado Descubrimiento de Conocimiento en Bases de Datos. Proviene de los términos usados en los sesenta por los estadísticos: "pesca de datos" (*data fishing)* y "dragado de datos" (*data dredging)*. En los años ochenta, comenzó a usarse el término "minería de bases de datos" (*database mining)*, pero al ser patentado el término como un producto comercial, evolucionó a "minería de datos", y tomó su acepción actual desde los noventa. Como puede verse en la mencionada publicación de Fayyad y su posterior revista, el concepto ya era bastante común a mediados de esa década. Han y colegas la definen como:

> […] La minería de datos es un subcampo interdisciplinario de la informática y 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 (Han *et al.* 2011, 15-21).

Hace poco más de una década, el término fue introducido en la bibliotecología como "bibliominería" (*bibliomining)*, concepto que será analizado más adelante.

En ese mismo año, Hayashi (1996, 41) presentó una ponencia denominada "Ciencia de los datos, clasificación y métodos relacionados". El definió ahí:

> […] La ciencia de los datos no es solo un concepto sintético para unificar estadísticas, análisis de datos y sus métodos relacionados; también abarca sus resultados. La ciencia de los datos tiene por objeto analizar y comprender fenómenos reales con 'datos'. Dicho de otra forma, el objetivo de la ciencia de los datos es revelar las características o la estructura oculta de los complicados fenómenos naturales, humanos y sociales utilizando datos, desde una perspectiva diferente de las teorías y métodos establecidos tradicionales.

Ésta es una de las tempranas definiciones de la ciencia de los datos. A partir del 2002, se creó una revista especializada en la ciencia de los datos precisamente con ese nombre: *Data Science Journal.*

En un texto del año 2000, Noboru Ohsumi menciona que desde 1992 él había señalado: "[...] es importante destacar que estamos de acuerdo en la necesidad de desarrollar, a través de la práctica, la investigación sobre la teoría y la aplicación del análisis de datos en una nueva 'ciencia de los datos'" (Ohsumi 2000).

Para profundizar en el desarrollo temprano de la ciencia de los datos, se recomiendan dos excelentes revisiones panorámicas: Murtaugh y Devlin (2018) y Cao (2017). Este último (Cao 2017, 15)

menciona que la primera conferencia que tuvo ya como tema la ciencia de los datos fue la Conferencia del IFCS sobre Ciencia de los Datos, Clasificación y Métodos Relacionados de 1996 (Hayashi *et al.* 1996). Él además presentó una definición completa y moderna de "ciencia de los datos" resumiendo diferentes visiones del concepto:

> [...] Desde la perspectiva disciplinaria, la ciencia de los datos es un nuevo campo interdisciplinario que sintetiza y se basa en la estadística, la información automática, la computación, la comunicación, la gestión y la sociología para estudiar a los datos y sus entornos –incluidos los dominios y otros aspectos contextuales, como los aspectos organizativos y sociales– con el fin de transformar los datos en percepciones y decisiones, siguiendo un pensamiento y una metodología datos-conocimiento-sabiduría (Cao 2017, 8).

Otro de los factores que sin duda contribuyeron a todo lo anterior fue el inmenso e incesante crecimiento de las capacidades de almacenamiento de datos a lo largo del siglo XX y lo que va del corriente. No puede pensarse en colectar grandes cantidades de datos si no se tiene dónde almacenarlos. Por lo mismo, el concepto de "grande" o "masivo" con respecto a los datos ha estado estrechamente ligado a las capacidades de los dispositivos para el almacenamiento de cada época y ha ido cambiando con cada avance. El punto de partida consistió en la invención por parte de Hollerith en 1890 de la tarjeta perforada para el censo de Estados Unidos. El correspondiente a la década anterior había tardado casi diez años en contarse y ante el crecimiento de la población –pasó de 50 a 63 millones–, se veía como imposible llevarlo a cabo en solo diez años de plazo. El proceso de los 63 millones de tarjetas perforadas se hizo con tabuladores mecánicos en solo tres años. Desde entonces y por décadas, las tarjetas de cartón se volvieron el medio idóneo para el almacenamiento de datos procesables con ayuda de máquinas tabuladoras, clasificadoras, de contabilidad, de registro unitario y, finalmente, computadoras. Útiles como fueron, no obstante las tarjetas también marcaban un límite físico: 1 Gigabyte de datos en tarjetas de cartón perforadas requería 12.5

millones de tarjetas totalmente llenas, pesaba 31.25 toneladas y ocupaba un volumen de poco más de 31 metros cúbicos.

El siguiente paso fue la invención de las cintas magnéticas a fines de los veinte,[5] las cuales también fueron muy utilizadas por varias décadas y llegaron a almacenar en sus mejores versiones unos 30 Megabytes por unidad. Fueron sucedidas por los tambores magnéticos, con capacidades de 50 a 100 Kilobytes; les siguieron los discos magnéticos, los cuales comenzaron con unos 4 Megabytes de capacidad en los años cincuenta hasta varios Terabytes en la actualidad, luego los cartuchos de cinta *Linear-Tape Open* o LTO, con capacidades actuales de hasta 24 Terabytes cada una, y posteriormente por las memorias de estado sólido. En la tecnología computacional existe un conocido principio denominado "Ley de Moore",[6] el cual básicamente establece que la capacidad del procesador central de las computadoras se duplica cada 18 meses. Mark Kryder –exdirector de tecnología del gran fabricante de discos duros Seagate– afirmó de forma semejante que la cantidad de almacenamiento de datos que pueden alojarse en una cierta área de un medio magnético también se duplica cada 18 meses. Aunque esto ya no sucede en la actualidad, los aumentos en su capacidad siguen siendo sorprendentes.

Aunado a esto, los costos bajaron a proporciones inusitadas: 1 Gigabyte de almacenamiento magnético en disco costaba más de cien mil dólares en 1980; esa misma capacidad cuesta poco menos de tres centavos de dólar en 2020. En cinta, puede bajar hasta 1 centavo de dólar por Gigabyte. A este respecto, Morris y Truskowsky (2003, 206) afirman que desde 1996 se llegó al punto de inflexión donde el almacenamiento electrónico tenía ya mejor costo/beneficio que en papel.

---

5 Fritz Pfleumer, ingeniero austro-germano, inventó en 1928 la forma de almacenar información magnética en una cinta. Sus principios siguen vigentes hoy en día.

6 Principio empírico establecido por Gordon E. Moore, cofundador de la empresa Intel, en abril de 1965.

Concepto

Busca siempre la simplicidad en la Ciencia de los Datos. La verdadera creatividad no hace las cosas más complejas; siempre las simplifica.
DAMIAN DUFFY MINGLE

El término "dato" tiene su origen etimológico en el sustantivo latino *datum*, que se traduce como algo dado o establecido. De forma simple, un dato es una representación simbólica de los atributos de una entidad, hecho o suceso que toma la forma de una variable cuantitativa o cualitativa; es la expresión mínima de contenido acerca de un tema. Cuando los datos se consideran y analizan en conjunto y en contexto constituyen una información; por esta razón, los datos se colectan y se agrupan. Cuando los datos eran procesados de forma manual, obviamente la cantidad que podía ser analizada era muy limitada. Con el advenimiento de las máquinas electromecánicas a partir de Hollerith, el procesamiento de datos se fue incrementando sustancialmente, y con la llegada de las computadoras, la cantidad de ellos fue creciendo todavía más en función de la capacidad de estos equipos hasta llegar al manejo de cantidades inmensas de los últimos años: los "datos masivos".

Con respecto al concepto *big data* y como antecedente del mismo, Lohr (2013) menciona que "[…] el término big data, que abarca la informática, la estadística y la econometría, fue comenzado a utilizar en conferencias impartidas por John Mashey de la empresa Silicon Graphics a mediados de los noventa" (Lohr 2013, s.p.).

Kenwright (1999) presentó en la Conferencia "Visualization '99", una ponencia que ya incluía ese término. Diebold (2013) afirma que los datos masivos (*big data*) deben entenderse a la vez como un fenómeno, un término y una disciplina. Él menciona que encontró en la literatura algunas pocas referencias anteriores al año 2000 al respecto, tanto académicas como no académicas, donde se utiliza el término pero no se conoce a fondo el fenómeno, y agrega que, por el contrario, en esos años algunos académicos eran conscientes del fenómeno emergente pero no usaban el término. Obviamente la disciplina se crearía después.

Actualmente se acepta como punto de partida del concepto una nota publicada en febrero del 2001 por Doug Laney, analista especializado en información del Grupo Meta, perteneciente al Grupo Gartner, quien planteó las características fundamentales de este concepto usadas ampliamente a lo largo del tiempo (Laney 2001). La definición de datos masivos del Grupo Gartner ha sido desde entonces y durante muchos años de gran aceptación; en ella se incluyen ya las tres características fundamentales de ese tipo de datos establecidas por Laney, conocidas como las tres 'V': "[…] los 'datos masivos' son activos de información de gran volumen, velocidad y variedad que exigen formas rentables e innovadoras de procesamiento de la información para mejorar el conocimiento y la toma de decisiones" (Gartner Glossary 2005).

Volumen se refiere a la inmensa producción y acumulación de datos a nivel mundial en cantidades inusitadas y siempre crecientes. *Velocidad* se refiere a la enorme tasa a la que se crean los datos diariamente: millones de páginas web, mensajes, redes sociales, noticias, correos, solo por mencionar algunos. La *Variedad* tiene que ver con todos los tipos imaginables de datos que se producen cada día a partir de cuantiosas fuentes y formatos: las innumerables formas de representar los datos crean un serio problema para interpretarlos. Se abundará en estos tres conceptos más adelante.

Continuando con las definiciones, la empresa Oracle, especialista en productos para manejo de datos define:

> […] Los datos masivos son conjuntos de datos cada vez más grandes y complejos, especialmente provenientes de nuevas fuentes de datos. Estos conjuntos de datos son tan voluminosos que el software tradicional de procesamiento de datos no puede manejarlos. Pero estos volúmenes masivos de datos pueden ser utilizados para resolver problemas de negocios que antes no hubiesen podido ser abordados" (Oracle, s.f., s.p.).

Dans (2011) lo precisa así*:* "[…] Los datos masivos se refieren al tratamiento y análisis de enormes repositorios de datos, tan desproporcionadamente grandes que resulta imposible tratarlos con las herramientas de bases de datos y analíticas convencionales".

Gantz y Reinsel (2012, 3), los expertos de la Corporación IDC definieron: "[…] los datos masivos son una nueva generación de tecnologías y arquitecturas diseñadas para extraer económicamente valor a partir de volúmenes muy grandes de una amplia variedad de datos, permitiendo su captura, descubrimiento y/o análisis a alta velocidad".

Martínez Musiño (2020, 96) elaboró una definición de datos masivos desde la perspectiva de las ciencias de la Información: "[…] es el fenómeno de generación masiva y constante de datos, cuyo tratamiento y organización requieren tanto de recursos tecnológicos como de programas especializados de computadora y herramientas de interpretación y análisis para lograr la cientificidad".

La ya mencionada revista estadounidense *Forbes*, especializada en el mundo de los negocios y las finanzas y editora de listas que despiertan gran interés en los sectores técnico y económico, publicó en septiembre de 2014 un artículo en el cual menciona doce diferentes definiciones del término (*Forbes* 2014).

De Mauro, Greco y Grimaldi (2016, 131) establecieron una definición sintetizada basada en 15 definiciones formales analizadas, las cuales clasificaron en cuatro grupos de características básicas: 1) información, 2) tecnología, 3) métodos, 4) impacto. A partir de estas características, elaboraron la siguiente definición: "[…] Los datos masivos son activos de información caracterizados por un volumen, una velocidad y una variedad tan grande que requieren

de tecnologías y métodos analíticos específicos para su transformación en valor". Aunque es una síntesis de muchas definiciones, puede verse que no varía mucho con respecto a las anteriores, y no aporta algo diferente.

Google los define 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 otros métodos avanzados de análisis de datos que extraen valor de los datos, y rara vez a un tamaño en particular de conjuntos de datos.

A manera de contraste, cabe mencionar que algunos autores han tratado de visualizar estas definiciones de los datos partiendo del enfoque opuesto; es decir, definir los "datos no masivos" (*small data*) para de ellos moverse hacia las definiciones de datos masivos. Como ejemplo, Pollock (2013 s.p.) definió: "[…] los datos no masivos son aquellos cuya cantidad se puede almacenar y procesar convenientemente en una sola computadora; específicamente, en un servidor de alto rendimiento". Este tipo de definiciones tiene como característica común la capacidad de la organización para manejar localmente sus datos dentro de sus propios equipos, sin recurrir a proveedores o servicios externos.

En conclusión, al igual que con muchos otros conceptos tecnológicos complejos de esta época, no existe una definición universal o consensuada del término "datos masivos" (*big data)*, como establecieron detalladamente Zhan y Widén (2017, 569). No obstante, al margen de las numerosas definiciones que existen del concepto, puede hacerse una caracterización con los elementos que se encuentran como común denominador entre ellas: 1) los datos masivos consisten en el tratamiento y análisis de conjuntos de datos tan grandes, variados, complejos y dispares, 2) producidos a una velocidad tan rápida y provenientes de muy diversas fuentes, 3) que los equipos, programas y procedimientos "tradicionales" de

procesamiento de información: servidores, bases de datos, buscadores, etcétera, no son suficientes y 4) por tanto requieren métodos, equipos y programas mucho más poderosos, sofisticados y especializados para compilarlos, analizarlos y correlacionarlos, 5) todo con el fin de poder extraer rápidamente de esos datos patrones, tendencias y asociaciones, principalmente del comportamiento e interacciones humanas y 6) de ahí estar en posibilidad de tomar decisiones fundamentadas que ayuden a las organizaciones, 7) lo cual otorga a los datos un enorme valor agregado.

A partir de esta caracterización, es posible ir analizando con mayor profundidad sus siete componentes para lograr un mayor entendimiento del concepto de los datos masivos. Más aún, como afirmara Diebold, se han convertido en toda una disciplina, la cual conviene explicar.

El núcleo fundamental en la caracterización es la colecta, organización, almacenamiento, y análisis de datos, cuyas características esenciales fueron enunciadas en las tres "V": volumen, velocidad y variedad. Conviene analizar con más detalle a qué se refieren esos tres elementos.

Volumen: el mundo ha experimentado en las últimas décadas un aumento inusitado de producción de información a nivel global, la cual incluye también a los datos. El primer conteo masivo al respecto es el conocido estudio realizado por Lyman y Varian (2000), quienes establecieron que durante ese año el mundo generó un Exabyte (EB) de datos. Si bien la cantidad exacta de datos por año es algo muy difícil de estimar y se encuentran variadas cuentas al respecto, existe una serie de estudios a los cuales se han hecho múltiples referencias, realizados por la organización IDC desde 2007 hasta 2020. Ellos dan cuenta del vertiginoso crecimiento del volumen de información digital. Estimaron que el mundo produjo 130 Exabytes (EB) de información en 2005; 1,227 en 2010; 2,837 en 2012; 8,591 en 2015, y 40,026 Exabytes en 2020 (Gantz y Reinsel 2007, 2012).[7] Esto significa que la cantidad de

7 1 Exabyte = 1,000 Petabytes = 1'000,000 Terabytes = 1,000'000,000 Gigabytes = 1'000,000'000,000 Megabytes = 1 Trillón europeo = 10^{18} bytes o caracteres.

información que se produce en el mundo se ha duplicado aproximadamente cada dos años durante las últimas dos décadas. Como referencia, considérese que un Exabyte de información equivale a 500 billones[8] de cuartillas de texto de 2,000 caracteres cada una, o dos billones de libros de 250 páginas cada uno, o mil billones de mensajes de correo electrónico de mil caracteres cada uno, o un billón de páginas web de un Megabyte cada una, o 333 mil millones de fotografías de alta resolución, o 250 mil millones de archivos de música mp3 de cuatro minutos cada uno, o casi dos mil millones de CD-ROM, o 3,500 billones de tuits de 280 caracteres cada uno. Es decir, el mundo está inundado de información y datos que crecen exponencialmente.

Velocidad se refiere a la enorme tasa a la que se crean los datos diariamente: millones de páginas web, mensajes, redes sociales, noticias, correos, videos e imágenes, solo por mencionar algunos. Obviamente, para producir los volúmenes de datos mencionados se requiere forzosamente que estos se produzcan a una velocidad vertiginosa. En los últimos dos años, se han producido en promedio 27 Exabytes por día. Esto es, 27 seguido de 18 ceros, o 27 trillones de bytes cada día. El sitio We are Social menciona que en 2020, dentro de los 4,500 millones de usuarios de Internet que existían en el mundo hay 3,800 millones de usuarios de redes sociales (We are Social 2020). Dentro de ese conjunto y a manera de ejemplo, Twitter por sí solo genera más de 7 Terabytes de datos (TB) diariamente, y Facebook 10 TB de datos cada día.[9] Se enviaron 200 mil millones de correos electrónicos cada día del último año y se suben diariamente a la red 100 millones de imágenes a Instagram. Gran parte del origen de esta velocidad proviene de los usuarios en sí mismos, pero además de esto, se crean ya muchos más datos acerca de ellos por parte de las redes en las que se mueven, y todavía más es creada por máquinas y objetos conectados a la Internet. Los dispositivos y objetos físicos se han aunado

8 Billón europeo; es decir, un billón es igual a un millón de millones.

9 1 Terabyte = 1,000 Gigabytes = 1'000,000 Megabytes = 1'000,000'000,000 bytes = 10^{12} bytes o caracteres.

a las personas y organizaciones para participar activamente en la producción del enorme flujo de datos que se incorpora diariamente al universo digital.

Finalmente, la *Variedad* está relacionada con todas las formas posibles para representar esos datos, pues el mundo produce todo tipo de datos imaginables en innumerables formatos: noticieros, música, archivos, fotografías, mapas, catálogos, redes sociales, chats, mensajería, tiendas, juegos electrónicos, llamadas, TV, radio, cine, eventos, tutoriales, sitios y portales, videos, blogs, correos; datos de salud, bancarios, comerciales, tributarios, académicos, periodísticos, etcétera. Estos datos son producidos tanto por personas y organizaciones como por máquinas.

El problema asociado con la variedad consiste en que existen múltiples formas de entender y clasificar los datos: de acuerdo con su forma, pueden ser cuantitativos o cualitativos; según su fuente pueden ser capturados, derivados, exhaustivos o transitorios; de acuerdo con su tipo, pueden ser primarios, secundarios o metadatos; según su representación o estructura, se consideran tres tipos de ellos: datos estructurados, no estructurados y semi-estructurados. Los primeros provienen de organizaciones y sistemas diseñados al efecto: datos oceanográficos y meteorológicos, sismológicos, astronómicos, bancarios, tributarios, etcétera, y por esa estructura bien especificada son aquellos con mayor facilidad de proceso, ya que son homogéneos, están normalizados y ordenados y con esos elementos se facilita su administración. Ejemplos de ellos son una fecha en formato dd/mm/aa, un número telefónico a diez dígitos o un ISBN (Joyanes 2013 y Olavsrud 2012). Los no estructurados provienen de contenidos web, redes sociales, foros, correos electrónicos, archivos de texto simples, hojas de cálculo, audios o videos, blogs, mensajes de correo de voz, mensajes instantáneos, etcétera. No tienen tipos definidos ni están organizados bajo algún patrón; tampoco son almacenados de manera relacional, o con base jerárquica de datos; no tienen un formato normalizado determinado, y no es fácil identificar su tipo o clase. Para poder procesarlos, es indispensable organizarlos y clasificarlos, y la única manera de hacerlo es cuando contienen metadatos de origen o estos pueden ser agregados con

cierta facilidad. Entre esos dos extremos, se encuentran los datos con estructura intermedia o semi-estructurados. Se abundará en todos ellos más adelante.

Muchos autores coinciden en que los datos masivos no pueden ser descritos integralmente con solo esas tres características, por lo que han ido agregando a las originales otras nuevas que ellos consideran relevantes. Marr (2015), Lomotey y Deters (2014), y Affelt (2015) establecen que son cinco las "V" o características; ellos agregan a las originales la *Veracidad*, la cual se refiere a la fiabilidad o desviación de los datos; otros autores que también la incluyen son Lawlor (2016) y Plale (2013). Esta característica se refiere a que al haber grandes cantidades y variedad de los datos, la calidad y precisión son menos controlables, y puede existir gran sesgo e inconsistencias en ellos; por ejemplo, los mensajes de Twitter con etiquetas *bash*, las "tendencias", las "noticias falsas" o *fake news*, etcétera. De acuerdo con Marr, los métodos de colecta y análisis de datos deben por tanto incluir mecanismos para su depuración. Affelt (2015, 21) llama a esta característica *Verificación,* pero se refiere a lo mismo: "el proceso por el cual bibliotecarios y profesionales de la información analizan las fuentes de datos y sistemas de recuperación para determinar la calidad de los mismos". Marr, Lomotey y Deters, y Affelt agregan también como quinta característica de los datos masivos al *Valor:* se refiere a la capacidad real de poder extraer información valiosa y útil del conjunto de los datos, pues de nada sirve colectarlos y procesarlos si no se puede obtener un beneficio tangible para la organización. Marr refiere que ésta es la característica más importante de las cinco, pues muchas organizaciones han caído en el atractivo de los datos masivos sin tener conocimientos y capacidades suficientes, haciendo fuertes inversiones en ellos con poco valor redituable para ellas. Las proporciones costo/beneficio son capitales en este aspecto. Muchos otros autores coinciden con estas dos características añadidas y aunque algunos agregan todavía más, como por ejemplo la *Variabilidad, Volatilidad, Vaguedad* o la *Complejidad*, hay un consenso acerca de que las cinco características descritas son las mínimas indispensables para tratar de forma adecuada a los datos masivos.

Al margen de las definiciones y características, es conveniente subrayar en este punto que la tendencia actual de la conceptualización de los datos masivos hace énfasis en que su importancia no debe estar en realidad en las técnicas y herramientas para el manejo de muchos datos, sino en cómo extraer valor de ellos, cómo hacerlos útiles a las organizaciones. Es indispensable pensar en cómo capitalizar esos datos para que su potencial pueda ser realmente aprovechado para la toma de futuras decisiones organizacionales críticas. Sólo al ser capaces de organizar y explotar eficazmente esta información, se podrá obtener una mayor inteligencia en la organización al permitir una mejor y más rápida toma de decisiones. La preocupación por un mejor y mayor aprovechamiento de los datos ha ido creando una serie de tendencias, especialidades, etcétera, alrededor de ello.

De hecho, en los últimos años ha surgido al respecto una corriente dentro de los datos masivos denominada *thick data* (datos densos o datos espesos), la cual establece que para ser explotados integralmente, los datos masivos requieren ser complementados con un enfoque cualitativo de ellos además del cuantitativo, que permita conocer contextos, sentimientos, historias, opiniones, emociones y los modelos del entorno de los sujetos estudiados. Usualmente, los datos masivos son procesados por matemáticos, estadísticos e informáticos y se basan en procesos matemáticos y computacionales; los "datos densos" son procesados por antropólogos, etnólogos, sociólogos y científicos sociales, y hacen énfasis en el uso de otras herramientas de investigación propias de esas disciplinas para el análisis de datos: encuestas, cuestionarios, entrevistas, grupos de discusión, diarios personales, etcétera. El concepto de los *datos densos* fue acuñado por Tricia Wang basada en los postulados de los antropólogos sociales Clifford Geertz (1973) y Gilbert Ryle (1949), quienes establecieron el concepto de "descripción densa" (*thick description)*: un conjunto de técnicas de investigación cualitativa y etnográfica para las ciencias sociales que tiene como propósito elaborar descripciones, contextos e interpretaciones detalladas de situaciones estudiadas por un investigador. En estrecha relación con esto, en años recientes se ha desarrollado la disciplina conocida como

"netnografía" –de *net*, la red, y *etnography*, etnografía–, la cual adapta técnicas etnográficas de investigación para estudiar procesos, relaciones y prácticas culturales, comunidades virtuales, fenómenos y dinámicas específicas que se encuentran en la red mundial a través del análisis de los datos disponibles en ella. Paul Gibbons (2015) resume todos estos conceptos espléndidamente en una sola frase: "El lado humano del análisis es el mayor reto para la implementación de los datos masivos".

Por otra parte, esta preocupación por un mejor y mayor aprovechamiento de los datos ha creado toda una especialidad profesional alrededor de ello dentro de la *ciencia de los datos*, denominada específicamente "análisis de datos" (*data analytics)*. Actualmente se han desarrollado estudios de posgrado, cursos, diplomados, etcétera, alrededor de este tema por parte de innumerables instituciones. Básicamente, esta especialidad tiene como fin ulterior lo mismo que la ciencia de los datos en general; es decir, detectar tendencias y patrones en los datos para plantear soluciones. Pero más puntualmente, esta especialidad se refiere a la tarea de identificar cuáles variables de la organización pueden ser relacionadas con ciertos datos y establecer así correlaciones para el planteamiento de preguntas y la eventual obtención de soluciones a través de técnicas específicas. Estas técnicas forman el núcleo del análisis de datos como especialidad. Se abundará en ello más adelante.

La importancia de los *datos masivos*

Una respuesta aproximada al problema correcto vale mucho más que una respuesta exacta a un problema aproximado.

JOHN TUKEY

Ya ha sido mencionada la importancia del análisis sistematizado de datos para optimizar la toma de decisiones en las organizaciones, en especial con datos masivos, y es indudable el interés que este tema ha despertado en las últimas dos décadas. Kalantari y colegas (2017) realizaron un estudio bibliométrico en el que mencionan que encontraron 6,572 artículos sobre el tema de datos masivos solo en el Web of Science. Xu y Wu (2019) encontraron en un estudio semejante 10,989 documentos sobre el tema; ambos estudios son un claro indicador del interés y desarrollo al respecto. Varias revistas dedicaron números completos al tema, como por ejemplo *Research Trends* tan temprano como 2012 (*Research Trends* 2012).

Desde 2015, la editorial Springer comenzó a editar una revista exclusivamente sobre este tópico: *Journal of Big Data*; desde 2011 apareció ya el primer *Glosario de datos masivos* (Warden 2011), y en 2013 el primer *Manual de datos masivos* (Plunkett *et al.* 2013). Halevi y Moed (2012) hicieron una revisión de la literatura existente en ese año acerca de los datos masivos como tema científico y de investigación, y puede verse de sus resultados que el tema ya había adquirido notoriedad desde entonces. Conviene ahora

ahondar en los principales usos de este tipo de datos en los diversos sectores en general para poder apreciar todas sus potencialidades, y de ahí partir para vislumbrar sus usos y beneficios en las bibliotecas.

El proceso, uso y análisis de datos masivos es utilizado hoy en día ampliamente en innumerables sectores de la sociedad, por lo que no se pretende aquí presentar una lista exhaustiva de todas sus aplicaciones prácticas, pero sí hacer una selección que sea lo suficientemente representativa e ilustre sus usos y beneficios.

1.- Sector financiero, banca, seguros
Algunos ejemplos de uso de los datos masivos en este sector son el análisis de riesgos financieros, modelos económicos, análisis actuarial para seguros, detección de movimientos fraudulentos con tarjetas de crédito, desarrollo de nuevos servicios personalizados para la banca y seguros, entre otros. Las plataformas bancarias digitales de hoy y todas las operaciones al respecto que pueden hacerse desde casa son resultado también del análisis de datos.

2.- Sector industrial
Algunos ejemplos de uso de los grandes en este sector son: la optimización de procesos de manufactura, el análisis de riesgos, diversificación o integración de negocios en corporaciones, pronósticos de demanda y consumo, datos puntuales provenientes de GPS o identificadores de radiofrecuencia RFID, programación y seguimiento instantáneo de entregas y envíos, manufactura controlada por robótica, impresoras 3-D, fabricación y retroalimentación de datos de productos de "Internet de las Cosas" y dispositivos personales o *Wearables*, Inteligencia Artificial, optimización de eficiencia energética, entre otras.

3.- Comunicaciones y transportes
Algunos ejemplos de uso de los grandes en este sector son: la predicción de demanda a corto y largo plazo de servicios de Internet y telefonía –servidores, centros de datos, canales, redes, ancho de banda, la nube, etcétera–, así como de servicios de transporte:

aviones, trenes, autobuses, carreteras, barcos, etcétera; cálculo de embarques, rutas, aeropuertos, puertos marítimos, y otros, para optimizar tiempos, envío, combustibles, estiba, etcétera; reprogramación instantánea de recursos para su optimización en función del uso y demanda: servidores, redes, canales, centrales y otros. Por ejemplo, la empresa de transporte Uber utiliza el análisis de datos masivos en tiempo real para cambiar dinámicamente las tarifas de sus viajes en función de la demanda a cada momento.

Se utilizan también para crear nuevos servicios para diferentes públicos objetivo y hacer recomendaciones de esos servicios a nuevos usuarios potenciales, medir el rendimiento de los recursos, etcétera. De hecho, la nueva tecnología de comunicaciones como el 5G[10] incluye mucho de análisis de datos masivos para asignación dinámica de recursos.

4.- Meteorología

Los servicios de meteorología, oceanografía, vulcanología, etcétera, compilan incesantemente grandes cantidades de datos para alimentar sus modelos de comportamiento del clima a largo plazo en cada parte del mundo, así como para eventos puntuales como huracanes, tornados, erupciones, avalanchas, etcétera. Estos modelos requieren de enormes cantidades de datos en tiempo real provenientes de innumerables sensores para poder ser actualizados con frecuencia y calcular trayectorias, hacer advertencias oportunas, etcétera.

5.- Gobierno

El sector gobierno es otros de los sectores que más ha aprovechado el análisis de datos. Se utiliza para muy diversas tareas: diseño, implementación de servicios para el ciudadano, seguimiento y análisis de actividades tributarias, diseño y seguimiento de servicios de salud, educación, transporte, etcétera. Grandes iniciativas,

---

10 5G es la quinta etapa de tecnologías de telecomunicaciones para celulares e Internet. Implica principalmente mayores anchos de banda (hasta 10 Gigabits por segundo) y mayor disponibilidad para muchos usuarios a la vez.

como la de "Gobierno abierto", se basan fuertemente en el procesamiento masivo de datos. Más puntualmente, el análisis de datos en este sector permite cambiar parámetros al instante para el control de tráfico, semáforos, accesos, vialidades, etcétera, en función de los sensores en calles, GPS, videocámaras y otros. Permite medir el flujo al momento de servicios como el alquiler de bicicletas u otros vehículos públicos para intercambiarlos y reposicionarlos adecuadamente entre sus estaciones. Permite reprogramar por cada etapa del día el número de unidades a entrar en servicio en metro, autobuses y otros transportes públicos. Facilita el envío de alertas oportunas ciudadanas con motivo de sismos, tormentas, inundaciones, etcétera. Permite contar con datos instantáneos de múltiples puntos acerca de contaminación atmosférica para tomar medidas a corto plazo, entre otros.

6.- Sector salud

En el sector salud un ejemplo aplicado reciente y trascendental de los datos masivos ha sido el diseño y desarrollo de las nuevas vacunas contra la pandemia; sin la concurrencia de estos datos intercambiados y analizados en tan corto plazo por variados grupos transdisciplinarios a lo largo de todo el planeta, esas vacunas no hubieran podido ser creadas y producidas en tan corto tiempo.

En los servicios de salud públicos, el análisis de datos masivos permite diseñar políticas y estructuras de salud a nivel internacional, regional y por país, visualizar el estado de salud de grandes sectores, diseñar y seguir campañas de vacunación, sanitarias, etcétera, seccionando por regiones; visualizar y seguir brotes de enfermedades y su tratamiento, predecir eventualidades de salud pública y asignar recursos para su atención; construir modelos puntuales de vigilancia epidemiológica; comprar y mantener grandes inventarios de medicinas y otros recursos, etcétera. Esto incluye colecta y análisis de millones de datos de diagnósticos e imágenes médicas, recetas, datos relacionados con alergias conocidas, demografía, descripciones clínicas y resultados de pruebas de laboratorio.

En los servicios de salud privados, el análisis de datos –como en otros sectores– permite diseñar la estrategia de mercadeo y

seguir las tendencias. Existen además aplicaciones más puntuales, como por ejemplo: medir cuáles tratamientos son más eficaces para determinadas afecciones; identificar los efectos secundarios de los medicamentos, y diseñar nuevos fármacos para diversos propósitos. Tanto en el sector público como en el privado, permite el seguimiento puntual e instantáneo de pacientes hospitalizados en función de sensores de datos en ellos.

7.- Comercio

Éste es uno de los sectores que más ha aprovechado el análisis de datos. Se utiliza para diseñar productos y hacerles seguimiento, coordinar logística de embarque y distribución a nivel local, regional y mundial, llevar gestión de existencias e inventarios. Permite diseñar y lanzar campañas de mercadotecnia segmentadas por grupos de usuarios y analizar con detalle eventos de compra, tendencias y preferencias. Permite hacer seguimiento instantáneo de campañas de venta, lealtad de usuarios, publicidad, etcétera.

Especialmente el comercio electrónico o e-comercio se ha beneficiado de esta herramienta. Los casos exitosos de Amazon, e-Bay, AliBaba, iStore, Mercado Libre, Wish, etcétera, son algunos claros ejemplos de organizaciones que utilizan el análisis de datos masivos para todas las etapas de sus negocios: diseño de productos, mercadotecnia, pago, envío, seguimiento, retroalimentación del usuario, inventarios, etcétera.

8.- Entretenimiento, turismo

Varios servicios del sector de entretenimiento como Spotify, YouTube, Netflix, etcétera, utilizan el análisis de datos masivos colectados de entre sus millones de usuarios en todo el mundo para dar recomendaciones informadas a los usuarios individuales. Amazon Prime ha estado experimentando con los datos masivos para ofrecer una experiencia al cliente presentando video, música y libros de Kindle en un solo lugar, "personalizando" además su oferta con datos anteriores del cliente. Los fabricantes de juegos electrónicos recogen datos masivos para medir el desempeño de sus productos, las preferencias de los usuarios y las tendencias. El análisis de

datos se utiliza también para identificar colectivamente el motivo de la suscripción o la anulación de la suscripción de un producto, así como el interés por un contenido determinado, etcétera.

Los grandes eventos de entretenimiento: juegos olímpicos, campeonatos de fútbol o de tenis, Supertazón, entre otros, utilizan el análisis de datos masivos en tiempo real para medir audiencia, preferencias, etcétera, y hacer ajustes sobre los tiempos de transmisión, mercadotecnia y más.

El sector turismo analiza datos masivos para observar y comparar ofertas y precios de viajes, analizar tendencias y preferencias de los usuarios, hacer recomendaciones a usuarios y prestadores de servicios, etcétera.

9.- Alimentos

Al igual que en otros sectores, el análisis de datos ayuda a diseñar la estrategia de mercadeo y a seguir las tendencias. Pero existen aplicaciones más puntuales; por ejemplo, encontrar patrones y similitudes ocultas que ayudan a los restaurantes a detectar a clientes potenciales. El procesamiento de imágenes y el aprendizaje artificial permiten identificar el lugar más buscado dentro de los restaurantes; los dueños de los restaurantes pueden así destacar el área con fines de publicidad y mercadeo. Muchos restaurantes manejan ya con base en datos su inventario inteligente o su sistema de gestión de existencias para ordenar el resurtido oportuno de las mismas. Cadenas de comida rápida los utilizan para predecir el número de usuarios que habrá a ciertas horas específicas de cada día para que los empleados puedan anticipar la preparación según la hora de demanda. Cuando la fila es larga, les permite modificar automáticamente las pizarras de menús para que se muestren solo los platillos que pueden ser preparados en poco tiempo.

10.- Educación

El sector educativo también se ha visto beneficiado del análisis de los datos masivos. Éste es un sector donde los alumnos y profesores dejan diariamente mucha huella de datos derivados de sus actividades. Entre sus principales usos se distinguen: el diseño de

políticas y estructuras educativas a nivel internacional, regional y por país, y visualizar el estado del sector educativo en grandes cortes. A las instituciones educativas les permiten el diseño y la implementación de nuevos planes de estudio, carreras, asignaturas, tareas, etcétera. Permiten también el seguimiento y la retroalimentación de los sistemas educativos. De hecho, la oficina de publicaciones de la Unión Europea señala este tipo de supervisión educativa como el mayor logro del análisis de datos masivos (Berendt *et al.* 2017). Ayudan también a los directivos y docentes en la producción de Sistemas de Gestión del Aprendizaje, así como conocer y retroalimentar el éxito de los estudiantes en sus cursos. Auxilian a los docentes en la creación de tareas, ejercicios y exámenes utilizando información que ya existe previamente en línea. Permiten también incorporar al proceso de enseñanza a las redes sociales, estudiando los blogs de los estudiantes, cargas o *uploads*, mensajes, "me gusta", etcétera, integrando estas herramientas para medir cuánto les interesa un tema o curso específico.

Parry (2018) estudió cómo las universidades usan datos masivos para atraer y retener a los estudiantes, ayudarlos a elegir sus cursos, y proporcionarles la asesoría pertinente. A los alumnos les permite visualizar los planes de estudios y mapas curriculares de una cierta carrera o plan, y con las recomendaciones pertinentes diseñar su ruta, tiempos, prioridades, etcétera, para la obtención del grado. Un amplio compendio de todas estas aplicaciones en educación puede verse en Baker (2015).

Gran parte de la producción cotidiana de datos de los estudiantes proviene de sus actividades de búsqueda de información académica para sus necesidades; obviamente esta parte tiene que ver muy de cerca con las bibliotecas. Por su naturaleza e interés particular, el análisis de los usos y aplicaciones de datos masivos en bibliotecas será tratado en un apartado especial.

Los anteriores no son todos los sectores donde se utilizan los datos masivos; son solo los principales. Además, si bien en esta lista se incluyeron las bibliotecas en el sector educativo, cabe resaltar que en realidad existen bibliotecas y centros de información en todos y cada uno de los sectores mencionados. Tampoco son todos

los posibles usos del análisis de datos dentro de ellos; es sólo una muestra representativa con algunos ejemplos puntuales para ilustrar todas las posibilidades actuales que ya existen en este tipo de análisis. Por esto es válido suponer que los datos masivos también pueden ser de utilidad e interés para las bibliotecas y su personal profesional y conviene por tanto adentrarse en su estudio.

Datos masivos en bibliotecas

> *En cierto sentido, el mundo de las bibliotecas es un microcosmos del mundo más amplio, ambos impulsados por la tecnología pero amedrentados por lo desconocido, cambiando hacia formas que la mayoría de nosotros entendemos poco.*
>
> MICHAEL GORMAN,
> *Our Enduring Values Revisited: Librarianship in the 20th Century*

Los datos masivos adquieren una presencia cada vez mayor en el ámbito de la información y en las organizaciones relacionadas a ella y por esa razón tienen ya un impacto en todo tipo de bibliotecas. Nicholson y Bennet (2016, 86) mencionan que "[…] el rápido aumento del volumen, velocidad y variedad de los datos de las bibliotecas generados por diferentes herramientas para ellas ofrece formas innovadoras de entender las interacciones con los usuarios en el entorno de estas organizaciones".

En realidad, aunque el fin último de las bibliotecas son los usuarios, hay numerosas aplicaciones de los datos masivos en prácticamente todos sus ámbitos y quehaceres. Existen varios autores que ya han reflexionado en este aspecto, indicando las posibilidades y los proyectos iniciados al respecto. El análisis de datos –en especial los masivos– puede usarse en general en bibliotecas de la misma forma que en otros sectores en alguno de los campos de los datos: minería de datos, aprendizaje de máquina o inteligencia artificial, estadística y visualización. Pero estas son solo grandes áreas de división de los datos de acuerdo con sus técnicas y herramientas. Al cruzarlas con las diferentes áreas de acción de las bibliotecas, esto se convierte en una matriz enorme con múltiples

subcampos de acción específicos. Por ejemplo, la minería de datos aplicada a bibliotecas o "bibliominería"[11] puede usarse para metadatos, para catálogos extendidos, para seguimiento de usuarios, para bibliometría y otros tipos de análisis de impacto de toda clase de materiales documentales, para pruebas de "usabilidad" de colecciones y servicios, para análisis de textos, por mencionar algunos usos. Lo mismo sucede con las demás áreas de análisis de datos al aplicarlas en la biblioteca.

La Federación Internacional de Biblioteca y Asociaciones (IFLA), en respuesta a las conclusiones de su anterior Informe de Tendencias del 2013, propuso la creación del Grupo de Interés Especial sobre Datos Masivos (*Big Data Special Interest Group)* durante su Congreso Mundial de Bibliotecas e Información o WLIC de 2014 en Lyon, Francia, con el propósito de que las bibliotecas se convirtiesen en parte proactiva del movimiento de los datos y no fuesen solo simples espectadoras (IFLA 2014). El grupo fue instaurado formalmente durante el WLIC 2015 en Ciudad del Cabo y ha realizado desde entonces una serie de estudios, eventos y documentos al respecto.

El principal antecedente del uso de datos masivos en bibliotecas –proyecto que continúa hasta la fecha– es el catálogo mundial Worldcat, operado por la organización OCLC, mismo que, de acuerdo con los datos del propio sitio, contenía en 2019 450 millones de registros catalográficos en casi 500 idiomas provenientes de casi 18 mil bibliotecas del mundo; el catálogo consigna también inventarios de 2,800 millones de obras en esas bibliotecas. Además, ha utilizado estructuras de entidad-relación para crear datos enlazados de sus existencias, identificando las entidades en sus registros y luego asignando relaciones entre ellas. Por supuesto, este catálogo no fue concebido en sus orígenes como un proyecto de

11 La bibliominería –*bibliomining*– consiste en las técnicas de minería de datos y de bibliometría que se utilizan conjuntamente para extraer patrones, tendencias, relaciones, etc., significativos a partir de sistemas y datos de biblioteca. Estas técnicas abarcan la identificación de temas u objetos de interés, la creación de un almacén de datos, su refinamiento, proceso, análisis, así como la obtención y evaluación de los resultados.

datos masivos, y no es producto de una sola institución y un momento; es un esfuerzo colaborativo y acumulado, que sin duda representa el arquetipo de los datos masivos en bibliotecas.

Otro esfuerzo colectivo bibliotecario de datos masivos es la biblioteca digital HathiTrust. Beth Plale (2016), co-directora del Centro de Investigaciones de este proyecto lo describió así en ese año: "[…] La colección es de gran tamaño. Para recorrer en 24 horas los 5 mil millones de páginas de sus 14 millones de libros digitalizados se necesitarían 14,000 computadoras funcionando simultáneamente". En 2020, el sitio HathiTrust consigna ya más de 17 millones de ítems digitalizados. Otro ejemplo significativo de datos masivos en el campo de las bibliotecas es el acervo denominado "Archivos de Internet" (*Internet Archives)*, el cual es un sistema sin fines de lucro que comenzó en 1996 para almacenamiento de páginas web con el fin de evitar su desaparición total, y luego se extendió a otros materiales digitales o digitalizados: a la fecha, informa que maneja 330 mil millones de páginas web, 20 millones de libros, 4 millones de audios, 4 millones de videos, 3 millones de imágenes, y 200 mil programas informáticos: en total, más de 45 Petabytes o 45 x 10^{15} bytes de datos (Internet Archive).

Varios autores coinciden en que existen dos grandes vertientes de uso de los datos masivos en las bibliotecas: una directa y otra indirecta. Jharotia (2016, 3) menciona que el efecto directo está en el uso de las herramientas especializadas para analizar los grandes conjuntos de datos provenientes de las bibliotecas en sí mismas. El efecto indirecto es a través de sus usuarios, quienes cada vez más requieren de utilizar productos y servicios derivados de los datos masivos en sus búsquedas de información. Por su parte, Olendorf y Wang (2017, 191-192) afirman algo semejante al respecto: la primera vertiente consiste en usar datos masivos en las bibliotecas como auxiliar en sus operaciones diarias; ellas pueden utilizar estos datos para mejorar sus colecciones, utilizar mejor el espacio, evaluar sus servicios, retroalimentar sus programas de instrucción y optimizar la información proporcionada a sus usuarios. Estos autores también coinciden en que la segunda vertiente en la que las bibliotecas pueden trabajar con datos masivos es la del ámbito de

servicios de búsqueda de información para sus usuarios. Además de ello, algunas bibliotecas ya están comenzando a proporcionar muchos servicios de datos –buena parte de ellos masivos– para investigadores y académicos, tales como diseño y planeación de gestión de datos, colecta de ellos, curaduría, almacenamiento y preservación de datos. Por su interés especial, este tipo de servicios se tratará con mayor profundidad más adelante.

Entrando más en detalle, existen muchas áreas de oportunidad en las bibliotecas respecto a los datos masivos. Blummer y Kenton (2018, 18-19) realizaron un estudio en el que revisaron la literatura especializada que existía al momento acerca de datos masivos específicamente en bibliotecas. Encontraron 76 documentos, de los cuales extrajeron los cuatro grandes temas tratados ahí: 1) gestión de datos masivos en las bibliotecas (29); 2) prestación de servicios de análisis de datos por parte de bibliotecarios (26); 3) documentos informativos acerca de datos masivos (13); 4) oportunidades de capacitación para los bibliotecarios en datos masivos (8). A su vez, encontraron ocho subtemas dentro del tema de gestión de datos masivos: 1) privacidad y gestión de datos; 2) habilidades de los bibliotecarios en la protección de la privacidad; 3) retos adicionales en la gestión de datos masivos; 4) evaluación y detección de necesidades para la gestión de datos masivos; 5) colaboración en la gestión de datos masivos; 6) los diversos factores que fomentan grandes proyectos de gestión de datos; 7) investigaciones y estudios acerca de datos masivos; 8) actividades de los bibliotecarios en datos masivos (Blummer y Kenton 2018, 9). Por supuesto, más documentos se han acumulado desde entonces.

Un número completo de la revista especializada en bibliotecas y tecnología *Library Hi Tech* fue dedicado exclusivamente al tema de los datos masivos en 2018; Liu y Shen (2018) elaboraron una reseña detallada de todos sus artículos al respecto. La afamada revista de bibliotecología *Library Journal* ha dedicado numerosos artículos al tema a lo largo de los últimos años. Véanse como ejemplos representativos *Promise and Problems of Big Data* de Steven (2013) y *What Governmental Big Data May Mean for Libraries* de Schwartz (2013).

Además de todos los documentos publicados, es perceptible también el interés que el tema ha despertado en la profesión bibliotecaria, visible en los numerosos eventos que han sido realizados alrededor de esta temática en años recientes. Un buen número de ponencias y sesiones de las conferencias anuales de la Special Libraries Association desde 2014 a la fecha han estado enfocadas en algún aspecto de los datos.[12] Lawlor (2016) hizo una reseña muy completa de las ponencias de la Conferencia Anual NFAIS[13] de ese año, donde se trataron los datos como parte relevante de la investigación y el conocimiento. La decimoquinta Conferencia Internacional de Ciencias de la Información y la Computación de la IEEE/ACIS en 2016 tuvo algunos páneles donde se trataban los posibles temas de investigación en datos masivos por parte de las bibliotecas (Wang *et al*. 2016).

Una gran parte de los usos de los datos masivos en bibliotecas se trabaja con técnicas provenientes del campo de la "Inteligencia Artificial" o IA. En primera instancia, muchas personas –bibliotecarios inclusive– se imaginan al oír el concepto de IA en la biblioteca algo así como una máquina en la contestadora telefónica respondiendo todo lo que se le pregunte, o un robot "asistente personal" instalado en el vestíbulo del edificio respondiendo a todo lo que se le pida, una especie de "Siri", "Alexa" o "Cortana" en la biblioteca.[14] Aunque en efecto existen algunas bibliotecas que han incursionado en esta variedad de aplicaciones, es meramente una curiosidad tecnológica sin mucha efectividad dado su limitado repertorio, y todavía seguirá en este estadio por buen tiempo. No obstante, sirven muy bien como aplicaciones "estandarte" para

12 Véase como ejemplo: 24th Annual Conference and Exhibition of the Special Library Association. 2018. http://slaagc.org/slaagc2018/.

13 NFAIS = National Federation of Advanced Information Services. Asociación sin fines de lucro de bibliotecarios, editores, académicos, informáticos, etcétera, de la Unión Americana.

14 Estas aplicaciones denominadas "asistentes personales" utilizan procesamiento de lenguaje natural de las personas para responder preguntas, hacer recomendaciones y ejecutar ciertas acciones mediante la conexión hacia un cada vez mayor conjunto de servicios web y otras aplicaciones.

llamar la atención de los usuarios hacia otros servicios de la biblioteca. Véase como ejemplo de ello Harada (2019).

Existen en la práctica muchas variantes y aplicaciones de la Inteligencia Artificial, desde máquinas que juegan ajedrez, robots para fabricación industrial, máquinas para control de tráfico, sistemas de reconocimiento de imágenes, automóviles sin conductor, etcétera. Se distinguen en la disciplina dos grandes campos de estudio: el campo teórico y las aplicaciones específicas. El primero tiene que ver con todo el concepto y teoría del comportamiento y capacidades "inteligentes" de las máquinas, su filosofía y deontología, etcétera. El segundo trata acerca de la construcción de aplicaciones puntuales para resolver ciertos problemas específicos; esta segunda variante es la de interés en las bibliotecas, y no es tan lejana como pudiera pensarse.

Existen muchas aplicaciones en el quehacer cotidiano. De hecho, cualquier computadora personal o teléfono inteligente dispone de una buena cantidad de aplicaciones utilizadas por el público de entre las cuales los usuarios no están conscientes que son de IA; por ejemplo, el sistema que da instrucciones para manejar en auto de un punto a otro de la ciudad, ofreciendo opciones de menor tiempo, menor distancia, ruta más económica, etcétera. Otro ejemplo muy visible de estas aplicaciones consiste en las sugerencias de sitios de compra –tales como Amazon–, que indican al usuario detalles personales como "las personas que compraron esto también adquirieron esto otro", o sugieren productos semejantes de acuerdo con las visitas previas del usuario. Otro ejemplo cotidiano de ello son los sistemas de sugerencia de ortografía y estructura que existen en los procesadores de texto. Un ejemplo más de usos aplicados de la IA son los programas de traducción de textos, de los cuales el más conocido es Google Translator, pero donde ya hay productos realmente impresionantes mucho más allá de la simple traducción literal, tales como *DeepL, Reverso Traslation, Wordlingo, BabelFish* o *Traslation2.*

Los antecedentes de la IA pueden remontarse hasta las figuras o muñecos "autómatas" que han existido desde más de hace

veinte siglos.[15] En la época moderna, se considera a Alan Turing como el padre de la IA. Él fue el inventor de la máquina Bombe para la decodificación de mensajes cifrados de los alemanes durante la Segunda Guerra Mundial, y diseñó la famosa "Prueba de Turing", un criterio mediante el cual puede evaluarse la inteligencia de una máquina cuando sus respuestas en la prueba no pueden diferenciarse de aquellas dadas por un humano. La acuñación del término "Inteligencia Artificial" se atribuye a John McCarthy en 1956, quien lo introdujo en la primera reunión sobre el tema; él estableció ahí: "[…] cada aspecto del aprendizaje o cualquier otra característica de la inteligencia puede, en principio, describirse con tanta precisión que se puede construir una máquina para simularlo" (Dartmouth… 1956 s.p.).

El concepto más básico de IA define cualquier tipo de inteligencia que no surge a través de procesos naturales, o donde la inteligencia puede ser entendida y medida de tal forma que puede ser recreada. Existen muchas definiciones del concepto, dependiendo del enfoque que se pretende y la variante de la disciplina que lo ocupa. Para el campo de acción de las bibliotecas interesa –más que la definición de Inteligencia Artificial– la definición especializada y el estudio de algunos de sus subcampos más utilizados en ellas. Contrario a lo que pudiese pensarse, el interés de las bibliotecas en el campo de la IA no es nuevo; hace tres décadas, en 1991, Charles Bailey ya elaboró una interesante y completa reseña de las aplicaciones de esta tecnología en las bibliotecas (Bailey 1991).

Existen numerosas aplicaciones de la IA en las bibliotecas, las cuales conviene estudiar y dividir de acuerdo con los grandes capítulos temáticos de las colecciones y los servicios de las bibliotecas que merecen tratarse más específicamente.

---

15 Herón de Alejandría escribió un libro en el siglo I de nuestra era donde explicaba la creación de mecanismos para entretenimiento que imitaban el movimiento humano o de animales, abrían puertas automáticamente, etc. Él detalla ahí su "Teatro automatón" con marionetas mecánicas. Ahmad, Muhammad y Hasan bin Musa ibn Shakir compilaron en el año 805 un libro que describe más de cien mecanismos y autómatas. Hay muchos textos más.

Datos masivos en taxonomías y metadatos de las bibliotecas

> *Pocas personas, fuera del mundo de los expertos bibliotecarios y conservadores de museos, saben lo manejables que pueden ser los hechos bien ordenados, por multitudinarios que sean, y lo rápido y completo que pueden recapitularse, hasta las visiones más raras y los asuntos más recónditos, una vez que se han colocado en un esquema bien ordenado de referencia y reproducción.*
>
> H. G. WELLS,
> "World Brain: The Idea of a Permanent World Encyclopaedia", 1938

Conviene en este punto ir analizando con más detalle y con ejemplos dónde pueden encontrarse los datos masivos en el ambiente de las bibliotecas. Como ya se mencionó, existen numerosas actividades, tareas y servicios en la biblioteca en los que los datos masivos pueden ser aplicados y aprovechados en beneficio de ellas: estudios y análisis de las colecciones y los servicios, de usuarios, de aprendizaje profundo, en sistemas expertos, traducción, OCR, en proyectos de análisis de textos con muy variadas herramientas para muy diversos propósitos; para asistentes robóticos, por mencionar algunos. Muchas vertientes específicas pueden encontrarse dentro de estas líneas generales.

En primera instancia, el uso de datos masivos en bibliotecas puede hallarse en el diseño y creación de esquemas de metadatos y de nuevas taxonomías de información. 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 de entes estériles con poca o nula utilidad. En la citada serie de estudios de la corporación IDC acerca de datos masivos, en el correspondiente al año 2014 ellos establecieron que se agregaban metadatos de una manera

sistematizada solo al 3 por ciento de la inmensa cantidad de datos que se estaba produciendo en el mundo (*The Digital Universe…* 2014). Los metadatos son importantes en todo tipo de estructura de información, pero se vuelven cruciales en el ámbito de los datos masivos, ya que informan todo acerca de los datos: qué son, quién los generó, cómo, cuándo, dónde y por qué se generaron. Y no solo informan de los datos en sí mismos, sino también de sus elementos asociados: transacciones, formularios, programas o aplicaciones, 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 y llegan a ser tan exhaustivos que en realidad hablamos ya de "metainformación". Zeng y Qin (2008, 15) establecieron cuatro tipos de estándares de metadatos usados actualmente en la práctica bibliotecaria:

- "Estructuras", como el Conjunto de Elementos de Metadatos del Núcleo de Dublín.
- "Contenido", como las Reglas Anglo-Americanas de Catalogación.
- "Valores", como la lista de encabezamientos de nateria de la Biblioteca del Congreso de Estados Unidos.
- "Intercambio", como el Formato Marc 21 para datos bibliográficos.

Como ya se mencionó, dependiendo de la naturaleza del dato y del emisor, los datos pueden ser estructurados, semi-estructurados o no estructurados. En función de ello, pueden existir muchos o pocos metadatos embebidos en cada tipo de datos, pero con frecuencia esos metadatos no son evidentes para todo el mundo; es necesario conocer a fondo su estructura y su esencia para poder extraer algo coherente de ellos. Por ejemplo, Schmarzo (2018) estableció que existe la sorprendente cantidad de veinte metadatos asociados a cada mensaje de Twitter, más allá del contenido en sí mismo. Pocas personas caen en cuenta de esta enorme cantidad de metadatos asociados a algo tan sencillo como un tuit de 280

caracteres. Es conveniente resaltar en este punto que para los que estudian este tipo de comunicación, lo que se dice en el mensaje como tal no tiene ningún valor estadístico, pero esos eventuales veinte metadatos representan una mina de oro para el análisis de este tipo de red social.

Desde tiempo atrás, los bibliotecarios han estado conscientes del gran valor de los metadatos en el mundo de la información y por este motivo están familiarizados con su diseño, creación y uso. Igualmente, han creado y utilizado desde tiempos inmemoriales toda clase de taxonomías de la información.[16] Como es sabido, Cutter y Dewey desarrollaron desde el siglo XIX sistemas al respecto con el objeto de unificar las reglas de registro y descripción en diferentes bibliotecas, y esos no fueron los primeros antecedentes de estos esfuerzos.[17] Ha sido un proceso incesante que en décadas recientes 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. Este concepto de "ontología" no es absoluto ni monolítico, y varía sensiblemente con el enfoque de la disciplina que los define. Además de esos diversos enfoques posibles, Souza y sus colegas (2011) han establecido diversos *niveles* o *profundidades* de "precisión ontológica"; del más simple al más complejo, ellos distinguen: 1) Lexicón

16 "Una taxonomía es un vocabulario controlado que se organiza en una jerarquía. Cada término designa una categoría, tipo o clase. Solo hay un tipo de vínculo, que significa 'es una variedad de' y corresponde a una relación de subclase. Hablando estrictamente, cada nodo en una taxonomía tiene exactamente un 'padre', pero el término 'taxonomía' a menudo se refiere a las jerarquías con múltiples 'padres'. A veces también se utiliza para referirse a las redes con más de un tipo de enlace" (Uschold y Grüninger 1996, 94).

17 Frederick Rostgaard escribía teorías de clasificación documental desde 1697. Los sistemas de organización documental del Vaticano, el de Jacques-Charles Brunet, el de Antonio Panizz (1841), el de William Harris (1870), el decimal de Dewey (1876) y el de la Biblioteca del Congreso de Cutter (1891) son 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.

o vocabulario con definiciones en lenguaje natural; 2) taxonomía simple formada de diccionarios de datos y jerarquías; 3) tesauro o taxonomía con términos relacionados; 4) modelo relacional, el cual consigna restricciones de tipos y relaciones arbitrarias entre entes, y 5) teoría axiomática completa.

Todos estos niveles de precisión y descripciones han sido explorados y desarrollados por profesionales bibliotecarios en tiempos recientes. Algunos otros autores dividen de otra forma los niveles o profundidad para las ontologías. En términos generales, clasifican como "ontologías ligeras" a aquellas que comprenden únicamente las etapas de vocabulario, clasificación o tesauro, y como "ontologías complejas" a las que ya incluyen axiomas, restricciones, etcétera. Milton (1998, 86-88) las divide entre ontologías centradas en la teoría y aquellas que son orientadas pragmáticamente. Para este autor, las primeras son aquellas creadas a partir de una cierta teoría científica, humanística o social y por tanto hacen énfasis en ella, mientras que las segundas son aquellas emanadas de la práctica consensuada de una disciplina, y por lo general se diseñan pensando en que eventualmente podrán ser utilizadas por sistemas informáticos. Estas últimas son por tanto las más comunes en la práctica de las ciencias de la información y están dirigidas a áreas específicas prácticas como es el caso de la Bibliotecología o la Archivística.

Un ejemplo interesante y actual de estos desarrollos son los modelos conceptuales subyacentes de las RDA (Recursos, Descripción y Acceso), el estándar de catalogación para la formulación de registros bibliográficos usado en bibliotecas, archivos, museos, etcétera. Como es sabido, RDA es un conjunto de directrices, elementos de datos e instrucciones para crear metadatos de recursos bibliotecarios y del patrimonio cultural correctamente formados de acuerdo con los modelos internacionales para aplicaciones de datos enlazados orientadas al usuario. Esos modelos conceptuales subyacentes de las RDA son los Requisitos Funcionales para Registros Bibliográficos o FRBR, los Requisitos Funcionales para Datos de Autoridades o FRAD, los Requisitos Funcionales para Datos de

Autoridades de Temas o FRSAD, y la ontología press,[18] avalados por la IFLA y consolidados con su "Modelo de Referencia de Bibliotecas" (*Library Reference Model)*. Si se desea abundar en estos modelos conceptuales, véase la página de la federación *IFLA's Bibliographic Conceptual Models (*https://www.ifla.org/node/2016), donde se encuentra una buena reseña de *todos ellos*.

El punto central de interés de todo ello consiste en que llevar de la teoría a la práctica cada uno de esos modelos conceptuales implica el manejo de grandes cantidades de datos. La manera más rápida y completa de construir los vocabularios con definiciones en lenguaje natural, las taxonomías simples –diccionarios de datos y jerarquías–, los tesauros o taxonomías con términos relacionados, los modelos relacionales con atributos, restricciones, relaciones; los modelos de requisitos funcionales, etcétera, para cada disciplina ha sido a través de la colecta, proceso y análisis de grandes cantidades de elementos al respecto. Dado que todos los elementos taxonómicos mencionados deben construirse para cada uno de los campos del conocimiento humano, la tarea apenas ha comenzado, y falta todavía la mayor parte por realizar: un campo potencial de estudio y desarrollo sumamente extenso.

Cabe subrayar en este punto que todos estos modelos conceptuales no son solo ideas teóricas utilizables en disquisiciones académicas: tienen innumerables usos y aplicaciones prácticas para optimizar el registro y la recuperación de información tanto dentro de las bibliotecas, como en el ámbito que la red mundial impone actualmente. De hecho, puede asegurarse que esas aplicaciones prácticas son parte fundamental de lo que mantendrá a las bibliotecas en un futuro cercano en el concierto de la información universal, y por eso conviene profundizar en su análisis con más detalle.

---

18 Press es una ontología formal diseñada para representar la información bibliográfica sobre recursos con continuidad, y más específicamente sobre publicaciones seriadas (revistas, periódicos, etc.). Tiene como objetivo proponer respuestas a problemas añejos con la aplicación de la familia de modelos FRBR a esas publicaciones seriadas y recursos continuos.

Datos masivos en los catálogos

> *Sabio es aquél que conoce las fuentes del conocimiento; dónde ha sido escrito y dónde puede hallarse.*
> Archibald A. Hodges

De inicio, muchos pudieran pensar que los datos masivos están en los catálogos, dados sus grandes volúmenes actuales. Enormes como pueden ser en la actualidad, sobre todo en bibliotecas como la Británica, la del Congreso de Estados Unidos, o la Nacional de Francia; o en grandes compilaciones como OCLC o HathiTrust, los datos masivos no se encuentran en realidad en sus catálogos, pero sí en toda la información asociada a ellos.

Los catálogos correspondientes a las colecciones de las bibliotecas poseen intrínsecamente una inmensa cantidad de datos vinculados entre ellos que conforman una red de datos masivos que no es evidente a simple vista. En esos catálogos, se encuentran embebidos innumerables autores –personas y organizaciones–, eventos, lugares, editoriales, épocas, temas, fechas, citas, etcétera, pero más importante aún: conforman entre todos ellos un inmenso entramado de interrelaciones que no son perceptibles o extraíbles fácilmente, y que no se dan unitariamente en un solo registro: solo existen en el conjunto.

Además, los catálogos de bibliotecas suelen estar separados por tipo de material: catálogos de libros, de revistas y sus tablas de contenido, de tesis, de imágenes, de audio, verticales, etcétera, lo cual dificulta todavía más percibir y establecer las interrelaciones entre los datos de diferentes catálogos, ya que generalmente son entes totalmente separados entre sí, especialmente cuando la biblioteca maneja catálogos para materiales "tradicionales" y para materiales digitales. Estas interrelaciones no evidentes son las que en realidad conforman un gran conjunto de datos masivos que sin ser el catálogo provienen de él, y que son materia de importantes eventuales estudios en las bibliotecas, ya que ofrecen inéditas y poderosas opciones para la búsqueda y recuperación de

información. Todos estos desarrollos tienen su origen en el Protocolo de la Iniciativa de Archivos Abiertos para la Recolección de Metadatos (*Open Archives Initiative Protocol for Metadata Harvesting*, OAI-PMH) del 2015 (Open Archives 2014), originalmente elaborado en 2002 con sucesivas actualizaciones a la fecha, el cual marcó las pautas para la colecta, el análisis y la interrelación de metadatos internamente en una biblioteca entre conjuntos de ellos, así como con los provenientes de editores, proveedores, etcétera. Numerosos proyectos mucho más evolucionados se han derivado desde entonces basados en esta iniciativa.

Además de ello, existen iniciativas alrededor del concepto de "Datos Enlazados" (*Linked Data*). Estos son un conjunto de estructuras y principios de captura y registro para compartir globalmente datos dispares interconectados en la web, que son legibles por máquina. La teoría básica de los datos entrelazados consiste en que los datos tienen más valor entre más puedan relacionarse con otros datos en el entorno global de la red mundial; cuantas más cosas, eventos, personas, lugares, etcétera, estén conectados entre sí de manera estructurada, más poderosa será la red de datos, independientemente de que provengan de diversas fuentes y sus formatos no sean homogéneos.

El propósito de este principio es potenciar el descubrimiento de conocimientos y la eficacia del análisis de datos. Las estructuras de este tipo de datos han sido definidas bajo los estándares HTTP, RDF y URI como una de las bases de la web semántica propuesta por Tim Berners-Lee y el Consorcio W3C desde 2006.[19] Existe un gran tema común entre los metadatos de bibliotecas y los postulados de la web semántica: cómo lograr que las relaciones implícitas que se encuentran en los metadatos tradicionales de las bibliotecas –que son obvias para los humanos– sean también

19 El término "datos enlazados", también llamados "datos vinculados" (*linked data*) es atribuido a Sir Tim Berners-Lee, considerado el creador de la World Wide Web, en su nota: "Linked Data Web architecture: Design Issues" (2006). Última actualización: 18/06/2009. Él menciona ahí un estilo de publicación en la Web con datos estructurados interrelacionados.

suficientemente explícitas para que las entiendan las máquinas. Dada la relevancia del tema, el Consorcio W3C creó un grupo de interés denominado Grupo Incubador de Datos Enlazados en la Biblioteca, el cual elaboró un detallado estudio al respecto cuyos resultados fueron publicados en un informe final. El documento expresa como objetivo:

> […] ayudar a aumentar la interoperabilidad global de los datos de las bibliotecas en la web, reuniendo a las personas involucradas en las actividades de la web semántica que están centradas en datos enlazados dentro y fuera de la comunidad bibliotecaria, aprovechando las iniciativas existentes e identificando vías de colaboración para el futuro (W3C 2011).

El documento hace reseña de algunos proyectos en curso al respecto y presenta interesantes recomendaciones acerca del tema. Heery mencionó como las principales similitudes entre los metadatos tradicionales de las bibliotecas y los datos enlazados:

> [...] lo que quizá sea el aspecto más llamativo de la web semántica para la comunidad bibliotecaria es la coincidencia entre los intereses tradicionales de la gestión de la información y de las bibliotecas –construir vocabularios, describir propiedades de los recursos e identificarlos, intercambiar y agregar metadatos– y las preocupaciones que están impulsando el desarrollo de las tecnologías de la web semántica (Heery 2004, 70).

Ávila menciona al respecto:

> […] la integración de Linked Data en el registro bibliográfico tiene dos propósitos esenciales. Por un lado, vincular los datos de las bibliotecas con otras fuentes de datos disponibles en la web. Por otra parte, propiciar la generación de un método para la óptima recuperación de la información en las bibliotecas, acorde a las demandas actuales de los usuarios (Ávila 2020, 8).

Dada la importancia del tema, la IFLA (2017) ya ha compendiado también el uso de los datos enlazados en bibliotecas.

Como ejemplos representativos de metadatos con datos enlazados, la Biblioteca Británica y la Biblioteca del Congreso de Estados Unidos ya han comenzado a estudiar estos vínculos de datos entre sus cientos de respectivas colecciones –que implican muchos millones de ítems– para tratar de modelar las interrelaciones entre personas, eventos, lugares, etcétera, contenidos en sus acervos. Véanse como ejemplos el "Modelo de datos para libros de la Biblioteca Británica" (British Library Data Model–Books*)*, y el "Servicio de Datos Enlazados de la Biblioteca del Congreso de Estados Unidos", *Library of Congress Linked Data Service* (Library of Congress s.f.), así como la teoría de este concepto en la visión de Hallo y colegas (2015). Básicamente, estas instituciones han extraído cantidades inmensas de datos de sus catálogos armando matrices de interrelaciones que se dan entre ellos, creando numerosos mapeos con esas interrelaciones para formar conjuntos inéditos de metadatos.

Springer Nature, la división en ciencias naturales de editorial Springer, está construyendo un proyecto denominado SN SciGraph alrededor del concepto de "datos abiertos enlazados" (Springer Nature s.f.). Este proyecto consiste en un "descubridor" de ciencias naturales, el cual compila datos de las publicaciones de esta editorial en este campo junto con los de otras instituciones académicas asociadas. La base de datos recopila información de artículos de investigación, libros y capítulos, conferencias, citas, instituciones, investigadores, etcétera, y hace vinculaciones semánticas entre ellos. Ellos afirman que eventualmente tendrán cerca de dos mil millones de elementos vinculados. La empresa OCLC también ha estado haciendo algunos desarrollos acerca de datos enlazados provenientes de sus catálogos; tiene un proyecto denominado WorldCat Linked Data, del cual ya ha publicado algunos subproductos: VIAF (*Virtual International Authority File)*, que es una estructura de catálogos de autoridad vinculados, y FAST (*Faceted Application of Subject Terminology)*, el cual vincula de manera especial encabezamientos de materia provenientes de la Biblioteca del Congreso de los Estados Unidos (OCLC s.f.).

Aunque todos estos proyectos todavía se encuentran en etapas incipientes debido al inmenso número de datos a extraer y

correlacionar, los resultados parciales obtenidos son muy interesantes y alentadores: ciertamente el concepto general de este tipo de iniciativas puede establecer dimensiones inéditas sumamente útiles y poderosas en el ámbito de las bibliotecas, y sin duda caen en el entorno del manejo de datos masivos, pues en efecto contienen las características establecidas de inicio para ellos: volumen, velocidad y variedad.

El *volumen* de datos que pueden contener las interrelaciones de todos esos elementos –autores, eventos, lugares, temas, épocas, editoriales, fechas, citas, etcétera– fácilmente puede ascender a varios millones, sobre todo si se involucran varios catálogos de una biblioteca simultáneamente: libros, revistas, tesis, mediatecas, etcétera. La *velocidad* a la que cambian los datos y por ende sus interrelaciones es muy rápida, ya que diariamente se agregan nuevos elementos a todos esos documentos interconectados pertenecientes a una cierta biblioteca y, en consecuencia, el mapa de las relaciones cambia. Finalmente, la *variedad* también es sumamente amplia, pues las formas estructurales de esos datos –por más que en las bibliotecas están normalizados– no son iguales entre catálogos diferentes, y pueden diferir todavía más con información que proviene de fuentes exteriores: tablas de contenido, sistemas de citas, catálogos de autoridad, ontologías externas, etcétera.

La afirmación de que este tipo de proyectos puede establecer dimensiones inéditas de gran utilidad práctica y potencia en el ámbito de las bibliotecas se ha venido consolidando cada vez más en los últimos años. Estudios realizados recientemente en innumerables bibliotecas en todo el mundo señalan que uno de los principales factores de satisfacción o frustración de los usuarios en una gran cantidad de ellas es precisamente su buscador interno de información dentro de sus catálogos por tres razones principales: en primer lugar, muchas bibliotecas aceptan el primer buscador que les es ofrecido como parte de su Sistema Integrado de Gestión Bibliotecaria (*Integrated Library System* o ILS), sin verificar si cuenta con los elementos mínimos requeridos: operadores booleanos términos exactos, palabras truncadas o "comodines", delimitación por fechas, idiomas u otros parámetros, etcétera.

Como regla general, esos buscadores de biblioteca realizan la recuperación más básica posible con todos los registros que contengan cualquiera de las palabras en cualquier posición, y presentan todo lo que se parezca a lo introducido por el usuario sin discernimiento; así recuperan enormes conjuntos de información poco pertinente y de poca utilidad e interés. En segundo lugar, las costumbres de búsqueda de los usuarios han cambiado sustancialmente debido a la web, y por lo general los buscadores usuales de biblioteca solo pueden hacer búsquedas sobre los campos "típicos" de los registros catalográficos: autor, título, tema, editorial, serie, etcétera, que los usuarios consideran ya muy limitados para encontrar lo que buscan. En tercer lugar, por regla general los buscadores operan de forma vertical en los catálogos, uno por uno, por lo que el usuario debe repetir su búsqueda en cada uno de ellos para abarcar todos exhaustivamente.

Conscientes de este problema y de los nuevos usos de los datos, muchas bibliotecas ya están agregando grandes conjuntos de datos adicionales a sus catálogos optimizando con ello radicalmente sus buscadores. Por ejemplo, capturan y agregan a ellos las tablas de contenido y el glosario de cada uno de sus libros asociando estos términos al registro catalográfico original. Esto potencia inmensamente la búsqueda, pues el buscador no dispone ya solamente de unas pocas palabras del autor, título o tema, sino muchas palabras contenidas en el índice o glosario de ese libro.

Algunas bibliotecas especializadas –por ejemplo, en literatura– no agregan tablas de contenido, ya que éstas casi no existen y aportan poco en ese tipo de publicaciones, pero extraen y agregan todos los personajes, lugares, épocas, eventos, etcétera, consignados en novelas, obras de teatro, y otras similares. Algo semejante puede hacerse en casi todas las disciplinas, adaptándolo al contexto y características de cada una de ellas; por ejemplo, en la química se puede extraer del índice y agregar al registro original fórmulas, procesos, sustancias, etcétera, útiles para el buscador; esto puede adaptarse para muchas otras disciplinas. Si la biblioteca cuida además que las búsquedas puedan hacerse horizontalmente a lo largo de varios de sus catálogos, y verifica que

su buscador cuente con los elementos pertinentes de búsqueda y acotamiento: operadores booleanos, términos exactos, palabras truncadas o "comodines", delimitación por fechas, idiomas u otros parámetros, cambios de formatos de despliegue de las fichas, etcétera, la biblioteca tendrá una herramienta formidable de búsqueda que competirá muy ventajosamente con búsquedas en la web o en otras fuentes. Con respecto a este último punto acerca de las opciones deseadas en los buscadores útiles para los usuarios, Markey (2007), Borgman (2007), Calhoun (2006) y De Rosa (2006), entre otros autores, han elaborado amplias y variadas listas al respecto. La Biblioteca Nacional de Nueva Zelanda preparó una interesante "lista de verificación" o *checklist* para que los bibliotecarios puedan verificar todos los atributos y capacidades de un cierto ILS que pretendan adquirir. La lista es muy completa y está bien desarrollada de acuerdo con los elementos considerados deseables hoy en día (New Zealand National Library 2017).

Algunas bibliotecas están agregando ya a su buscador diversos elementos interrelacionados como los anunciados previamente: autores, eventos, lugares, citas, etcétera, construidos además horizontalmente a lo largo de varios catálogos. He aquí el inmenso valor de las interrelaciones en los catálogos de bibliotecas: los modelos de datos interrelacionados permiten que el buscador informe al usuario que un cierto autor buscado –por ejemplo, en el catálogo de libros– aparece en otros catálogos, o que ese autor es citado en otros textos, sean libros, artículos, tesis, etcétera. O que ese autor publica generalmente con ciertos otros autores. Lo mismo puede hacerse con temas u otros parámetros. El buscador puede así informar también que las personas que consultaron cierto libro también consultaron otros relacionados, apuntando hacia ellos; las posibilidades son infinitas. El punto central es que ciertos datos adicionales agregados a los catálogos pueden potenciarlos enormemente. Esa extracción de datos, sus interrelaciones y sus agregados no puede hacerse por métodos manuales o computacionales básicos: requieren de un tratamiento especial que se acerca sin duda a los datos masivos por su volumen, su velocidad de cambio y sus diversas estructuras.

Estos proyectos aplicados son sólo una muestra de las posibilidades; Schilling (2012) elaboró una buena reseña acerca de cómo transformar metadatos en datos enlazados dentro de las bibliotecas. Ávila (2020-2) desarrolló un excelente compendio acerca de los modelos y plataformas de datos enlazados en las bibliotecas. Los proyectos mencionados al respecto: el Modelo de datos para libros de la Biblioteca Británica, el Servicio de Datos Enlazados de la Biblioteca del Congreso de Estados Unidos y los catálogos expandidos mencionados previamente son solo unos pocos ejemplos de todo lo que puede lograrse en la práctica con el uso de datos masivos y metadatos para la optimización mayúscula de catálogos en bibliotecas. Cabe resaltar que además este tipo de desarrollos sin duda son ejemplos totalmente aplicados y tangibles del concepto teórico de las bibliotecas semánticas y los datos enlazados.

Datos masivos en los estudios métricos de la información documental

> *Ya no estamos en la era de la información. Estamos en la era del manejo de la información.*
>
> Chris Hardwick

Los datos masivos se encuentran también en 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.[20] 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.

20 Bibliometrics, Archival metrics, Informetrics, Librametry, Scientometrics, Webometrics y Altmetrics, respectivamente.

Los factores que rigen el impacto de los artículos en revistas académicas por supuesto han tenido y tienen fuertes repercusiones en la comercialización, demanda y precios de esas revistas, así como en el prestigio, estímulos, y reconocimiento de sus autores y sus instituciones. Por este motivo, han sido desde hace décadas elementos de gran interés para las editoriales, los centros de investigación, las bibliotecas, etcétera, y dentro de esas organizaciones para los autores, editores, bibliotecarios y responsables de políticas de investigación, en todas las disciplinas. Estos factores se mantuvieron estables durante décadas, pero en los últimos años han sido cuestionados, y han entrado en juego nuevos elementos que han iniciado un cambio radical en su conceptualización: las revistas y los repositorios de investigación en acceso abierto, las redes sociales entre investigadores, el desglose de las estructuras de citación en categorías diferenciadas más finamente, por mencionar algunos. De hecho, la webmetría y la altmetría como especialidades de la métrica documental han surgido en épocas recientes con la web.

Además de todo ello, las nuevas técnicas de análisis de datos han introducido nuevas aproximaciones y enfoques a los estudios de la información documental. Gorbea (2013, 154) mencionaba desde ese año al respecto:

> […] los indicadores bibliométricos con frecuencia han sido utilizados para fundamentar el nivel de desarrollo científico alcanzado por determinada disciplina, institución o país, práctica que ha propiciado las propuestas de políticas científicas y de información sustentadas en el reconocimiento de altos niveles de productividad, impacto, visibilidad y crecimiento de la literatura científica generada en estas instancias. Esto ha acarreado consigo que siempre aparezcan mejor representados por este tipo de indicadores aquellas disciplinas, fuentes de información, instituciones y países que se encuentran en la denominada 'corriente principal' […] el comportamiento anterior ha suscitado controversia sobre la validez y utilidad de este tipo de indicadores en la evaluación de la ciencia, motivo por el cual el estudio, reconocimiento y definición de los indicadores bibliométricos en la literatura especializada es extenso.

Las grandes capacidades y herramientas actuales de colecta, manejo y uso de datos han ido ampliando las posibilidades de este tipo de modelos y técnicas para crear nuevos campos de aplicación y nuevos tipos de estudios al respecto. Moed (2012, 4) menciona que

> […] la disponibilidad de artículos de investigación de texto completo en formato electrónico trae la oportunidad de realizar análisis textuales de todo el contenido de un artículo, y no solo de los metadatos extraídos mediante la indización de bases de datos. Los contextos de las citas pueden analizarse lingüísticamente, y se pueden realizar análisis de sentimiento para revelar cómo el autor que cita valora una obra citada.

Diversos autores, como Small y Klavans (2011), y He y Chen (2018) establecieron que utilizando técnicas de minería de datos en textos, análisis visual, etcétera, era posible obtener resultados interpretables de los diferentes contextos de las citas en revistas científicas, siempre y cuando los datos estuviesen suficientemente estructurados. Esto permitía diferenciar las citaciones en esas revistas en "pesos" y categorías diferentes para elaborar diferentes tipificaciones de esas citas. He y Chen mencionan asimismo que los contextos de citación también se utilizaron con el fin de caracterizar publicaciones para diversos usos aplicados, tales como el resumen de publicaciones (Qazvinian *et al.* 2010), la generación de artículos de encuesta (Mohammad *et al.* 2009) y la recuperación de información (Huang *et al.* 2015). Lourdes Feria (2020) describió un interesante ejemplo aplicado de bibliominería para diagnóstico de usuarios en México.

Otros autores, como Duy y Vaughan (2006, 2005 y 2003) han establecido que, teniendo volúmenes considerables de datos, el análisis del acceso y uso de las revistas científicas es una alternativa más exacta para medir su impacto en lugar del tradicional análisis de citas; ellos también realizaron estudios para verificar hasta qué punto los índices de impacto y uso de ciertas revistas científicas coincidían con aquellos proporcionados por los editores. Esos estudios y sus resultados son muy útiles también en las bibliotecas para retroalimentar la eventual selección de suscripciones a las

revistas, y ayudar a la toma de decisiones acerca de la renovación o cancelación de ellas. Más aún, se han comenzado ya a hacer análisis de impacto no tan solo de revistas, sino inclusive de libros; véase al respecto Halevi, Nicolas y Bar-Ilan (2016).

Halevi (2014) compiló un resumen muy interesante de los tipos de uso de los datos masivos en bibliometría, y los categoriza en cinco: citas, referencias, palabras clave, uso, y análisis de textos completos. Ella agrega que:

> […] la disponibilidad de los datos y las capacidades tecnológicas dieron lugar a una fuerte proliferación de bases de datos bibliométricas y mejores instrumentos de análisis de datos para el desarrollo de: indicadores de evaluación científica más sofisticados y personalizados; mediciones del comportamiento de investigadores y editores de revistas; indicadores de impacto social de la investigación, tanto en su valor académico como en su contribución a la ilustración del público en general; creación y análisis de macro-conjuntos de datos mediante la combinación de múltiples conjuntos de ellos.

Otros autores han investigado al respecto los patrones internos, la distribución geográfica, las revistas e instituciones influyentes, la colaboración internacional entre los autores, las instituciones y los países o territorios de las grandes publicaciones de datos. Se ha estudiado la coincidencia y las correlaciones de las palabras clave de los autores de las publicaciones. Existen estudios bibliométricos acerca de diversas disciplinas que han realizado con datos masivos lo mismo que habían hecho tiempo atrás con datos reducidos: detectar las tendencias y temas de actualidad, sus dispersiones, los grupos de investigación, etcétera. La diferencia consiste en que al emplear cantidades mucho mayores de datos y al usar nuevas herramientas de análisis, se obtienen resultados diferentes a los que se lograban con herramientas "tradicionales": en especial, se pueden obtener aproximaciones más finas y detalladas.

Cabe resaltar en este punto que el análisis de textos o minería de textos no se limita solo a los estudios de metría. Con la utilización de la minería de datos, técnicas lingüísticas, estadísticas, de

aprendizaje de máquina, recuperación de información, comprensión del lenguaje natural, razonamiento basado en casos, y otras más, este tipo de estudios puede ayudar a las personas y organizaciones a obtener nuevos conocimientos extrayendo información significativa a partir de grandes cantidades de textos documentales sin estructura 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. Así, 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, tendencias en textos, por citar algunas.

Como puede verse, el uso de datos masivos ha abierto nuevas posibilidades y campos de interés para este tipo de metrías y sus profesionales dentro de las bibliotecas.

Datos masivos en el aprendizaje de máquina en bibliotecas

> *La ambigüedad no es hoy en día la falta de datos, sino un diluvio de datos.*
> Paul Gibbons, 2015

El siguiente campo de acción del análisis de datos en la biblioteca consiste en uno de los subcampos de la Inteligencia Artificial (IA): el denominado "aprendizaje de máquina" (*machine learning*), también llamado "aprendizaje automático",[21] en el cual se diseña y programa un cierto sistema específico de este tipo de IA para que sea susceptible de ser enseñado, entrenado o preparado para

21 El término "aprendizaje de máquina" se atribuye a Arthur Samuel en 1959, pionero en juegos de computadora e Inteligencia Artificial. Él lo definió como "la capacidad de las computadoras de aprender sin necesidad de una programación explícita".

realizar diversas acciones opcionales sin la intervención humana directa; estos sistemas específicos reciben datos que pueden interpretar, y de los cuales extraen patrones significativos; dependiendo de esos datos y sus interpretaciones, un cierto sistema responderá en una u otra forma (*Artificial Intelligence…* 2018, 6).

Todo programa de cómputo puede tomar decisiones simples de acción de acuerdo con las llamadas "proposiciones condicionales" o "expresiones condicionales", las conocidas como *if… then… else*. La diferencia básica con los sistemas de aprendizaje de máquina consiste en que estos últimos utilizan además decisiones complejas de ponderación de resultados conforme van procesando más y más datos, asignando "pesos" e iterando una y otra vez las posibles decisiones al retroalimentar con nuevos datos que reacomodan la importancia y el orden de los resultados después de cada iteración.

El aprendizaje de máquina es similar a la minería de datos en el sentido de que ambos son procesos para explorar grandes conjuntos de ellos con el fin de descubrir patrones y correlaciones; la diferencia principal estriba en que el aprendizaje de máquina llega hasta la predicción de patrones y no se queda solo en el descubrimiento de los mismos. La experiencia en este tipo de sistemas ha comprobado que entre más específica y puntual sea la tarea que realiza un sistema, podrá realizarla de manera más precisa; los sistemas que pretenden tener campos de acción muy amplios tienden a perder precisión. Si bien el término "aprendizaje de máquina" sugiere que ésta se está enseñando a sí misma, es necesario resaltar que en realidad las técnicas de aprendizaje más comunes en ellas son supervisadas por personas, y requieren enormes cantidades de trabajo y conocimiento humano agregados, así como el cuidadoso diseño y retroalimentación de los datos de entrenamiento.

El aprendizaje de máquina es usado hoy en día no 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, análisis de usuarios y tendencias, etcétera.

Muchos de ellos ya han empezado a construir aplicaciones prácticas de aprendizaje de máquinas en diversas vertientes: en primer lugar, se distingue el análisis y la síntesis de documentos. Consiste en programas que pueden "leer" cierto documento y extraer información a partir de él. Como ya se mencionó, los sistemas de este tipo se construyen para campos de acción –entiéndase documentos– muy específicos; no existe todavía el "sistema interpretador" universal para todo tipo de documentos. Los programas al efecto se construyen para interpretar específicamente cierto tipo de ellos: los hay para textos, para imágenes, para video, etcétera; y dentro de estos existen todavía más especializaciones: existen ya programas que han sido construidos para leer libros, otros para revistas, los hay para tesis, para correos electrónicos, etcétera. Existen programas para interpretar fotografías; otros para mapas, para pinturas famosas, para partituras, etcétera. Su función es extraer cierto tipo de información específica de esos documentos; ésta puede consistir en elementos tan variados como un resumen del texto, o qué tipos de personas u objetos aparecen en la imagen, o ciertos metadatos como personas, eventos, lugares, editorial, etcétera; esa información extraída puede ser utilizada posteriormente para ciertos proyectos o usos en la biblioteca.

Es conveniente resaltar en este punto que estos sistemas pueden convertirse en poderosos auxiliares para la biblioteca, pero siguen sin poder sustituir totalmente a la actividad humana en los proyectos. No existe a la fecha un sistema que pueda leer libros y construir sus fichas catalográficas completas a partir de ellos de forma sistemática y confiable, pero sí extraer suficiente información coherente de los mismos para proporcionar elementos valiosos para las personas, como los catalogadores, o para sistemas, como los "descubridores de biblioteca" (*library discoverers)*. Este uso es uno de los puntos de posible aplicación de estos sistemas que merece mayor reflexión en las bibliotecas. De hecho, uno de los cuestionamientos más radicales en ellas hoy en día es si deben seguirse construyendo los catálogos al estilo "tradicional" o debe ya cambiarse hacia nuevas estructuras de registro y recuperación documental. Una interesante reflexión de ello puede verse en Bourg (2017).

Otro ejemplo práctico de programas de "aprendizaje de máquina" muy conocido en las bibliotecas son los denominados OCR (*Optical Character Recognition* o Reconocimiento Óptico de Caracteres), los cuales, como es sabido, interpretan texto que ha sido escaneado en forma de imagen para convertirlo en formatos de texto interpretables por computadora, tales como txt, doc, rtf, odt, pdf, etcétera. Este tipo de programas pertenecen al campo de la IA dado que su función es leer e interpretar letras a partir de una forma gráfica, tal como los seres humanos lo hacen, y caen en el subcampo del "aprendizaje de máquina", ya que estos programas pueden "aprender" retroalimentándose de lo que las personas les van indicando, como errores de interpretación, fuentes tipográficas antiguas y descontinuadas, manchas y sus eventuales correcciones.

Las bibliotecas han usado estos dispositivos desde hace mucho más tiempo de lo que se cree. A principios de los años treinta –mucho antes de las computadoras– se inventó el reconocimiento de caracteres asociado a microfilmes para búsqueda rápida. Para fines de la década de los cuarenta, este tipo de sistemas era ampliamente utilizado en bibliotecas y archivos. Por ejemplo, Alicia Perales (1962, 21-22) consigna el Microfilm Rapid Selector, el cual almacenaba consecutivamente fichas con información en un microfilme. Adjunto a cada una de ellas se guardaba simultáneamente un patrón de marcas blancas codificadas en el filme a semejanza de perforaciones en tarjetas, el cual se buscaba después ópticamente por una máquina por medio de células fotoeléctricas para encontrar la combinación de marcas codificadas que contenían la información deseada en las fichas, a una tasa de hasta 36 mil fichas por hora.

Un dispositivo semejante también muy usado en bibliotecas durante las décadas de los cincuenta y posteriores fue el Kodak Minicard. La diferencia es que éste almacenaba cada registro individualmente. A mediados de los setenta, se introdujeron comercialmente los primeros programas OCR para interpretar textos con computadoras. En un principio eran muy burdos e ineficientes; no obstante, las bibliotecas los empezaron a adquirir y a utilizar para la conversión de textos debido a su utilidad. Con los años, estos programas han logrado niveles de eficiencia muy avanzados,

ya que sus características de aprendizaje y corrección han evolucionado enormemente hasta hacerlos muy eficaces y por esa razón prácticamente todos los proyectos de digitalización en bibliotecas y archivos incluyen el uso de un sistema OCR.[22] Más aún, dado que otras variantes de estos sistemas pueden también interpretar y convertir texto escrito al habla, así como texto escrito al alfabeto Braille, son muy utilizados en bibliotecas con departamentos tiflológicos para el servicio a personas que requieran de ellos. Pocos bibliotecarios han reflexionado en el hecho de que los sistemas OCR y tiflológicos que poseen en sus bibliotecas son añejas aplicaciones prácticas de IA, y en especial del aprendizaje de máquina.

Datos masivos y sistemas expertos en bibliotecas

> *Aunque los sistemas expertos pueden crear nuevas funciones para los bibliotecarios y liberarlos para otras tareas de alto nivel, esos sistemas invadirán de alguna manera sus dominios profesionales. Por eso, los bibliotecarios deben familiarizarse con la investigación actual y las aplicaciones acerca de los sistemas expertos que pueden afectar a las bibliotecas.*
>
> S. E. B.,
> "The Cutting Edge", *American Libraries* 1983

Otra de las subdivisiones de la IA utilizada ampliamente en bibliotecas son los denominados "sistemas expertos". Estos han despertado el interés de los bibliotecarios desde los ochenta, y desde entonces pueden encontrarse numerosos textos acerca de ellos que tratan, desde entonces, indización basada en el conocimiento,

22 Hoy en día, los sistemas de OCR más utilizados en proyectos de bibliotecas y archivos son: OneNote, Google Drive, SimpleOCR, FreeOCR, PhotoScan, OmniPage, Abby FineReader y Capture2Text. Por supuesto, existen más.

procesamiento de lenguaje natural, catalogación, consulta recuperación de información, etcétera.

Los sistemas expertos son programas informáticos que utilizan principios y métodos de la Inteligencia Artificial para resolver problemas dentro de un campo especializado que usualmente requeriría de la experiencia de personal experto. Incorporan los conocimientos técnicos acumulados por las personas expertas en un tema y se diseñan para funcionar lo más parecido a ellas. Básicamente, contienen una *base de conocimientos* de hechos y relaciones representados en forma de datos, y tienen la capacidad de hacer inferencias basadas en ellos. Los creadores de estos sistemas utilizan diversas técnicas para la adquisición de esa base de conocimientos, como el análisis de protocolos y procedimientos escritos, la descripción verbal de tareas realizadas por una persona, los cuestionarios, encuestas y entrevistas; el descubrimiento y documentación del conocimiento tácito dentro de la organización, así como la observación de procesos y su simulación.[23] Éste es un filón de gran valor en las bibliotecas, pues mucho del conocimiento de los bibliotecarios acerca de la gestión y explotación de información cae dentro de este rubro. El conocimiento tácito o interno de los bibliotecarios es su conocimiento acumulado, el conocimiento generado por su experiencia, aquel inherente al personal de biblioteca que ha sido interiorizado mediante diferentes procesos. Olivares (2020) hizo una reseña muy completa acerca de este tema.

Los sistemas de aprendizaje de máquina y sistemas expertos pueden usarse en muchas otras aplicaciones prácticas en las bibliotecas:

- Algunas bibliotecas extraen y guardan masivamente datos acerca de las búsquedas de sus usuarios para aprender más acerca de la lógica y formas que ellos utilizan para

23 El primer sistema experto fue desarrollado en 1965 en la Universidad de Stanford por Edward Feigenbaum y Joshua Lederberg. Se denominó "Dendral" y fue construido para el análisis de compuestos químicos.

aproximarse a la información y de esta forma mejorar los catálogos internos, los "descubridores" de información, etcétera.

- Muchas bibliotecas estudian específicamente el "lenguaje natural" que los usuarios –como todas las personas– utilizan para comunicarse, para tratar de enseñar a los computadores a entender ese lenguaje. Bajo estos principios, la máquina puede entender los conceptos clave del lenguaje dentro de una pregunta y su posible solución a través del procesamiento por IA del lenguaje natural. El objetivo de estas aplicaciones es diseñar y crear programas que analicen el lenguaje llano que una persona usa para aplicarlos a la extracción de información de textos, la recuperación de información en bases de datos y catálogos, la traducción automática, el reconocimiento y síntesis del habla, etcétera. Todas estas aplicaciones requieren de grandes cantidades de datos.
- De igual forma, muchas bibliotecas guardan información de búsquedas previas de los usuarios con objeto de "personalizar" la página de cada uno de ellos, "recordando" lo que han buscado con anterioridad estableciendo patrones, tal como lo hacen los sitios de comercio electrónico. Al guardar este tipo de información de cada usuario, el sistema puede posteriormente 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".
- Las páginas "personalizadas" de biblioteca se construyen por lo general de tal forma que permiten a cada usuario guardar la forma y aspecto de su página, los formatos para despliegue, sus búsquedas anteriores, etcétera, por lo que la presentación y comportamiento de la página de cada usuario pueda ser distinta, a su gusto y conveniencia.
- Algunas bibliotecas extraen datos de las redes sociales de sus usuarios conectadas a los servicios de la biblioteca, para recibir sugerencias de adquisición de obras, detectar

> "temas de tendencia" (*trend topics*), contar "me gusta" y otros eventos similares acerca de sus servicios o informaciones, verificar eficacia y dar seguimiento de sus servicios, medir "usabilidad" de nuevos servicios y opciones, detectar fallas o problemas, diseñar nuevos tutoriales, y muchos otros usos más.

Puede verse un interesante compendio de aplicaciones específicas de aprendizaje de máquina para bibliotecas en la obra *Artificial Intelligence and Machine Learning in Libraries* (2018).

Norman Jacknis (2017) resume la interacción entre bibliotecas e Inteligencia Artificial de una manera espléndida:

> […] La cuestión no es elegir 'entre la IA o las bibliotecas', sino que ambas se refuerzan mutuamente en aras de proporcionar mejores servicios a los usuarios. En lugar de ser puramente 'Inteligencia Artificial', estos nuevos servicios serían lo que empieza a ser una nueva palabra de moda: el AI (Aumento de la Inteligencia) para los seres humanos.

Las bibliotecas como grandes repositorios de datos

Es fácil mentir con datos. Es muy difícil decir la verdad sin datos.
ANDREJS DUNKELS

Durante la última década, las comunidades académicas y de información cayeron en cuenta de que los datos recopilados a lo largo de investigaciones científicas, periodísticas, sociales, etcétera, tenían un valor agregado después de concluidos sus proyectos al poder ser reutilizados posteriormente por otras personas, ya que sin duda un cierto conjunto de datos provenientes de una investigación es susceptible de ser analizado desde nuevos enfoques y perspectivas por grupos diferentes, y eventualmente nuevos resultados pueden obtenerse de esos datos. A partir de esta consideración, ellos no son ya únicamente la materia prima que produjo información, sino un objeto de información en sí mismos con un valor inherente propio, y por esa razón requieren de un tratamiento específico.

Además de ello, la tendencia mundial de la divulgación de los resultados de investigación comenzó a experimentar un cambio: ya no era obligación de los investigadores publicar únicamente en revistas "de prestigio". Cada vez más, los organismos de financiación académica gubernamental comenzaron a requerir que tanto los resultados como los datos de las investigaciones que se

hiciesen con fondos oficiales se hiciesen públicos.[24] Pero hacer públicos conjuntos de datos requiere de método y normalización. Hasta ese momento, prácticamente cada investigador o grupo diseñaba sus formatos, formas, depósitos, etcétera, para sus datos en cada investigación; cualquier método era bueno si funcionaba para el proyecto. Súbitamente, las comunidades de investigadores se vieron ante la necesidad de comenzar a manejar sus datos de una manera sistematizada y estandarizada con miras a guardarlos y accederlos después.

Los investigadores se vieron 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. Las instituciones de investigación –en especial aquellas en universidades– tuvieron la necesidad de comenzar a crear repositorios de datos de sus proyectos: esto significó un nuevo impulso a la ciencia y la gestión de los datos. Muchas instituciones acudieron a sus bibliotecas para asesorarse al respecto y para que ellas comenzaran a alojar esos conjuntos de datos. Así llegaron los repositorios de datos a las bibliotecas. A partir de ese momento, en palabras de Rafael Ball (2019, 2): "[…] el trabajo de la biblioteca ya no se centra solo en libros, revistas, y catálogos, sino también en todo tipo de datos –estructurados y no– así como sus formas: textos, metadatos, imágenes, acervos de audio y video, datos de investigación y software".

Diversas organizaciones bibliotecarias comenzaron a esbozar estos nuevos retos, tales como la "Asociación de Bibliotecas Universitarias y de Investigación" de Estados Unidos (*Association of College and Research Libraries*, la cual es una subdivisión de la ALA) (Tenopir *et al.* 2012, 2015), y la Liga de Bibliotecas Europeas

24 Existe ya una serie de iniciativas denominada "Open Data" (Datos Abiertos) la cual, de manera semejante a las revistas en acceso abierto, impulsa la creación y difusión de repositorios de datos abiertos. Además deben regirse por los principios denominados FAIR acerca de los datos: Findable, Accessible, Interoperable and Reusable; esto es, los datos deben ser localizables, accesibles, interoperables y reusables.

de Investigación" (LIBER, *Ligue des Bibliothèques Europeénnes de Recherche*) (Tenopir *et al.* 2016). La IFLA también realizó estudios detallados de los temas relacionados con datos en bibliotecas; en el último fascículo del año 2016 y primero del 2017 de su revista, esta organización compiló cerca de una veintena de textos y reflexiones acerca del tema, y lo dividió en cuatro grandes rubros: las necesidades de los investigadores, las habilidades requeridas de los bibliotecarios, los posibles servicios a ofrecer y la alfabetización en datos (IFLA s.f.). A partir de estos estudios preliminares, la IFLA creó una iniciativa al respecto denominada "Proyecto del Curador de Datos" (*Data Curator Project*) (IFLA s.f.-2). Su objetivo principal era determinar las funciones y responsabilidades de los profesionales de las bibliotecas que ya estaban trabajando en estas tareas 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, 251) hicieron una lista muy representativa y concisa de las principales actividades requeridas a los bibliotecarios a este respecto: 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, y 5) alojar y mantener repositorios en sus acervos.

Todas estas nuevas necesidades, conceptos y soluciones dieron origen a una nueva especialidad en el mundo de la información denominada "Gestión de Datos de Investigación" (o *Research Data Management*, RDM). Whyte y Tedds (2011) 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". Básicamente, la RDM trata todos los aspectos relacionados con la gestión, el almacenamiento y la distribución de datos provenientes de investigaciones: ciclo de vida de los datos, colecciones de ellos; captación, limpieza, consistencia y normalización de datos, sus formatos, metadatos, los repositorios y servicios para su consulta,

anonimización y seguridad de datos, su preservación, las habilidades y los roles necesarios para quien los opere, alfabetización en datos para los investigadores y hasta citación de los mismos.

Pinfield, Cox y Smith (2014) señalan que hay 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. La "gestión de datos", y en especial los de investigación, es una actividad multidisciplinaria, pero es claro que la Bibliotecología y los bibliotecarios deben estar entre los profesionales que los manejan. Por supuesto requiere de nuevos conocimientos, capacitación y entrenamiento, pero ciertamente los bibliotecarios tienen las bases profesionales adecuadas para esta tarea.

El interés en el tema de datos y bibliotecas ha ido creciendo enormemente en años recientes. La Association of Research Libraries (ARL) y la National Science Foundation (NSF) crearon hace pocos años una unión especial para desarrollar conjuntamente proyectos en lo que se conoce ahora como e-Ciencia; entre ellos los estudios y desarrollos en RDM. Algunos autores ya han tratado el tema de las bibliotecas en la RDM, como Cox y colegas (2017), Alvaro y colegas (2011), Matusiak (2019), y Lewis (2010). El Instituto de Investigaciones Bibliotecológicas y de la Información (IIBI) de la UNAM organizó en noviembre de 2018 un evento exclusivamente dedicado al manejo de datos y su relación con las bibliotecas: el Segundo Congreso de Estudios de la Información: Manejo de datos (*El manejo de datos…* 2020).

Muchas universidades e institutos de investigación están creando ya repositorios de datos al respecto; muchas bibliotecas y sistemas bibliotecarios están ya trabajando en este sentido. Un ejemplo muy representativo de ello es el repositorio de datos de la Red de la Biblioteca Nacional de Medicina de la Unión Americana Network of the National Library of Medicine (NNLM). Este repositorio fue creado por esta biblioteca para que los investigadores de instituciones asociadas a ella que así lo deseen puedan alojar ahí datos resultados de sus proyectos, obviamente en el área de salud (NNLM, s.f.). Otro ejemplo interesante son las guías creadas por las

bibliotecas del Instituto Smithsoniano para la creación y depósito de datos en repositorios, las cuales abarcan una amplia variedad de características debido a la enorme diversidad de intereses y disciplinas que abarca ese organismo (Smithsonian Libaries s.f.). En el campo de las ciencias sociales, puede citarse como ejemplo el Consorcio Inter-Universitario para Investigación en Ciencias Políticas y Sociales (Inter-University Consortium for Political and Social Sciences o ICPSR) de las bibliotecas del Instituto Tecnológico de Massachusetts (MIT), considerado el mayor repositorio a nivel mundial en este campo del conocimiento. Además de los repositorios de datos propios de instituciones académicas, que son ya muy numerosos, comienzan a proliferar algunos otros para alojamiento de datos en general, tales como Zenodo, Dryad o Dataverse.

En América Latina, la Comisión Económica para América Latina y el Caribe (CEPAL) de la ONU está asociada a un proyecto denominado Leaders Activating Research Networks (LEARN) como parte del programa de investigación e innovación Horizon de la Unión Europea. Este proyecto tiene como fin el impulso y desarrollo de proyectos de gestión de datos de investigación (GDI). Derivado de esta iniciativa, el sistema brasileño Scielo de revistas académicas de esta región ha instalado recientemente una versión piloto de un repositorio de datos precisamente para los investigadores y datos provenientes de esta zona geográfica. El repositorio está construido sobre Dataverse (Scielo Data s.f.). Algunos países de la región han comenzado a legislar al respecto y/o a construir repositorios de datos en universidades e instituciones académicas afines: Argentina, Brasil, Chile, Colombia, México y Perú (Andaur 2016). Existen ya además algunos registros o catálogos a nivel mundial que informan acerca de los repositorios de datos científicos, como por ejemplo "re3data", el cual ofrece información acerca de más de dos mil repositorios de este tipo (Re3data s.f.).

Como puede verse, hoy en día la "gestión de datos de investigación" es un nuevo campo de acción que ofrece nuevas y amplias oportunidades para los profesionales de la información que requiere de nuevas habilidades y conocimientos específicos; entre aquellos que son más susceptibles de ser formados en esta nueva

aproximación de la información, están sin duda los bibliotecarios debido a su bagaje y experiencia profesional características.

Pero la gestión de datos para el público no se limita a aquellos provenientes de investigación; como ya fue mencionado, en años recientes se ha ido gestando un conjunto de iniciativas bajo el denominador común de "Datos Abiertos" (*Open Data)*, las cuales tienen como fin impulsar la creación, difusión y uso de repositorios de datos abiertos de todo tipo. Estas iniciativas alrededor de los datos son la continuación de otros movimientos previos en favor de la apertura: *software* libre, Gobierno Abierto, revistas académicas abiertas, etcétera. Ello debido a que, más allá de los datos de investigación académica, los datos abiertos ocupan cada vez más un lugar preponderante en el mundo moderno: permiten una comprensión más completa de los problemas globales y las cuestiones universales, como enfermedades, educación, inseguridad, empleo o hambrunas. Son un factor fundamental de los principios de Gobierno Abierto con transparencia y rendición de cuentas; empoderan a los ciudadanos y por lo tanto fortalecen la democracia. Pueden agilizar los procesos y las estructuras sociales que gobiernos y sociedades han construido. Pueden apoyar de manera sobresaliente movimientos en favor de la igualdad racial, de género, etcétera. En resumen, pueden ayudar a transformar la forma en que entendemos el mundo moderno y nos relacionamos con él (ODC s.f.).

Ya existen algunos proyectos representativos al respecto en operación; por ejemplo, los Datos Abiertos del Banco Mundial (World Bank Open Data*)*, cuyo repositorio contiene más de 3,000 conjuntos de datos globales sobre desarrollo, economía, etcétera, en forma abierta. Igualmente, el Repositorio de Datos Abiertos de la Organización Mundial de la Salud, el cual engloba la información estadística sobre este tema proveniente de sus casi 200 miembros. Se encuentra también el Portal de Datos Abiertos de la Unión Europea, con 12 mil conjuntos de datos provenientes de gobiernos, agencias, instituciones, etcétera, de esa zona geográfica. Puede verse también el proyecto DBPedia de Wikipedia, el cual permite buscar y explorar semánticamente las relaciones y

propiedades de entre 4.6 millones de elementos de esa enciclopedia, tales como personas, lugares, eventos, etcétera. Existe también el proyecto Registry of Open Data on AWS *Resources* (RODA o Registro de Datos Abiertos en Recursos de la plataforma AWS) de Amazon, el cual permite buscar en un solo sitio datos abiertos que se hayan capturado en esa plataforma. Similar al anterior se encuentra el Explorador de Datos Abiertos de Google (Google Public Data Explorer), el cual permite buscar en múltiples bancos de datos abiertos a nivel mundial. A estos habría que agregarles numerosos bancos de datos abiertos específicos de agencias como la Administración Nacional Oceanográfica y Atmosférica (NOAA) y el Centro Nacional de Investigación Atmosférica (NCAR), los cuales colectan y distribuyen datos climáticos, meteorológicos, etcétera, de toda Norteamérica; y de forma semejante los servicios sismológicos, vulcanológicos, censales, etcétera, a nivel regional o de numerosos países.

Pero todavía hay más: ya se han mencionado las iniciativas alrededor del concepto de Datos enlazados en bibliotecas. Cuando estos datos se instalan como abiertos; esto es, pueden utilizarse y distribuirse libremente, se denominan Datos Abiertos Enlazados (Linked Open Data o LOD). Teóricamente, se conforma como una nube virtual de datos en la que cualquier persona puede acceder a cualquier dato autorizado así como agregar nuevos, lo cual proporciona un entorno abierto, estructurado e interoperable que favorece que los datos puedan ser creados, interconectados y consumidos a escala global.

Nótese que son dos conceptos relacionados pero diferentes: los datos abiertos se ponen a disposición de todos sin necesidad de estar entrelazados con otros; los datos pueden vincularse sin que tengan que estar disponibles libremente para su uso y distribución. En suma, los datos pueden ser abiertos pero no enlazados, y pueden estar enlazados pero no ser abiertos. Cuando se conjuga a la vez que los datos estén enlazados y sean abiertos, entonces son datos abiertos enlazados. Al igual que con otros tipos de proyectos de datos, ya existen propuestas alrededor de datos abiertos enlazados específicamente en bibliotecas (véase el compendio de

documentos al respecto resultado de la Reunión Satélite de IFLA en Francia en 2014, http://ifla2014-satdata.bnf.fr/).

Los ejemplos mencionados de grandes proyectos alrededor de datos abiertos ya están operando; diariamente se suman nuevos a la lista. El punto central es que el diseño, la gestión y la explotación de este tipo de datos abarca un campo infinitamente mayor que la "gestión de datos de investigación". Aun si esos proyectos no se gestan o insertan directamente en una biblioteca, requieren indefectiblemente de personal con experiencia y conocimientos en el manejo de datos. Buena parte de esos desarrollos ya involucran a personal proveniente de bibliotecas, pero sin duda podrían ser más. El personal profesional bibliotecario tiene sin duda en esos proyectos de datos abiertos –más allá de los de investigación académica– grandes oportunidades de desarrollo profesional.

Todos los anteriores usos presentados acerca de datos masivos en bibliotecas no son hipotéticos; ya existen y son utilizados en alguna biblioteca del mundo. Son las nuevas teorías y descubrimientos que se van convirtiendo en aplicaciones tecnológicas en este entorno. El contexto de las bibliotecas digitales sigue cambiando. Pierre Piganiol, en una fecha tan temprana como 1971, resumió esto espléndidamente:

> […] la información no debe construir una estructura muerta: el cuerpo de conocimiento está en continua evolución, y para poder predecir e influir en el futuro es vital que la información contenga al menos las semillas del progreso y los descubrimientos del mañana. Lo que distingue a la ciencia de la información moderna de la documentación tradicional es precisamente la introducción de este elemento heurístico (Piganiol 1971 13).

La cara negativa de los datos masivos en bibliotecas

Porque si somos observados en todos los asuntos, estamos constantemente bajo la amenaza de corrección, juicio, crítica, incluso plagio de nuestra propia singularidad. Nos convertimos en niños, encadenados bajo ojos vigilantes, constantemente temerosos de que –ya sea ahora o en el futuro incierto– los patrones que dejemos atrás serán traídos de vuelta para implicarnos… Perdemos nuestra individualidad, porque todo lo que hacemos es observable y registrable.

BRUCE SCHNEIER.
"El eterno valor de la privacidad", 2006.

Como cualquier otro desarrollo tecnológico –aunado a sus múltiples usos y beneficios–, los datos masivos también tienen sus riesgos, problemas y desventajas, los cuales son bastante significativos, y por este motivo deben ser conocidos y estudiados por los bibliotecarios para poder evitarlos o, al menos, reducirlos.

Los datos masivos son difíciles de administrar, en parte por su inmenso volumen inherente y en parte porque hay una falta generalizada de conocimiento en cómo manejarlos adecuadamente; a nivel mundial hay muy pocos expertos en datos. Esta falta de conocimiento y de personal calificado por lo general implica mal planteamiento de objetivos y técnicas, duplicación de datos, inconsistencia o sesgo de los mismos, mala selección de herramientas de análisis, interpretaciones erróneas, etcétera, con las

consecuencias negativas subsecuentes. Es muy fácil perderse en un mar de datos; además, los proyectos de datos masivos requieren de un cierto presupuesto y presentan grandes retos técnicos. Obviamente todo esto es un inconveniente; pero sin duda el aspecto más negativo del uso de los datos masivos es su posible abuso en la privacidad y confidencialidad de datos personales.

Todas las anteriores aplicaciones enunciadas de los datos masivos –que no son ni buenas ni malas en sí mismas– pueden ser y son ya utilizadas para enormes beneficios de las personas; desgraciadamente, conllevan un enorme problema: dado que con frecuencia implican colecta de datos personales, introducen también grandes riesgos a su privacidad. Al compilar y almacenar grandes cantidades de datos –si se incluyen datos personales entre ellos–, se introduce el riesgo de que estos sean usados para propósitos distintos a la estadística y al procesamiento masivo, y que empiecen a ser utilizados para fines poco éticos o hasta ilegales.

Aun cuando el compilador de los datos no los utilice mal, existe el riesgo de fugas de información de sus servidores de datos, deliberadas o por error, o ataques a los mismos para extraer información por parte de terceros con intenciones aviesas. Esto no es un simple problema de protección de datos: implica desde principios y decisiones éticas, hasta manejo de grandes intereses comerciales, legislaciones y normas al respecto, responsabilidades técnicas y administrativas, gobernanza de datos, rendición de cuentas y seguridad informática. Hay componentes diferentes en ello: la protección de datos es esencialmente una cuestión técnica que implica mayormente asegurar los datos contra el acceso no autorizado: quién y cómo los cuida. La privacidad de los datos va todavía más allá: es una cuestión ética y legal que implica aspectos todavía más profundos: quién puede tener datos personales, durante cuánto tiempo, quién define a aquellos que los pueden acceder, quién puede acceder a ellos autorizadamente y en qué circunstancias, quién los puede modificar, a quién y cómo pueden ser transferidos, etcétera. La protección de los datos es un requisito necesario más no suficiente para lograr un fin mayor: la privacidad de los datos.

La Regulación de Protección de Datos Personales de la Unión Europea (EU's General Data Protection Regulation o GDPR) establece que se entiende por "datos personales":

> […] toda información relativa a una persona física identificada o identificable; esto es, aquella persona que puede ser identificada, directa o indirectamente y en particular, mediante referencia a un identificador como un nombre, un número de identificación, datos de localización, un identificador en línea o uno o varios factores específicos de la identidad física, fisiológica, genética, mental, económica, cultural o social de esa persona física (GDPR s.f., s.p.).

La ley mexicana al respecto define los datos personales como "cualquier información concerniente a una persona física identificada o identificable. Se considera que una persona es identificable cuando su identidad pueda determinarse directa o indirectamente a través de cualquier información" (México. Ley General… 2017). La GDPR –vigente desde el 2018– tiene como antecedente a la "Directiva de e-Privacidad" del 2002, también de la Unión Europea, y es considerada actualmente como la regulación más avanzada a nivel mundial en beneficio de la privacidad de los usuarios (Eur-Lex s.f.).

Sin duda alguna, los datos son hoy en día un inmenso negocio a nivel mundial, y entre ellos, los datos personales lo son todavía más. Grandes empresas transnacionales como Google, Facebook, Twitter, etcétera, tienen ganancias multimillonarias que provienen en gran parte de la venta de datos; este hecho en sí no va contra la ética o la ley. Existen innumerables aplicaciones legítimas y éticas del uso y/o venta de datos que crean negocios, comercio, publicidad, aplicaciones gubernamentales, puestos de trabajo, etcétera, todo dentro de principios legal y moralmente aceptables.

El problema está en que la línea entre lo legal y lo ilegal, lo ético y lo que no lo es, en realidad es difusa y sus límites no son claros. No es el caso aquí discutir todos los malos usos que se le pueden dar a los datos, pero quedan como claros ejemplos el escándalo de la empresa Cambridge Analytica, que usó datos personales extraídos de Facebook para crear un sesgo en las elecciones de la Unión Americana en el 2016; Facebook enfrentó pérdidas

multimillonarias al respecto. Están también las demandas, sanciones y multas que han recibido y siguen recibiendo las empresas Google, Apple y Amazon en Europa por manejo dudoso y abusivo de los datos que colectan. Como estos ejemplos, hay miles al respecto; todos hemos recibido una infinidad de llamadas telefónicas y correos en las cuales nos ofrecen bienes y servicios que nunca solicitamos, y hasta herencias africanas. El punto es que existen muchas entidades interesadas en apropiarse de datos personales y existen muchas otras dispuestas a entregarlos, no siempre dentro de lo legal y ético. La tentación al respecto ha sido y sigue siendo enorme.

Derivado de ello, hoy en día más de cien países poseen algún nivel de legislación y regulaciones acerca de privacidad y protección de datos (CNIL 2019). Algunos países han llevado las regulaciones a niveles muy altos en beneficio de sus ciudadanos, como es el caso de aquellos pertenecientes a la Unión Europea, y hay otros que, en la práctica, aunque han expedido leyes para protección de datos personales son conocidos por ejercer una fuerte vigilancia y censura gubernamental a sus ciudadanos, como Rusia o China. Hay países que no ejercen censura sobre sus ciudadanos, como los Estados Unidos, pero se cuestiona mucho hasta dónde ejercen vigilancia sobre sus ciudadanos; esto no es historia del pasado. En reconocimiento de lo anterior, la Asamblea General de las Naciones Unidas adoptó en 2013 y 2014 varias resoluciones sobre el "Derecho a la privacidad en la era digital", y conminó a todos los países a que "respeten y protejan el derecho a la privacidad, incluso en el contexto de las comunicaciones digitales [y a] adoptar medidas para poner fin a las violaciones de esos derechos y creen las condiciones necesarias para impedirlas, como cerciorarse de que la legislación nacional pertinente se ajuste a sus obligaciones en virtud del derecho internacional de los derechos humanos" (ONU 2013).

Austin (2016) estableció que a nivel mundial había entonces diez grandes temas de atención y debate acerca de la privacidad y protección de datos personales: 1) Las regulaciones de ubicación física y jurisdicción de servidores de datos; 2) el Internet de las Cosas y cómputo ubicuo; 3) las regulaciones oficiales de privacidad; 4) las regulaciones que flexibilizan las obligaciones de

cumplimiento en temas de y privacidad y protección de datos; 5) la vigilancia gubernamental; 6) el desarrollo de nuevos estándares de ciberseguridad; 7) los datos masivos; 8) el nuevo marco mundial de referencia acerca de transferencias de datos; 9) las recientes leyes y regulaciones acerca de seguridad de datos, y 10) las nuevas regulaciones de la Unión Europea sobre protección de datos personales.

Dentro del campo de nuestro interés, las bibliotecas y los archivos, las actividades de compilación y uso de datos –en especial los masivos– también implican con frecuencia la colecta de datos personales. Esto conlleva responsabilidades que el bibliotecario debe conocer, así como habilidades que debe adquirir para su correcto uso. Ello es inevitable y no fue originado en lo absoluto por los datos masivos: ha existido desde hace mucho tiempo y simplemente ahora se ha acentuado más con la tecnología. Durante largo tiempo, las bibliotecas han defendido el derecho a la privacidad de sus usuarios: la American Library Association (ALA) adoptó desde tan atrás como 1939 un conjunto de principios conocido como la "Declaración de Derechos en la Biblioteca". Entre varios derechos, definió ahí que lo que cualquier persona decida leer no compete a nadie más, y no hay razón válida para que gobiernos, organizaciones o personas interfieran o se enteren de ello. En su versión actual, establece en su inciso VII:

> […] Todas las personas, independientemente de su origen, edad, antecedentes o puntos de vista, tienen derecho a la privacidad y la confidencialidad en el uso de su biblioteca. Las bibliotecas deben abogar por la privacidad de las personas, educarlas y protegerlas, salvaguardando todos los datos de uso de la biblioteca, incluida la información de identificación personal (ALA 1939).

En su "Interpretación de la Carta de los derechos a la privacidad", la ALA establece claramente:

> […] la privacidad es esencial para el ejercicio de las libertades de expresión, de pensamiento y de asociación. La falta de privacidad y confidencialidad disminuye las opciones de los usuarios,

> suprimiendo así el acceso a las ideas. La posibilidad de vigilancia, ya sea directa o a través del acceso a los registros del discurso, la investigación y la búsqueda, socava una sociedad democrática (ALA 2002).

La IFLA también se pronunció al respecto desde hace muchos años. Sus principios partieron de acuerdo con los Artículos 12 y 19 de la Declaración Universal de los Derechos Humanos, entre los cuales la libertad de acceso a la información y la libertad de expresión han sido principios fundamentales para la profesión en las que –muy especialmente– se considera a la privacidad como parte indispensable de la salvaguarda de esos derechos.

El Artículo 12 de la Declaración Universal incluye a la privacidad como un derecho humano, y establece que "[…] nadie será objeto de injerencias arbitrarias en su privacidad, su familia, su domicilio o su correspondencia". La privacidad, por tanto, es fundamental para el acceso y uso de información sin temor a consecuencias. El Código de Ética de la IFLA recoge estos principios y explícitamente establece que:

> […] Los bibliotecarios y otros trabajadores de la información respetan la privacidad personal y la protección de datos personales que por necesidad sean compartidos entre los individuos y las instituciones […] La relación entre la biblioteca y el usuario se basa en la confidencialidad; los bibliotecarios y otros trabajadores de la información tomarán las medidas apropiadas para garantizar que los datos de los usuarios no sean compartidos más allá de su proceso original (IFLA 2012).

Esta organización agregó además en su "Manifiesto de la IFLA sobre Internet": "[…] Los servicios bibliotecarios y de información […] tienen la responsabilidad de […] esforzarse por garantizar la privacidad de sus usuarios y que los recursos y servicios que utilicen sean confidenciales" (IFLA 2015). Igualmente, la "Declaración de IFLA sobre el Acceso a la información de Identificación Personal en los Registros Históricos" establece:

> […] Los bibliotecarios deben reconocer la obligación de vigilar la legislación de sus gobiernos en lo que se refiere a la confidencialidad de los registros de datos. Especialmente, los bibliotecarios deben apoyar la necesidad de que las leyes de privacidad protejan a los usuarios de la biblioteca de abusos tales como agencias gubernamentales que vigilan sus hábitos de lectura y de investigación (IFLA 2008).

Recientemente, la IFLA dedicó todo un número completo de su revista al tema de la privacidad (IFLA 2018).

Como puede verse, la IFLA es otra de las instituciones que también ha establecido desde bastante tiempo atrás principios fundamentales del respeto a la privacidad, la protección de datos personales y la confidencialidad en la relación entre el usuario y el servicio bibliotecario o de información. A partir de ellos, muchos otros países y organismos a nivel mundial han recogido y hecho suyos estos principios; por ejemplo, en México, además de la ley, el Código de Ética Profesional del Colegio Nacional de Bibliotecarios establece*:* "[los profesionales de la bibliotecología] guardarán total reserva de los hechos respecto a la información solicitada o recibida, a los datos personales del usuario, así como a materiales consultados o prestados, a menos que lo autoricen los interesados". Muchos otros códigos deontológicos recogen estos principios, además de las leyes; esto habla de su obligación tanto ética como legal.

Los principios mencionados hablan de *privacidad* y de *confidencialidad* de los datos personales. Es necesario aclarar la diferencia dentro de las bibliotecas: surgieron desde hace muchas décadas con la *privacidad*, la cual significa en una biblioteca el derecho de todo usuario a leer y consultar lo que desee sin que los temas de su interés sean examinados o escudriñados por terceros. La *confidencialidad* proviene del hecho de que una biblioteca entre en posesión de datos personales que hagan identificable al usuario y por tanto debe tomar las medidas para evitar su acceso no autorizado. Es decir, la *confidencialidad* es un proceso que protege, entre muchas otras cosas, la *privacidad*. Esta última es un derecho de todo usuario; la confidencialidad es una obligación ética y legal que tiene la biblioteca de proteger ese derecho.

Desde que estos principios fueron establecidos hace casi un siglo, innumerables bibliotecas en el mundo construyeron dentro de sus servicios bibliotecarios y de información políticas y mecanismos de privacidad, así como las medidas para mantener la confidencialidad de los usuarios. Durante el siglo pasado, antes del mundo digitalmente globalizado, esta tarea resultó relativamente sencilla, ya que los procesos en la biblioteca que recolectaban datos personales eran muy contados: la consulta de catálogos, índices y otros impresos en soportes tradicionales no dejaba huella asociada del usuario y sus intereses. Los únicos puntos de registro entre usuario y material consultado eran, por una parte, las papeletas de préstamo, las cuales fueron siempre destruidas una vez que el libro era devuelto, y por otra parte las tarjetas de préstamo, en las que era práctica común sustituir el nombre del usuario por un número de credencial que hacía imposible para terceros la asociación de nombres con lecturas. Las bibliotecas que llevaban registros de préstamo microfilmados destruían periódicamente esos registros, precisamente con miras a la confidencialidad. Todas las bibliotecas que extraían datos para estadísticas de uso de las colecciones, lo hicieron siempre de forma anónima, igualmente debido a principios de privacidad. En las primeras décadas de la automatización de los servicios de consulta y documentación ofrecidos por las bibliotecas a través de sus propios sistemas de cómputo, éstas tenían control total sobre el acceso y, por tanto, sobre la privacidad de sus usuarios. Dado que el material a consultar estaba dentro de sus computadores, las bibliotecas podían controlarlo totalmente y evitar acceso o transferencia no autorizados a los datos sensibles. Como regla general, las bibliotecas se negaron siempre a proporcionar información personal de cualquier tipo a personas u organizaciones externas.

Desafortunadamente, entre los males que trajo la globalización de datos en el mundo digital, están la invasión masiva a la privacidad y a la confidencialidad de datos personales. La libertad de información, la privacidad y la confidencialidad se han visto seriamente amenazadas en las dos últimas décadas debido al auge de la recopilación a gran escala de este tipo de datos, la vigilancia

electrónica y la intercepción de comunicaciones digitales. Esto es, especialmente sensible en los datos masivos, y de ahí su cara negativa en las bibliotecas. Por supuesto, éste no es un problema exclusivo de ellas, pero sin duda también les afecta grandemente y por eso requiere de su atención meticulosa.

Durante la reunión anual 2019 del Foro Económico Mundial en Davos, se dieron a conocer los datos preliminares de un estudio mundial sobre la percepción de la privacidad de datos por parte del público (Ipsos 2019). Como principales hallazgos se destaca que: 1) la gran mayoría del público manifiesta muy bajos niveles de confianza respecto al uso de datos personales por parte de empresas y gobiernos a nivel mundial; 2) sólo un tercio del público tiene conocimientos aceptables acerca de la privacidad de datos; 3) dos tercios del público se sentirían más cómodos si supieran claramente cómo las organizaciones usan y comparten sus datos.

Lo anterior viene a colación porque si las bibliotecas quieren estar entre las instituciones que el público percibe como "confiables" en lo relativo a privacidad y confidencialidad de datos personales, es indispensable que manejen y compartan los datos personales de forma totalmente eficiente y segura, y además que hagan ese proceso muy transparente hacia los usuarios, de tal forma que efectivamente puedan crear en ellos la percepción de confiabilidad en la seguridad de datos dentro de la institución. De otra forma, las bibliotecas pasarán indefectiblemente a formar parte de ese gran conglomerado de organizaciones percibidas como "poco confiables" por el público, ya sea debido a que no manejan bien los datos personales o a que –a pesar de hacerlo bien– no lo comunican adecuada y transparentemente al público. Gorman (2000, 36) señala que el respeto a la privacidad personal "es uno de los ocho valores fundamentales" para la confianza entre la biblioteca y sus usuarios. Además de lo anterior, no hay que olvidar que en muchos países –como es el caso de México– las leyes disponen que todas las organizaciones que guardan datos personales cumplan con una serie de disposiciones al respecto. Por lo mismo las bibliotecas son sujetos obligados.

Esto no es trivial: debido al inmenso auge de servicios y productos de información digitales con los que la biblioteca y sus usuarios interactúan en la actualidad, es necesario cuidar numerosos aspectos de los datos de los usuarios. A diferencia de las papeletas del siglo pasado, hoy en día hay numerosos puntos de eventual recolección de datos personales en la biblioteca, aún sin que ella se lo proponga; la ALA señala nada menos que una lista de 32 posibles puntos al respecto: abarcan los registros electrónicos de préstamo, las bitácoras de búsquedas en catálogos, los elementos asociados a ellas como historiales de búsqueda, cachés, *cookies* y certificados; los correos electrónicos y los servicios de diseminación selectiva, por mencionar algunos (ALA 2007). Además de estos puntos "típicos" de colecta de datos, existirán aquellos derivados de proyectos especiales de datos masivos, como los ya mencionados acerca de estilos y preferencias de búsqueda de los usuarios, personalización de páginas web, extracción de tendencias en redes sociales, etcétera.

Por si ello fuera poco, una gran parte del problema de la privacidad ha sido introducido por causa de la interacción de publicaciones y servicios de información digitales y en red provenientes de un proveedor externo a la biblioteca. Cada vez más, las obras que se consultan, los descubridores, los servicios de búsqueda y documentación, tablas de contenido, etcétera, provienen de terceros comerciales. Cuando un usuario decide adquirir o se suscribe a bienes y servicios de información directamente de manera personal, es común que el proveedor le imponga condiciones que con frecuencia atentan, entre muchas otras cosas, contra su privacidad. Si esa persona quiere aceptar estas condiciones cuando accede a un texto digital a un proveedor, es su gusto y decisión. El problema se vuelve grave y atañe a las bibliotecas cuando los proveedores pretenden que estas condiciones sean extendidas a ellas, para quienes por supuesto son del todo inaceptables. Entre todos los inconvenientes de esos productos y modelos comerciales, el aspecto de la privacidad y el respeto a los datos personales se ha ido convirtiendo en uno de los temas más candentes, ya que las bibliotecas y sus usuarios enfrentan serios retos al respecto:

- Fuera del ámbito de la institución, los proveedores de contenidos y servicios de información comerciales utilizados por las bibliotecas pueden –y de hecho, lo hacen– recopilar datos sobre las búsquedas, actividades y transacciones de los usuarios, o ponen como condición para la prestación de sus servicios o contenidos que las bibliotecas recopilen y les transfieran datos.
- Los servicios en la nube que alojan sistemas bibliotecarios pueden recopilar, almacenar y transferir datos de los usuarios al margen de la institución bibliotecaria o de información. Con frecuencia la biblioteca desconoce cómo y dónde se procesan y almacenan esos datos en la nube. Recuérdese que una enorme proporción de estos servicios se aloja en servidores bajo otras jurisdicciones legales.
- La inmensa mayoría de las aplicaciones (*apps*) que se ofrecen para dispositivos móviles recopilan datos sobre la identidad, localización, preferencias y costumbres de sus usuarios. Muchas de esas aplicaciones comerciales son utilizadas por servicios bibliotecarios o de información con regularidad, y obviamente esas empresas comparten con terceras partes los datos que compilan. Una gran cantidad de esas aplicaciones son en apariencia "gratuitas", pero debe tenerse en mente que cuando un usuario no paga por un producto o servicio en la red, indefectiblemente el producto es él. En estos casos, los usuarios siempre "pagan" la aplicación con sus datos.

Si bien la privacidad de los usuarios de una biblioteca no es una tarea fácil, tampoco es imposible: es un problema con solución. Aunque éste se ha agravado sensiblemente con el desarrollo de la tecnología, no es un problema que se resuelva a partir de ella: como muchos otros en la biblioteca, es cuestión mayormente de método y procedimiento. El problema de la protección de datos –que es la parte tecnológica– debe ser atendido, pero es solo un componente menor. Como ya fue establecido, la parte mayor contempla aspectos todavía más amplios.

En primer lugar, para poder desarrollar método, cada biblioteca debe construir unas "políticas de privacidad y protección de datos" específicas para su contexto y características. Las políticas proporcionan a la biblioteca el gran cimiento estructural para la planificación y los programas de acción destinados a proteger la privacidad y los datos personales de sus usuarios. Deben contemplar las cuestiones éticas y legales que constituyen el marco de referencia de la organización, y deben establecer los grandes rubros acerca de quiénes definirán las cuestiones al interior de ella acerca de privacidad y los datos personales, quiénes harán y actualizarán los planes y programas institucionales al respecto, quiénes lo supervisarán, quiénes definirán a aquellos que pueden acceder a los datos, quiénes definirán cómo se transfieren, y quiénes los resguardarán. Las políticas se redactan a nivel teórico y macro-institucional, y por esa razón tienden a ser mucho más estables en el tiempo. Por supuesto las políticas deben ser consistentes con la legislación vigente aplicable a cada país.

Partiendo de las políticas como base, la biblioteca puede comenzar a desarrollar procedimientos, guías, buenas prácticas, estándares, etcétera. Los procedimientos y las guías son las versiones prácticas que instrumentan los conceptos esbozados teórica y gruesamente en las políticas, detallando acciones preestablecidas y secuenciales que cubran toda una variedad de procesos y secciones de la biblioteca. A diferencia de las políticas, los procedimientos y las guías son específicos, y por tanto pueden y deben cambiar con cierta frecuencia conforme se requiera. Por este motivo, no deben incorporarse elementos de los procedimientos y guías dentro de las políticas o viceversa. Las políticas establecen el por qué, el qué y el quién; los procedimientos y guías establecen el cómo, cuándo, dónde y en su caso detallan los quiénes. La experiencia acumulada irá generando las "buenas prácticas" y los estándares.

Para construir el conjunto de políticas y procedimientos, se sugiere partir de los principios y las técnicas recomendados por la "gobernanza de datos". Con ellos, la organización puede desarrollar con detalle todos los aspectos acerca de quién puede colectar y tener datos personales en la organización, quién puede acceder a ellos autorizadamente y en qué circunstancias; quién los puede modificar,

a quién y cómo pueden ser transferidos, y en la parte de protección de datos, quién los custodia. Para poder asegurar que todos los puntos de colecta de datos han sido revisados y se instrumenten sendos procedimientos al respecto, conviene diseñar y realizar "auditorías" de todos los servicios que ofrece la biblioteca, tanto los internos como aquellos proporcionados a través de proveedores; por supuesto esto incluye a los proyectos de datos masivos. A partir del conjunto de estos elementos documentales, no habrá ya en la biblioteca proceso o conjunto de contenidos que no tenga un responsable, así como una serie de procedimientos y guías para el manejo adecuado de cada grupo de datos personales.

Al respecto de la realización de auditorías en la biblioteca acerca de los puntos de eventual colecta de datos personales con el fin de diseñar procedimientos y responsables para su protección, se presentan aquí los principales puntos:

- Confidencialidad de los registros de préstamo y reserva de obras.
- Confidencialidad en la búsqueda en catálogos internos de la biblioteca. Con periodicidad y método se deben borrar historiales, *cookies*, cachés, etcétera.
- Revisión y certificación por parte de la biblioteca de las condiciones de privacidad de las aplicaciones (*apps*) usadas generalmente en la biblioteca.
- Revisión y certificación de las condiciones de privacidad de los proveedores de bienes y servicios documentales a los que la biblioteca está suscrita. Éstas deben rechazar siempre a aquellos proveedores y servicios que no cumplan con las condiciones mínimas de privacidad de usuarios, y advertir a otras bibliotecas al respecto.
- Revisión y certificación por parte de la biblioteca de las condiciones de privacidad de los servicios que la biblioteca instale en la nube.
- Seguridad y privacidad especial para usuarios jóvenes y/o niños.

- Seguridad de las computadoras de la biblioteca para evitar la introducción de *malware* tendiente a espiar o a extraer información de usuarios.
- Seguridad en la red interna de la biblioteca, en especial la inalámbrica.
- Seguridad de las redes sociales de la biblioteca y a las que accede.
- Cortafuegos y cifrado de datos sensibles de los registros de usuarios colectados por la biblioteca.
- Advertencia, asesoría y capacitación a los usuarios por parte de la biblioteca acerca de riesgos de seguridad al utilizar buscadores, servicios y herramientas externos a la biblioteca.
- La biblioteca debe anonimizar al máximo los datos usados por ella para sus proyectos, estadísticas o retroalimentación.

Este último elemento consignado en la lista es el que se ha vuelto la "regla de oro". Los datos que no se tienen no pueden fugarse ni ser extraídos. En todo servicio o aplicación que diseñe o construya la biblioteca, debe siempre cuestionarse de antemano cuáles datos deben ser colectados para el mismo. En una inmensa mayoría de casos, sucede que los proyectos funcionan sin incluir datos personales o con un mínimo de ellos. Nunca, en ningún proyecto o servicio, debe recabarse este tipo de datos si no es indispensable. En muchos procesos de la biblioteca que recaban datos para análisis es totalmente factible compilarlos sin registrar los datos sensibles de los usuarios; un ejemplo típico de ello es la estadística de uso de las colecciones de una biblioteca universitaria. En estos casos usualmente se registran datos de la obra como la signatura topográfica, autores y título, y puede registrarse del usuario su carrera, semestre, edad, etcétera, los cuales se obtienen de su número de credencial. Pero si la captación final de datos omite este último número y otros datos sensibles –no son necesarios en lo absoluto para la estadística–, la biblioteca puede realizar extensos análisis para el propósito deseado sin necesidad de captar datos personales que estarían eventualmente en riesgo.

En los servicios que requieren identificar al usuario, como es el caso del préstamo, reserva de obras, consulta a publicaciones provenientes de un proveedor, etcétera, la biblioteca debe asegurarse siempre de que recaba solo el mínimo de datos personales, aquellos estrictamente necesarios. En muchos casos, pueden utilizarse ciertos métodos para ocultarlos en las aplicaciones; esto es, "anonimizar" los datos, como lo han hecho las bibliotecas desde largo tiempo atrás. El antiguo ejemplo mencionado de la sustitución de números de credencial en lugar del nombre del usuario en las tarjetas de préstamo sigue siendo vigente en el mundo digital. Utilizar "alias" a todo lo largo de los sistemas y ficheros informáticos de la biblioteca –ya sean numéricos o textuales– en lugar del nombre del usuario sigue funcionando muy bien al registrar préstamos, reserva de obras, apartado de tiempo de computadora, consulta de catálogos, acceso a servicios de proveedores y muchos otros puntos que requieren identificar a un cierto usuario. Por supuesto, al interior de la biblioteca existe un registro principal que tiene todos los datos del usuario, pero ése debe encontrarse centralizado en un solo lugar, en una buena "bóveda" informática, atrás de algunos cortafuegos y otros mecanismos de seguridad y de preferencia en ficheros cifrados.

En lo relacionado con bienes y servicios adquiridos a proveedores, la ya mencionada Regulación de Protección de Datos Personales de la Unión Europea o GDPR, además de ser considerada en la actualidad como el mecanismo regulatorio más avanzado acerca de la privacidad de datos, es también uno de los aliados más valiosos para las bibliotecas en esta tarea, ya que incluye todas las regulaciones esenciales hacia los proveedores. Gradualmente se va convirtiendo en un marco de referencia mundial, y por esta razón se recomienda a las bibliotecas que se encuentran fuera de la región europea que revisen siempre las condiciones de servicio de sus proveedores para verificar hasta qué grado ellos cumplen con este marco regulatorio. Los puntos esenciales que la GDPR cubre en este aspecto y que es recomendable que las bibliotecas verifiquen meticulosamente son:

1) Privacidad por diseño: las empresas deben construir y ofrecer sus procesos de negocio incluyendo la privacidad de los datos desde el principio, y no como una idea tardía. Siempre deben tener un responsable de la protección de datos y la privacidad, fácilmente identificable, e independiente de la operación.
2) Consentimiento explícito: los usuarios deben siempre poder aceptar o rechazar explícitamente los términos y las condiciones de privacidad, incluyendo la aceptación de *cookies*, antes de acceder a un servicio o producto.
3) Restricción de datos: las empresas no pueden colectar datos sensibles como raza, religión, afiliación política u orientación sexual.
4) Derechos de acceso y portabilidad: los usuarios tienen derecho a solicitar cuál información personal se ha recabado de ellos, y a solicitar su información personal a otras empresas que la hayan recibido de segunda o tercera mano.
5) Derecho al olvido: los usuarios pueden solicitar que sus datos sean borrados de una cierta lista.
6) Notificación de intrusiones: las empresas deben informar siempre a los usuarios cuando haya una violación de datos dentro de las 72 horas de su descubrimiento.
7) Revisión de la jurisdicción: los datos personales deben residir en servidores conocidos en una cierta jurisdicción identificada, y no deben ser trasladados arbitrariamente fuera de ella sin consentimiento expreso de los usuarios.[25]

Como ha podido verse, existen metodologías relativamente sencillas para reducir e inclusive eliminar el uso de datos personales a lo largo de muchas de las aplicaciones y servicios de la biblioteca. No es conveniente duplicar registros de datos personales a lo

25 El Tribunal de Justicia Europeo estableció en julio de 2020 que no es válido que los proveedores transfieran datos personales de Europa a Estados Unidos (https://curia.europa.eu/jcms/upload/docs/application/pdf/2020-07/cp200091es.pdf).

largo de cada departamento, sección, servicio o proyecto en la biblioteca. Al minimizar en ella el número de puntos donde se manejan datos personales, la cantidad de lugares a cuidar se reduce sensiblemente, lo cual facilita enormemente la tarea a los responsables de ello.

El conjunto de políticas, procedimientos, guías, estándares, auditorías y metodologías aquí consignadas hace evidente el concepto ya mencionado anteriormente de que la mayor parte de todo ello consiste en método y procedimiento, y solo una mínima parte son elementos tecnológicos. Lo más importante sigue siendo *cómo* se hacen las cosas, y no *con qué* tecnología se hacen. En efecto, existen técnicas y herramientas tecnológicas para incrementar la privacidad en las organizaciones cuyo uso puede contemplarse de inicio en las bibliotecas, pero es de suma importancia recalcar que de ninguna forma, bajo ninguna circunstancia, pueden sustituir a un buen método de trabajo. Las herramientas tecnológicas son complementos, no sustitutos.

Finalmente, la biblioteca debe hacer conciencia y capacitar con cierta periodicidad tanto a su personal interno, como a sus usuarios en materia de privacidad y confidencialidad de datos personales, seguridad informática y temas afines para que vayan aprendiendo a manejarlo de una manera adecuada. Con respecto al personal, todos los nuevos empleados, becarios, voluntarios, asistentes, etcétera, deben ser concientizados desde el inicio de que deben proteger los derechos de privacidad de los usuarios de la biblioteca. Con respecto a los usuarios, aun cuando más de la mitad de la humanidad se conecta ya de alguna forma a la red y al mundo digital, sigue existiendo un enorme analfabetismo funcional en esta temática. Recuérdese el segundo hallazgo ya mencionado presentado en el Foro Económico Mundial de 2019: solo un tercio del público tiene conocimientos aceptables acerca de la privacidad de datos. Si bien no ha sido una práctica generalizada, la mejor manera de que la biblioteca puede reforzar sus iniciativas en materia de confidencialidad es alfabetizar en lo posible y periódicamente a sus usuarios y a su personal en lo relativo a esta temática.

Las herramientas para los datos masivos

Es un error capital teorizar antes de tener datos. Insensiblemente uno comienza a torcer los hechos para adaptarlos a las teorías, en lugar de elaborar teorías que se adapten a los hechos.

ARTHUR CONAN DOYLE,
Un escándalo en Bohemia. 1891

Para el manejo de datos masivos, existe hoy en día un sinnúmero de "herramientas"; se denomina así genéricamente a los sistemas, programas y aplicaciones informáticos, metodologías, algoritmos, "areneros",[26] etcétera, tanto comerciales como de acceso abierto. A la fecha no existe la gran aplicación universal que abarque todas o al menos la mayoría de las necesidades al respecto; en cambio, existen múltiples herramientas puntuales con altas especialidades para cada tipo de propósito, producidas además por variados fabricantes. Ningún proyecto de datos es igual a otro; cada uno tiene su contexto y características propias. Así, al construir un cierto proyecto de datos muy rara vez es posible adquirir un único producto para su solución; siempre, indefectiblemente, se requiere armar un conjunto de herramientas informáticas para su implementación y solución. Por ello, uno de

26 Un "arenero" (del inglés *data sandbox*), en el contexto de los datos masivos, es una plataforma de prueba y desarrollo que se utiliza para que las organizaciones simulen en pequeña escala conjuntos de datos, su comportamiento, análisis y resultados para verificar cómo se comportará el conjunto ya en explotación plena.

los conocimientos clave de inicio en el campo de los datos masivos consiste en adquirir una buena idea general de todas las posibilidades y tipologías entre esas herramientas, para así estar en posibilidad de seleccionar la adecuada para cada caso y necesidad. Las hay de muchos tipos; como una lista indicativa y general, existen aplicaciones para:

- Extracción de datos de muy diversas fuentes: de redes sociales, de programas específicos, de la red, etcétera.
- Extracción de datos a partir de textos, imágenes o voz.
- Posterior análisis de los datos y descubrimiento de patrones.
- Visualización, graficación y/o presentación de los resultados en formas coherentes.
- Análisis de datos textuales.
- Manejo y gestión de datos masivos (en grandes volúmenes).
- Inteligencia Artificial; en especial aprendizaje de máquinas.

Por su especialización, ciertas herramientas se han vuelto de uso común para ciertos sectores: comercial, financiero, de salud, educativo; y por supuesto, también las hay para uso específico en bibliotecas y otras organizaciones de información.

Al ser tan numerosas, ninguna persona conoce o domina todas las herramientas existentes, pero es indispensable que todo profesional de la biblioteca conozca en términos generales la oferta, así como las posibilidades y capacidades de aquellas, para así poder descartar o aproximarse a una eventual selección de alguna de ellas, al igual que lo hace con Sistemas Integrados de Gestión Bibliotecaria o ILS, buscadores o descubridores especializados, catálogos automatizados u OPAC,[27] sistemas de automatización de bibliotecas, y otros semejantes. Ningún responsable de biblioteca domina o conoce a profundidad la totalidad de los existentes, pero

27 Se denomina Online Public Access Catalog (OPAC) al Catálogo de Acceso Público en línea, conocido también como Catálogo Automatizado de la Biblioteca. Consiste en una base de datos en línea de los registros de los materiales documentales que posee una cierta biblioteca o grupo de bibliotecas.

generalmente conoce lo suficiente de todos como para eventualmente poder seleccionar y adquirir adecuadamente uno de ellos; principalmente, aprende qué es lo que hay que preguntar, en qué hay que fijarse y cuáles pruebas hay que hacer para llegar a una adecuada selección de un eventual sistema.

Además de ello, hoy en día existen muchos productos, sistemas, aplicaciones, etcétera, alrededor de los datos que ya están siendo desarrolladas por grandes instituciones de ese sector y que han puesto a disposición de bibliotecas más pequeñas, las cuales pueden aprovechar todos estos productos y servicios en su beneficio con poca o nula inversión. Ya han sido mencionados los ejemplos del Servicio de Datos Enlazados de la Biblioteca del Congreso de los Estados Unidos, el de la Biblioteca Británica, la Red de la Biblioteca Nacional de Medicina también de ese país (NNLM), las guías de las Bibliotecas del Instituto Smithsoniano para la creación y depósito de datos en repositorios, etcétera. Como estos, existen ya múltiples ejemplos de desarrollos de grandes sistemas bibliotecarios acerca de datos masivos que pueden ser utilizados en beneficio de sistemas o bibliotecas más pequeñas sin grandes inversiones. Éste se ha ido convirtiendo gradualmente en un gran filón para que ellas adquieran buenas aplicaciones al respecto. Igualmente, los responsables de bibliotecas deben comenzar a conocer y estudiar todas estas posibilidades en el campo de los datos masivos para distinguir las opciones y así poder tomar iniciativas sin grandes erogaciones.

Además de lo anterior, muchos de los proveedores usuales de las bibliotecas también están haciendo desarrollos al respecto de los datos masivos, con quienes las bibliotecas pueden hacer alianzas estratégicas para su aprovechamiento e impulso, o simplemente adquirir en condiciones preferenciales algunos de los productos emanados de esos desarrollos para su beneficio. Ya se han mencionado como ejemplos el proyecto SN SciGraph de datos abiertos enlazados de la división de Ciencias Naturales de Springer Nature, o los estudios textuales de OCLC.

Fuera del ámbito de las bibliotecas –pero no por esto ajeno a ellas– existen también múltiples opciones de herramientas

aprovechables. Hoy en día se encuentra una gran cantidad de proveedores dedicados a alguna de las facetas del proceso y análisis de los datos. Existen dos grandes vertientes en este aspecto: por un lado, los proveedores que ofrecen algún producto específico para que sea adquirido por una organización y lo integre a sus proyectos, y por el otro los proveedores que suministran todo en forma de "paquetes" ofreciendo equipos, programas, procesamiento, almacenamiento, etcétera, en el esquema conocido como la nube. Todas las principales plataformas de esta modalidad ya están ofreciendo una serie de servicios relacionados con los datos masivos; algunos gratuitos y otros de paga. Esta segunda vertiente de servicios en la nube comienza a ser cada vez más utilizada, ya que permite a los usuarios adquirir grandes recursos con relativamente menores inversiones.

El manejo de datos masivos en la nube

> *Si crees que has visto esta película antes, tienes razón. El cómputo en la nube se basa en el modelo de tiempo compartido que usamos hace años antes de poder comprar nuestras propias computadoras.*
>
> David Linthicum.
> *Cloud computing and SOA convergence in your enterprise. 2009*

Existen muchas definiciones y conceptualizaciones de "la nube", pero no es el caso aquí entrar en un estudio detallado. Para comprensión del concepto y para ponerlo de manera simple, el "cómputo en la nube" consiste en un conjunto de recursos informáticos de equipo, programas y aplicaciones, almacenamiento, procesamiento, comunicación, información, etcétera, que pueden ser rápida y ubicuamente suministrados como servicio vía una red por un cierto proveedor y ampliamente escalados en función de las necesidades de un cierto usuario. La diferencia primordial de este

esquema con sus predecesores consiste en que por varias décadas el modelo comercial del suministro de equipo de cómputo, programas, comunicaciones, etcétera, fue manejado como la provisión de productos. Bajo este nuevo concepto, el modelo de negocio del cómputo en la nube –o simplemente "la nube"– consiste en la entrega de insumos informáticos como un servicio en vez de como un producto, a través de recursos compartidos sobre una red, y en el cual equipos, aplicaciones, almacenamiento, información, datos, infraestructura, etcétera, son proveídos al igual que los servicios comunitarios de agua, electricidad o gas, pagando solo lo que se consume.[28]

Originalmente, hace aproximadamente una década, se dividieron las variantes del cómputo en la nube en tres *modelos de servicio* básicos, también llamados "capas":

- *Software* como servicio *(Software as a Service* o simplemente SaaS*)*. En este modelo de servicio, el usuario utiliza aplicaciones o programas que se ejecutan en un servidor remoto de un proveedor en red –en la nube– y no gestiona ni controla infraestructuras o plataformas en la que dichas aplicaciones corren, como el tipo y/o modelo de servidores, sistema operativo, redes, etcétera.
- Plataforma como servicio (*Platform as a Service* o PaaS). En este modelo, el usuario puede desarrollar sus propias aplicaciones o sistemas, ya sea para su red local o en la nube, y para ello renta el acceso a una plataforma de programación, lo cual significa que puede seleccionar y utilizar sistema operativo, librerías, compiladores, paquetes, capacidad de almacenamiento, etcétera.
- Infraestructura como servicio *(Infrastructure as a Service* o IaaS*)*. En este modelo de servicio, el proveedor proporciona

28 Véanse para más detalle al respecto Juan Voutssás (2013). "Documentos de archivo en *la nube"* (https://iibi.unam.mx/voutssasmt/documentos/legajos17_nube_corto.pdf) y Alejandro Delgado (2013) "La nube" (https://iibi.unam.mx/p_a/archivistica/AGN%20legajos16-delgado.pdf).

> a sus clientes toda una infraestructura básica de cómputo y telecomunicaciones, normalmente bajo un esquema que proporciona recursos variables bajo demanda operados por el cliente a voluntad. Esta infraestructura es un ensamble de equipo, programas, redes, ayuda, seguridad, capacitación, etcétera.

Posteriormente, con la evolución de la nube, muchas otras variantes de modelos de servicio han sido agregadas. Entre otros, existen ahora además Contenidos como Servicio (CaaS), Red como Servicio (NaaS), Seguridad como Servicio (SaaS), Preservación como Servicio (PaaST) y, por supuesto, nuevos servicios dedicados expresamente a los datos: Datos como Servicio (DaaS), así como Datos y Plataforma como Servicio (DaPaas).

En la etapa anterior a la nube, las organizaciones alojaron y procesaron sus propios datos en un sistema informático y autónomo de su propiedad. El inconveniente de ese modelo consistía en que, a medida que los datos se volvían cada vez más voluminosos y complejos, su operación y mantenimiento se volvía a la vez más costoso. Con el modelo de Datos como Servicio (DaaS) en la nube, los datos son compilados, organizados y suministrados por un proveedor como un servicio rentado, y están fácilmente accesibles a través de la red. La gestión y el seguimiento de clientes se volvieron así un servicio rentado en plataformas externas a la organización. Pero los Datos como Servicio son solo la parte más evidente de los recursos utilizados en el ambiente de la nube: como se desprende de lo anterior, puede rentarse adicionalmente todo tipo de infraestructuras: equipo, almacenamiento, red; *software* para los proyectos: desde sistema operativos, hasta programas y aplicaciones generales y específicas para datos, librerías, etcétera.

Además, los servicios en la nube se van diversificando y sofisticando incesantemente, por lo que hay que estar atento a estudiar las variantes que van surgiendo, con el fin de evaluar su eventual selección o descarte. Como ejemplo de ello, se encuentra el llamado Cómputo en el borde *(Edge computing)*. En los casos de renta de Datos como servicio (DaaS), el envío de esos datos generados

por dispositivos a un servidor central en la nube llega a causar problemas de ancho de banda y tiempos de demora. El "cómputo en el borde" preconiza una alternativa más eficiente: procesar y analizar los datos en la ubicación física del usuario, en la fuente de datos o cerca de ellos. Al no tener que viajar los datos a través de toda la red hasta un servidor central para su proceso, el tiempo de demora se reduce significativamente, y se obtienen servicios más rápidos y confiables. No está clara todavía la utilidad y eficiencia de esta modalidad, pero Gartner afirma que para el año 2025 este tipo de variante de servicio en la nube comprenderá el 25 por ciento de la actividad en este ambiente, por lo que conviene volver a evaluarla de tiempo en tiempo (Van der Meulen 2018). Como ésta, hay otras variantes surgiendo con cierta frecuencia.

Por su naturaleza masiva, es común hoy en día que una buena parte del manejo de datos se haga en algún recurso en la nube. Por tanto, una de las primeras decisiones a tomar en este aspecto consiste en decidir qué proporción de servicios se va a utilizar en ese ambiente y cuáles de ellos, y también la contraparte: qué proporción de datos o aplicaciones será utilizada dentro del ambiente informático propiedad de la biblioteca o de la organización de la cual depende. Esto debe formar parte del plan desde el principio, ya que tiene que ver directamente con costos, tiempos, infraestructura, seguridad y privacidad. Aunque en general los servicios en la nube se adquieren por economía de escala, esto no necesariamente es una regla absoluta: muchas bibliotecas que ya cuentan con infraestructura informática propia pueden utilizar sus remanentes de capacidad para proyectos de datos con costos muy reducidos, sin necesidad de adquirir muchos servicios externos. Los estudios y las comparaciones costo-beneficio deben ser realizados desde el inicio para estar en capacidad de tomar las decisiones adecuadas en este aspecto.

Ello hace que las combinaciones de infraestructura propia y rentada sean muy amplias: significa que puede rentarse almacenamiento pero no proceso; *software* pero no equipo; proceso pero no datos, etcétera. Como en muchas otras iniciativas tecnológicas en la biblioteca, una aproximación comúnmente utilizada en ellas

consiste en adquirir de inicio plataformas gratuitas o de uso abierto para empezar a probar las ideas, desarrollar aplicaciones, verificar objetivos, etcétera. Obviamente estas plataformas son limitadas en cuanto a alcance o características por su propia naturaleza gratuita. Pero sirven muy bien para validar los proyectos. Una vez estables, y si la biblioteca se ha convencido de su utilidad y beneficio, puede cambiarse a las versiones de paga de esas plataformas.

Hoy en día prácticamente todos los gigantes del mundo de las Tecnologías de Información y Comunicaciones –Google, Amazon, Microsoft, Apple, Facebook, Oracle, IBM, etcétera– ofrecen algún tipo de producto o servicio relacionado con la gestión y el análisis de datos, algunos de estos en tipo abierto o gratuito con fines promocionales y otra buena parte de paga. Las bibliotecas se han aprovechado de estas facilidades desde hace algunos años, y las han incorporado a sus quehaceres. Las posibilidades son sumamente extensas: desde extracción e inserción mínima de datos para servicios puntuales, hasta grandes proyectos relacionados con ellos. Por supuesto no todo lo que ofrecen esas grandes corporaciones tiene que ver con datos, pero una parte muy significativa de sus productos y ganancias provienen de este sector. Aquí se revisarán los servicios y productos relacionados cercanamente con los datos.

Más allá de su famoso buscador, la empresa Google comenzó ofreciendo algunos servicios básicos en la nube, y gradualmente fue incorporando nuevas herramientas, aplicaciones, servicios, etcétera: todo un abanico de posibilidades de procesamiento y servicios tecnológicos; entre ellos, comenzó a haber en algún punto manejo de datos. En tiempos recientes, decidió integrar todo lo relacionado con servicios en la nube en un solo sitio para darle mejor visibilidad y acceso, y creó la estructura denominada Google Cloud. En este sitio ofrece una gran cantidad de posibles servicios de todo tipo en la nube, como los ya mencionados: acceso a servidores, almacenamiento, programas, etcétera. Muchos de esos servicios y productos tienen que ver o pueden utilizarse cercanamente con el manejo de datos. Por ejemplo, Google Charts; esta herramienta de acceso abierto se implementa fácilmente incrustando código de programación JavaScript en el código HTML

de un cierto sitio web y permite extraer datos de él, conectarse a una base de datos, clasificar, filtrar, modificar y/o visualizar sus datos. La nube de Google también ofrece acceso a súper procesadores llamados Tensor Process Unit (TPU) –usados en su buscador, traductor, identificador de fotografías, el asistente de Google y Gmail– con el fin de que los clientes desarrollen aplicaciones de Inteligencia Artificial en equipos de capacidad ampliada.

Otro ejemplo típico de extracción y uso de datos con la plataforma Google son los desarrollos conocidos como "mezclas" (*mashups)*. Básicamente, consiste en una página web que usa y combina datos, presentación y funcionalidades procedentes de una o más fuentes para crear nuevos servicios enriquecidos de forma fácil y rápida, integrando a la vez aplicaciones, fuentes de datos y sitios abiertos a través de una interfaz gráfica. Con ello, una biblioteca puede colectar un cierto conjunto de datos y luego integrarlos con la aplicación de acceso abierto Google Maps, y gracias a ello desplegar al usuario visualizaciones que presenten gráficamente una serie de datos e información propios de esa biblioteca, con la presentación, apariencia y facilidades típicos de los mapas de esa empresa. Google Mashup Editor es otra herramienta de esta plataforma que permite hacer mezclas de todo tipo con datos de las bibliotecas; pueden verse múltiples casos prácticos de servicios bibliotecarios hechos con mezclas en Engard (2009 y 2012) y en Stephens (2011). Existen otras plataformas que pueden ser utilizadas para crear una mezcla: Amazon, Facebook, Twitter, LibraryThing, Flickr, Ebay, YouTube, JournalTOC, etcétera. A la vez, existen múltiples programas o "editores" provenientes de proveedores para construir la mezcla, tanto en versiones libres como pagadas; entre ellas: Yahoo Pipes, Microsoft Popfly, IBM Mashup Starter Kit, Intel's Mash Maker, y DreamFace Interactive. Como se mencionó, las versiones libres de ellos son opciones adecuadas y suficientes para comenzar, a pesar de ser básicos; una vez dominada la técnica, conviene migrar a editores con costo para poder construir mezclas más complejas y sofisticadas, puesto que ofrecen más opciones y detalle.

Del amplio número de productos y proveedores de esta aplicación, se desprende la importancia de que el responsable de la

biblioteca comience a conocer y estudiar todas las posibilidades para distinguir las opciones y así poder tomar iniciativas en el campo de los datos de manera concertada. Dado que todas las plataformas ofrecen servicios parecidos, para iniciar adecuadamente la aproximación a ese conocimiento pueden utilizarse dos formas: por un lado, estudiando lo que ofrece cada una de las plataformas al respecto en cada una de sus variantes, y así conocer la oferta total de una cierta plataforma, y por otro, estudiando las herramientas por tipo para conocer todas las opciones en una cierta variedad de servicio en todas las plataformas y así poder hacer comparaciones entre ellas. Las dos aproximaciones funcionan bien en la práctica, y es cuestión de gustos.

Bajo el primer enfoque o aproximación, es decir, por empresa o plataforma, prácticamente todos los principales y conocidos proveedores de tecnología ofrecen diversas herramientas de tratamiento y análisis de datos: Google, Amazon, Facebook, Twitter, Apple, IBM, Microsoft, LinkedIn, YouTube, Flickr, Ebay. Entre ellas, se distinguen como las principales plataformas en la nube Google Cloud Platform, Amazon Web Services o AWS, Microsoft Azure e IBM Cloud, las cuales se irán estudiando con más detalle posteriormente. A éstas habría que agregarles algunos grandes proveedores de este tipo de herramientas no tan conocidos por el público en general, como Apache, Oracle, Hana o Gaia-X. El estudio de las herramientas ofrecidas por cada una de estas plataformas es la primera manera de aproximación hacia ellas. Es decir, se selecciona una empresa y se revisan todas las posibles herramientas de manejo y análisis de datos que ésta puede proporcionar con el fin de obtener un panorama general de toda la oferta de datos de esa organización. Y así, se continúa haciendo lo mismo con las demás empresas.

El segundo enfoque o aproximación consiste en ir analizando las herramientas por cada tipo de ellas, comparando entre las diversas ofertas proporcionadas por diversas empresas. Este segundo enfoque será utilizado para la revisión de posibilidades. Para este propósito y de manera arbitraria dividimos las herramientas en los siguientes tipos:

- Los manejadores de bases de datos, tanto SQL como NoSQL.
- Los manejadores de datos documentales.
- Las herramientas de "normalización" y mapeo de datos.
- Las herramientas para el análisis de datos masivos, con el fin de extraer patrones o tendencias de ellos.
- Las herramientas de visualización, interpretación o presentación de resultados.
- Las herramientas de Inteligencia Artificial.
- Herramientas de aplicación específica, como las mencionadas para mezclas, etcétera.

Los manejadores de bases de datos

> *Los Petabytes nos permiten decir que 'la correlación es suficiente'. Podemos dejar de buscar modelos. Podemos analizar los datos sin hipótesis para ver qué nos pueden mostrar. Podemos introducir las cifras en enormes servidores computacionales y dejar que los algoritmos estadísticos encuentren lo que la ciencia no puede ver.*
>
> Chris Anderson,
> antiguo editor en jefe de la
> revista *Wired*, 2018

Se han mencionado los manejadores de bases de datos "SQL" y "NoSQL". La diferencia fundamental entre ellos es que los primeros fueron diseñados para manejar datos estructurados y los segundos se especializan en datos no-estructurados. Como ya se ha mencionado, de acuerdo con la naturaleza de los datos y de su emisor, se distinguen tres tipos de ellos: datos estructurados, semi-estructurados y no estructurados. Es importante estudiar estos conceptos de forma básica para comprender su importancia y diferencias.

Los datos estructurados tienen una forma y un formato bien definidos, por lo que pueden representarse fácilmente de forma homogénea en tablas con renglones y columnas. Cada renglón se

conoce como un "registro"[29] y cada columna como un "campo". Una secuencia contigua de columnas o "campos" conforma un renglón o "registro". Provienen de un "modelo de datos", también llamado "esquema de datos"; es decir, una forma preestablecida de cómo se pueden procesar, almacenar y acceder; en consecuencia, son más fáciles de gestionar o manejar. Gracias a su estructura, cada campo contiene un único elemento fácilmente identificable y por tanto se puede acceder a él por separado o conjuntamente con los datos de otros campos.

Los "modelos de datos" utilizan ciertos elementos típicos, divididos en dos categorías; elementos estáticos: campos, registros, arreglos, cadenas de caracteres y elementos dinámicos: listas, pilas, colas, árboles. Cada uno de estos términos tiene un significado particular dentro del manejo de datos. De forma breve podemos mencionar algunos ejemplos en el campo de las bibliotecas para comprender mejor estos significados particulares*:* Un caso típico desde esta conceptualización de "registro" es cada entrada del "directorio de usuarios" de la biblioteca. Cada "registro" del mismo contiene los datos de un usuario, en forma de "campos" perfectamente definidos y preestablecidos con los datos de cada uno de ellos: número de credencial, nombre, escuela, carrera, dirección, fecha de nacimiento, teléfono, correo electrónico o algo semejante. La adición consecutiva de todos estos registros para cada usuario puestos en esta forma se convierten en el directorio. Todos y cada uno de esos campos tienen un tipo predefinido por un "modelo de datos": numérico, alfanumérico, fecha, etcétera; así como un cierto formato con longitud también predefinida: diez caracteres

29 En inglés *record*. Este concepto de "registro" procede de la informática y difiere de los significados de la ciencia bibliotecaria y la archivística. Desde esta conceptualización, un "registro" es un conjunto de "campos" consecutivos en forma de renglón. Cada campo contiene datos de un tipo determinado –numérico, alfabético, fecha, etc.–. Un conjunto de registros consecutivos conforma un "archivo" o "fichero".

para un teléfono, ocho caracteres para la fecha de nacimiento en formato dd/mm/aa, cinco dígitos para el código postal, treinta caracteres alfanuméricos para el correo electrónico, o algo parecido.

Un "arreglo" es una colección finita de elementos del mismo tipo en formato definido; es decir, son homogéneos y su característica distintiva es que están ordenados por medio de un índice. Un ejemplo típico de un *arreglo* es una tabla predefinida con las claves numéricas de los nombres de las diferentes bibliotecas dentro de una universidad. Para mantener homogéneos los datos, durante los procesos se captura su clave y no su nombre, y solo eventualmente cuando se desea se asocia nuevamente la dupla clave-nombre y se despliega el nombre completo. De esta forma, se ahorra espacio y se mantiene homogeneidad en los nombres.

A su vez, los "campos" o ciertos elementos de un "arreglo" pueden contener "cadenas de caracteres": secuencias de caracteres alfanuméricos con un cierto significado. En su expresión más simple, son elementos cortos, como el nombre de una persona, una dirección postal, un correo electrónico, etcétera. Cuando estas cadenas son cortas y precisas, se pueden manejar como datos estructurados.

Los elementos de tipo "dinámico": listas, pilas, colas, árboles, se utilizan en los sistemas informáticos como estructuras auxiliares para almacenamiento y recuperación; un ejemplo típico de ello son las "listas de palabras" que se construyen para recuperación en el catálogo de un OPAC; en realidad, el sistema de cómputo guarda listas donde cada palabra ha sido traducida a un número y apunta a todos los números de fichas que la contienen; de esta forma se ahorra tiempo y espacio y se incrementa sensiblemente la velocidad de recuperación. Se realizan operaciones semejantes con pilas, colas, árboles, etcétera; cada uno con un propósito específico.

Como puede verse, en este tipo de datos su acceso y manejo se facilita debido a su estructura. Desde hace décadas, se crearon teorías, principios, lenguajes, algoritmos, herramientas, etcétera, para manejarlos. Toda la teoría inicial de las bases de datos se creó alrededor de la tipología antes mencionada. Desde entonces, los Data Base Management Systems (DBMS o Sistemas Manejadores de Bases de datos) jerárquicos y relacionales se diseñaron y

construyeron en innumerables versiones y por múltiples fabricantes alrededor de estos conceptos. Para más información acerca de bases de datos relacionales, véase Oracle (s.d.). Todos estos manejadores pertenecen a una gran categoría denominada "manejadores SQL", ya que la "forma" o lenguaje común que permite realizar consultas sobre este tipo de datos se denominó desde un inicio Structured Query Language (SQL) o Lenguaje de Consulta Estructurado. De ahí el nombre de Bases de Datos SQL.

Los datos *no estructurados* tienen una forma interna pero no están estructurados mediante modelos o esquemas de datos predefinidos. No necesariamente son numéricos o textuales, y son generados tanto por personas como por máquinas. Es muy difícil normalizarlos, pues no tienen tipos definidos ni están organizados bajo algún patrón. No obstante, en muchos casos se desea procesar este tipo de datos, y por lo mismo es necesario poder organizarlos, clasificarlos, almacenarlos, buscarlos, borrarlos, etcétera, de alguna manera, y es indispensable encontrar formas de hacerlo. Un ejemplo típico de estos datos son los contenidos de una página web: la página en sí tiene una estructura interna en formato HTML, pero en ella sus contenidos están dispersos y es muy difícil identificar los datos de sus partes: textos, imágenes, audios, videos, botones, cajas, etcétera, que están embebidos en ella. Extraer información a partir de sus datos internos es una tarea muy complicada. Lo único que contiene la página como metadatos son sus etiquetas "meta" de HTML: palabras clave, descripción, etcétera, pero estos son datos mínimos que dificultan sobremanera extraer el resto de ellos de la página para su posterior procesamiento.

Otro ejemplo típico de este tipo de datos son los contenidos en videos. Si estos no vienen acompañados ya de una "ficha" o registro con sus metadatos, es sumamente difícil extraer información de ellos: ¿Quiénes son las personas que aparecen ahí? ¿Cuáles son las fechas, lugares, formato, duración? ¿Quiénes lo produjeron, dirigieron, editaron, escribieron, fotografiaron? ¿De qué trata el video? ¿Cuál es su temática? En ciertos casos esta información aparece escrita en créditos en el video, pero en muchos otros no. Y aun cuando aparece, es necesario que alguna persona observe el

video para extraer esa información personalmente, pues es difícil lograr que una máquina lo haga debido a que la forma de asentarlos nunca ha sido normalizada. Existen muchos otros tipos de documentos cuya extracción se dificulta si no vienen acompañados ya de sus metadatos: textos de procesadores de oficina, presentaciones de Power Point, audios, fotografías, blogs, mensajes de voz, textos o imágenes digitalizadas sin metadatos, etcétera. Por su naturaleza, se desarrollaron bases de datos especiales para almacenar y manejar estos contenidos en estructuras denominadas bases de datos *no relacionales*. Este tipo de bases de datos se conocen como "bases de datos NoSQL" *(No*SQL *Data Bases* o bases de datos No-solo SQL). La mayor diferencia contra los datos estructurados consiste en la facilidad de analizar unos y otros. Para los datos estructurados, existen ya herramientas analíticas maduras, pero las herramientas para el análisis y extracción de datos no estructurados todavía están en etapas de incipiente desarrollo y les falta mucho camino por andar.

En un estadio intermedio entre estos dos tipos de datos se encuentran los datos semiestructurados. Como su nombre lo indica, contienen estructura y forma en algunas partes de su contenido, pero en otras no. El ejemplo más típico de este tipo de datos dentro de una biblioteca es una ficha catalográfica; como es sabido, contiene una serie de campos, cada uno con una etiqueta que lo identifica seguida de unos indicadores, y luego el contenido en sí del campo; también puede incluir algunos marcadores: coma, porciento, dos puntos, etcétera. Una ficha catalográfica completa es una "cadena de caracteres" compleja, ya que está formada a su vez por subcadenas; esto es, cada uno de los campos de la ficha, siendo todas ellas también "cadenas de caracteres": número de clasificación, autor, título, pie de imprenta, materia, etcétera. Por sus dimensiones y forma, ya no se considera un dato estructurado, como es el caso de una cadena de caracteres corta o simple. Se considera dato semi-estructurado, ya que sí contiene divisiones preestablecidas e identificables –las señaladas por las etiquetas–, pero dentro de ellas los elementos textuales pueden ser muy extensos y complejos, como los autores corporativos o los títulos, no

fácilmente separables. Además, su longitud muy extensa y totalmente variable hace impráctico definir campos de una longitud fija al efecto.[30]

Otro ejemplo típico de este tipo de datos son los mensajes de correo electrónico, los cuales por su forma cuentan por un lado con una serie de campos bien definidos y preestablecidos: nombre y dirección e IP del emisor, nombre y dirección del o los destinatarios, fecha, asunto, número de identificación del mensaje, anexos, etcétera. No obstante, el contenido en sí del mensaje no está estructurado; es simplemente un texto en forma de una cadena de caracteres sin mayor definición, y es la parte más importante. Dado que toda esta parte carece de estructura, es difícil extraer datos de esta sección. Otro ejemplo adicional de datos de este tipo son los documentos XML. La parte de las etiquetas definidas tiene una buena estructura; la parte de los contenidos de esas etiquetas no necesariamente. La gran ventaja es que esa estructura basada en etiquetas es altamente flexible y adaptable a variadas necesidades para homogeneizar en lo posible ciertas formas de datos de muchos tipos de documentos, y esas estructuras son legibles por máquinas, lo cual facilita el trabajo.

Los sistemas manejadores de bases de datos no estructurados NoSQL pueden manejar también los datos semiestructurados, ya que estos manejadores se diferencian de aquellos relacionales porque no separan los datos del modelo o esquema de organización de datos. Esto los hace una mejor opción para almacenar y procesar la información que no cabe fácilmente en el formato de tabla o registro; este es el caso de los textos de longitudes muy variables. Además, permiten un intercambio de datos más fácil y fluido

30 Un ejemplo de ello es el campo "título"; existen unos tan cortos como *Ella*, de Rider Haggard, hasta otros tan largos como "Western Central Atlantic Fishery Commission: Report of the fifth session of the Scientific Advisory Group, Puerto Morelos, Mexico, 28-29 October 2011 = Rapport de la cinquième session du Groupe Scientifique Consultatif, Puerto Morelos, Mexique, 28-29 october 2011 = Informe de la quinta sesión del Grupo Asesor Científico, Puerto Morelos, México, 28-29 de octubre de 2011".

entre diferentes bases de datos. De hecho, las fichas de los catálogos de bibliotecas son un excelente ejemplo de este tipo de datos. Por la longitud extensa y muy variable de algunos de sus campos, siempre fue difícil manejarlos en forma de una tabla de renglones y columnas de longitud fija y predeterminada en un sistema informático, y por tanto hubo que imaginar desde un principio otras maneras de poder representarlos y almacenarlos eficientemente. El desarrollo de las "etiquetas" por parte de MARC para los campos de longitud variable fue una de las grandes aportaciones metodológicas a este tipo de datos.[31] Los manejadores de datos especializados en catálogos de bibliotecas son un ejemplo añejo de este tipo de herramientas.

Lógicamente, entre más estructurados estén los datos, más fácil es su manejo. Gran parte del problema consiste en que en el mundo digital se calcula que los datos estructurados representan solo un 5 por ciento de lo que se produce. Los semiestructurados abarcan un 15 por ciento, y el resto son datos no estructurados; se le agregan metadatos a un 3 por ciento de ellos (*The Digital Universe…* 2014). Obviamente lo opuesto también se cumple: entre menos estructura tengan los datos, es más difícil extraer de ellos información de valor; he ahí la gran importancia de los metadatos. Como es sabido, estos son los "datos acerca de los datos"; sin ellos, no sería posible obtener algo utilizable de esos conjuntos de datos, en especial los masivos, y se convertirían en una masa amorfa y estéril con poca o nula utilidad. De aquí se desprende la gran importancia que tienen los proyectos ya mencionados de extracción y correlación de metadatos masivos como los de la Biblioteca Británica o la del Congreso de Estados Unidos, y se reafirma la importancia de que las bibliotecas agreguen metadatos regularmente a su información y colecciones.

---

31 El origen de la estructura de "etiquetas" y "marcado de textos" proviene del metalenguaje General Markup Language (GML), desarrollado originalmente por la empresa IBM para edición de sus manuales. El proyecto MARC perfeccionó la idea y la volvió de aplicación universal en los años sesenta.

Como puede deducirse de lo anterior, las herramientas para la gestión y el análisis de datos masivos varían en gran medida en función del tipo de datos que pueden manejar, y de ello se desprende la importancia de que los bibliotecarios estudien y comprendan de inicio las características de los diferentes tipos de datos para estar en posibilidad de seleccionar las herramientas adecuadas para cada caso. Como puede verse, no existe ni el manejador universal de datos, ni el analizador universal de ellos, por lo que la selección y el uso de la herramienta adecuada para cada caso son fundamentales. Debe tenerse en mente también que algunas de ellas se encuentran en acceso abierto y otras son productos y servicios comerciales. Tanto unas como otras tienen sus ventajas y desventajas.

En lo relacionado con los manejadores de bases de datos, sin ser una lista exhaustiva, se encuentran:

Para datos estructurados, los manejadores que han sido usados tradicionalmente para estos propósitos: Oracle, MySQL, PostgreSQL, Microsoft Access, SQL Server, FileMaker, MariaDB, RDBMS, OpenOffice Base (https://www.openoffice.org/product/base.html). A su vez, las grandes plataformas en la nube ofrecen servicios basados en estos manejadores y compatibles con ellos, como por ejemplo CloudSQL de Google (https://cloud.google.com/products/databases/), AWS Relational DataBase Service (https://aws.amazon.com/es/rds/) y AWS Aurora de Amazon (https://aws.amazon.com/es/rds/aurora/), Azure Database (https://azure.microsoft.com/en-us/services/sql-database/) y Azure SQL Database (https://azure.microsoft.com/en-us/product-categories/databases/); de Microsoft, Hive de Apache (https://hive.apache.org/).

Para datos no estructurados, se encuentran los siguientes manejadores de bases de datos NoSQL, casi todos ellos en forma de servicio en la nube: AWS DynamoDB de Amazon (http://www.genbetadev.com/programacion-en-la-nube/amazon-lanza-dynamodb-una-base-de-datos-NoSQL-desarrollada-integramente-por-ellos), Cloud BigTable de Google (https://cloud.google.com/products/databases/), HBase (https://docs.microsoft.com/es-es/azure/hdinsight/hbase/apache-hbase-overview) y Cassandra de Apache

(https://cassandra.apache.org/), Azure CosmosDB de Microsoft (https://docs.microsoft.com/es-es/azure/cosmos-db/introduction), Voldemort de LinkedIn (https://www.project-voldemort.com/voldemort/), Redis (https://redis.io/) y Riak (https://riak.com/).

Por ser de especial interés para las bibliotecas, se destacan aquí algunos de los manejadores específicos para bases de datos "de texto" o documentales: CouchDB o Cluster Of Unreliable Commodity de Apache (https://docs.couchdb.org/en/stable/), MongoDB (https://www.mongodb.com/es), BaseX (http://www.basex.org/), etcétera. Como en los demás manejadores tipo NoSQL, en este tipo de herramientas los datos no se almacenan en tablas, sino que la base de datos está compuesta por "documentos" de longitudes muy variables que a su vez funcionan como objetos. Normalmente van asociadas a otras herramientas para extracción y análisis de información a partir de textos, las cuales se verán más adelante. Como sucede con muchos otros casos, algunas de ellas son productos o servicios comerciales y otras están en código abierto.

Las herramientas de normalización y mapeo de datos

> *Puedes tener datos sin información, pero no puedes tener información sin datos.*
>
> Daniel Keys Moran

Además de los manejadores de bases de datos en sí mismos –relacionales o no–, para el manejo de datos masivos se requiere previo a su explotación de un conjunto de herramientas que permitan realizar ciertos procesos preparativos sobre ellos: normalizar los datos, "mapearlos" o acomodarlos en conjuntos lógicos, depurarlos, dividirlos en segmentos manejables, etcétera. Éste es uno de los pasos previos importantes de la "curaduría de datos" para ordenarlos, desechar datos duplicados, irrelevantes, excesivos, imprecisos, etcétera, y ponerlos en condiciones idóneas para su proceso y análisis. Entre este tipo de herramientas, destacan:

Hadoop, de la fundación desarrolladora de *software* Apache. Es una herramienta en código abierto para procesamiento en general de datos. Para muchas organizaciones, es actualmente el programa informático de referencia para el manejo flexible de grandes volúmenes y variedades de datos. Permite agrupar, manejar y procesar grandes conjuntos de datos masivos estructurados, semiestructurados y no estructurados.

Al implementar esta herramienta en una cierta organización, se puede comenzar a organizar y procesar los datos de forma normalizada para su posterior explotación, ya que permite "mapear" los conjuntos de datos a estructuras lógicas manejables. Mapear datos significa básicamente homogeneizar datos representados de forma distinta en diferentes sistemas; por ejemplo, los nombres de los países del mundo: en un sistema pueden estar capturados con sus nombres completos en inglés United Arab Emirates, United Kingdom, Mexico, United States of America, etcétera. En otra aplicación con siglas: UAE, UK, MEX, USA. En otro sistema, con sus siglas de dominio de Internet: AE, UK, MX, US; en otra más con sus nombres en español: Emiratos Árabes Unidos, Reino Unido, México, Estados Unidos de América, etcétera. Al "mapear" estos conjuntos de datos, se "normalizan" a una cierta forma única para una cierta aplicación con el fin de hacerlos manejables; de otra forma serían inutilizables. En muchos procesos de datos como la integración, la migración, la sincronización, la automatización de almacenes de datos, la extracción automatizada de datos, etcétera, se requiere indefectiblemente de la depuración de los datos no utilizables y la normalización de los que se van obteniendo: he ahí la importancia del "mapeo". Este tipo de herramientas se requiere para estas tareas.

Existen otras herramientas parecidas: para "limpiar" o refinar los datos, existen herramientas como Open Refine; si los datos se vuelven demasiado voluminosos, se utilizan herramientas como MapReduce para distribuirlos en varios conjuntos lógicos más manejables. Adaptive MapReduce es la versión de la empresa IBM para este tipo de tareas de distribución y desagregación; Talend Open Studio es otra herramienta de código abierto para la integración y/o sincronización de datos (https://www.talend.com/es/resources/

introduction-talend-open-studio-data-integration/). Permite elaborar "conectores" o puntos de unión entre todos los sistemas fuente y destino de la organización por medio de modelos de integración de datos con el fin de homogeneizarlos y normalizarlos. De hecho, los conjuntos de datos ya normalizados que van a ser utilizados por varios departamentos o áreas conforman una categoría especial dentro de la gestión de datos denominada "datos maestros" de la organización, y por esa razón reciben un trato especial.

Cuando el desarrollador de un proyecto de biblioteca está construyendo su idea, después de seleccionar su manejador de base de datos debe pensar a continuación cómo va a organizar sus datos para un cierto propósito, y posteriormente cómo y con qué va a explotarlos. Éste es el momento para seleccionar todas las herramientas pertinentes para preparar, organizar, normalizar y depurar sus datos, y para ello cuenta con una serie de ellas de entre las cuales debe elegir para este objetivo. Las anteriormente mencionadas son algunos ejemplos al respecto. No debe considerarse que éste es un paso superfluo y por tanto obviarse, pues partir de datos desorganizados y excesivos puede causar demoras y fracasos.

Las herramientas para el análisis de datos masivos con el fin de extraer patrones o tendencias

El valor de los datos masivos no está en los datos masivos, está en su análisis.
GARY KING,
Universidad de Harvard

El siguiente tipo de herramientas para los datos masivos son aquellas especializadas para analizarlos y extraer de ellos información, tendencias, patrones, etcétera, de acuerdo con los objetivos de cada proyecto. Hablando en general, esta área contiene el núcleo de los datos masivos; esto es, detectar tendencias y patrones en ellos para plantear soluciones. Por su naturaleza existe una enorme variedad y subdivisiones de las aplicaciones informáticas para una infinidad de propósitos diferentes. Al igual que con los tipos anteriores,

las hay de múltiples fabricantes, plataformas, y especialidades; comerciales y de acceso abierto. La siguiente lista no pretende ser exhaustiva, sino ofrecer un panorama representativo de la muy amplia variedad y propósitos de esta categoría de herramientas:

BigQuery de Google (https://cloud.google.com/bigquery?hl=es), el cual es un servicio en la nube en forma de un almacén de datos de alta escalabilidad, que puede analizar enormes cantidades de ellos con técnicas SQL para extracción de información. Esta empresa también ofrece una herramienta denominada Google Analytics (https://analytics.google.com/analytics/academy/), la cual permite obtener información agregada acerca de las consultas a un cierto sitio web según su audiencia, comportamiento, tendencias, etcétera. Se utiliza ampliamente con fines de promoción, evaluación y seguimiento de páginas web.

Infosphere Streams (https://www.ibm.com/developerworks/library/bd-streamsintro/index.html), plataforma desarrollada por IBM. Está diseñada para descubrir patrones significativos a partir de flujos dinámicos de datos, en ventanas de minutos a horas. De la misma empresa se encuentra ThinkUp, una herramienta de código abierto para el análisis de datos que permite extraer información de Twitter, Facebook y Google+.

BigInsights es una plataforma analítica que permite a las empresas convertir complejos conjuntos de información a escala de Internet en conocimientos. Consiste en una distribución empaquetada de Apache Hadoop con un proceso de instalación muy simplificado y herramientas asociadas para el desarrollo de aplicaciones, el movimiento de datos y la gestión de *clusters* o segmentos de ellos. Un *cluster* es una plataforma de varias computadoras sincronizadas para lograr alto rendimiento.

System Applications Products High-Performance Analytic Appliance o SAP Hana (https://www.sap.com/spain/products/hana.html), junto con SAP Predictive Analytics, tiene la capacidad de integrar y analizar grandes cargas de trabajo de datos para ser analizados en tiempo real.

Oracle Big Data Appliance (https://www.oracle.com/engineered-systems/big-data-appliance/). Es una plataforma desarrollada

por la empresa Oracle de amplia cobertura de funciones adquirir, organizar y analizar grandes cargas de datos de diversas fuentes a gran velocidad.

Azure HDInsight (https://docs.microsoft.com/en-us/azure/hdinsight/). Es un servicio en la nube de Microsoft basado en Apache Hadoop, el cual permite interactuar con muchos programas de esa plataforma como Apache Spark –para análisis de datos en tiempo real–, Apache Hive, Apache Kafka, Apache HBase con el fin de procesar y analizar múltiples datos.

Splunk (https://www.softtek.com/es/tecnologias/splunk). Se especializa típicamente en el aprovechamiento de datos provenientes de máquinas de varias fuentes diferentes, como sitios web, aplicaciones, Internet de las Cosas y sensores.

Statistical Package for the Social Sciences (SPSS) (https://www.ibm.com/products/spss-statistics). Es una de las herramientas de *software* más clásica y antigua para el análisis estadístico por computadora. Este sistema fue creado en 1968 por científicos de la Universidad de Chicago para grandes computadores o *mainframes* y se convirtió en pionero de este tipo de herramientas. En 1975 se creó la empresa SPSS Inc. al efecto, la cual fue adquirida por IBM en 2009. Actualmente se maneja la versión 26 de esta herramienta para múltiples plataformas. A pesar de ser el más antiguo, este *software* sigue siendo uno de los más utilizados para análisis estadístico debido a la enorme cantidad de pruebas y análisis que puede realizar; cuenta además con interfaces a otras herramientas como SAS, Matlab, Statistica, etcétera, así como a rutinas en lenguaje R. Existe una versión libre de esta herramienta bastante aceptable denominada PSPPire (https://www.softpedia.com/get/Office-tools/Other-Office-Tools/PSPP.shtml).

Statwing (https://www.statwing.com/). A semejanza del anterior, es una herramienta de *software* utilizada para hacer análisis estadístico clásico en conjuntos de datos para extraer de ellos los elementos típicos de esta ciencia: parámetros, regresiones, correlaciones, etcétera.

LibInsight de la empresa Springshare (https://springshare.com/libinsight/). Es una herramienta de *software* utilizada para colecta

y análisis de estadísticas específicamente en bibliotecas. Almacena todos los datos de la biblioteca en una única plataforma y utiliza técnicas de análisis de conjuntos de datos cruzados para optimizar la toma de decisiones basada en ellos.

DisplayR (https://www.displayr.com/migration/). Es una herramienta de *software* de propósito general que incluye módulos para análisis estadístico, aprendizaje de máquina, análisis de datos y su visualización con una interfaz basada en lenguaje R.

"R" (https://www.r-project.org/) es un lenguaje de programación desarrollado por la Fundación para la Computación Estadística (Foundation for Statistical Computing). Es muy utilizado para estadística por medio de computadoras y para la minería de datos, así como para visualizaciones con gráficas. Con este lenguaje pueden desarrollarse rápidamente programas y aplicaciones para análisis de datos.

El lenguaje SAS fue diseñado para operar principalmente sobre tablas de datos: tiene variadas opciones para su lectura, transformación, combinación, resumen, y despliegue, así como para múltiples análisis estadísticos de los datos. Sus principales módulos a este respecto son: a) SAS/STAT, con procedimientos para realizar los análisis estadísticos típicos (regresiones, correlaciones, etcétera); b) SAS/ETS para el análisis estadístico de series temporales; c) SAS/IML para implementar lenguajes alternativos similares a Octave, Matlab[32] o R; d) SAS/OR para la resolución de problemas del tipo de Investigación de Operaciones, y e) SAS/GRAPH para generar gráficos. Cuenta además con módulos adicionales para otras tareas, como SAS Enterprise Guide para capacitación, SAS Data Integration Studio para mapeo de datos y SAS Enterprise Miner para minería de datos (https://www.sas.com/en_us/solutions/analytics.html#).

Finalmente, cabe mencionar que el clásico y antiguo Lenguaje de Consulta Estructurado o Structured Query Language (SQL), a pesar de su simplicidad, o tal vez debido a ella, sigue siendo ampliamente utilizado para extraer y analizar datos en Sistemas de

32 Matlab es un lenguaje de programación especializado para desarrollar proyectos que conllevan fuerte cálculo numérico. GNU Octave es similar pero de acceso abierto.

Gestión de Bases de Datos Relacionales (DBMS). A la fecha, se recomienda que toda persona que desee iniciar el aprendizaje de herramientas para gestión o manejo de datos en bases de datos comience aprendiendo el lenguaje SQL.

Las herramientas para el análisis de textos en datos masivos con el fin de extraer información o tendencias

El objetivo es convertir los datos en información y la información en perspicacia.
Carly Fiorina,
ex presidenta de Hewlett Packard

Como ha sido mencionado, el análisis de textos o minería de textos es uno de los rubros de particular interés en el ambiente de las bibliotecas debido a las múltiples aplicaciones que pueden dársele a este tipo de datos en este campo: identificación de textos, extracción de elementos de ellos, categorización y/o taxonomía de textos, extracción de conceptos, entidades, relaciones, eventos y otros metadatos; traducción, etcétera. Existen numerosos productos desarrollados expresamente para el análisis de información proveniente de textos en muchas de sus variantes, por lo cual son de particular interés para el ambiente de las bibliotecas. Nuevamente, sin ser una lista exhaustiva, destacan entre ellos:

Los programas de la fundación de *software* Apache. Esta organización ha desarrollado varios programas para el propósito de análisis de textos y extracción de información valiosa de ellos. El núcleo de la arquitectura lógica de estos *softwares* consiste en el concepto de documentos que contienen campos hipotéticos. Esto les permite ser independientes del formato del archivo informático: pdf, html, txt, doc, etcétera. Entre ellos, los principales son: Apache Lucene, el cual es una librería de *softwares* para texto completo basada en el lenguaje Java, la cual proporciona una plataforma de búsqueda e indexación de elementos dentro del texto (https://lucene.apache.org/). Se encuentra también Apache

OpenNLP, una herramienta de acceso abierto basada en técnicas de aprendizaje de máquina para procesamiento de lenguaje natural (https://opennlp.apache.org/). Se complementan con Apache UIMA, *software* de gestión de información no estructurada para capturar texto simple e identificar entidades internas como personas, lugares, organizaciones o relaciones, como *se localiza en* o *está asociado con* (https://uima.apache.org/).

Google Cloud Natural Language API. Utiliza el almacenamiento y proceso en la nube de Google con técnicas de aprendizaje de máquina para encontrar la estructura y el significado de textos no estructurados. Extrae información acerca de personas, lugares o eventos, opiniones en las redes sociales, etcétera (https://cloud.google.com/natural-language/).

Textalytics, un *software* desarrollado por Daedalus. Extrae con facilidad buena parte del contenido de todo tipo de documentos, especialmente en redes sociales (https://www.programmableweb.com/api/textalytics).

IBM SPSS Text Analytics. Esta aplicación permite capturar datos de encuestas, extraer conceptos clave, proponer resultados y categorizar respuestas (https://www.ibm.com/support/pages/downloading-ibm-spss-text-analytics-surveys-401). Se complementa con IBM Watson Natural Language Understanding, el cual es un servicio en la nube que utiliza técnicas de aprendizaje de máquina para extraer metadatos de los textos, tales como entidades-relación, sintaxis, palabras clave, categorías y opiniones (https://www.ibm.com/cloud/watson-natural-language-understanding).

Microsoft Azure Text Analytics API. Este *software* descubre ideas en textos no estructurados utilizando el procesamiento del lenguaje natural, y no requiere de experiencia en sistemas de aprendizaje de máquina. Identifica y extrae frases y entidades clave como personas, lugares, organizaciones, opiniones, entre otros, con el propósito de comprender temas y tendencias comunes en una amplia variedad de idiomas (https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/).

General Architecture for Text Engineering (GATE). Es una herramienta de acceso abierto basado en aprendizaje de máquina

para encontrar estructura y significado de textos no estructurados en diversos formatos: html, pdf, doc, text, etcétera. Detecta y extrae información acerca de entidades como personas, lugares, organizaciones, opiniones o eventos (https://gate.ac.uk/).

DiscoverText es un conjunto de herramientas de *software* simples y avanzadas en la nube que permite evaluar con rapidez y precisión grandes cantidades de textos no estructurados, así como los metadatos asociados provenientes de encuestas, chats, correo electrónico, comentarios públicos, Twitter, fuentes RSS y otras formas de datos de texto (https://discovertext.com/).

Semantria, Semantria API y Semantria for Excel. Similar al anterior, es un conjunto de herramientas de *software* en la nube para evaluar grandes cantidades de textos no estructurados, provenientes de encuestas, chats, correo electrónico, comentarios públicos, Twitter y otros textos similares. Cuenta con una versión agregable a Excel para análisis de datos contenidos en ese *software* (https://www.lexalytics.com/semantria).

Lexalytics Salience. Herramienta para extracción de información de contenidos y metadatos de textos en lenguaje natural de redes sociales, en especial Twitter (https://www.lexalytics.com/salience/server).

Provalis Research Text Analytics. Ensamble de las herramientas de *software* Prosuite, QDA Miner y Wordstat para minería de textos, captura de datos desde diversas fuentes, análisis de contenido, extracción de entidades como palabras clave, frases y temas; clasificación, identificación de patrones y tendencias, visualización, mapas, etcétera (https://provalisresearch.com/).

SAS Text Miner. Herramienta de *software* basada en el aprendizaje de máquina para extracción de temas o las entidades y relaciones de palabras clave a partir de datos no estructurados, especialmente textos, y a partir de ellos elaborar modelos predictivos (https://www.sas.com/en_us/software/text-miner.html).

Text2Data. Es un conjunto de herramientas de *software* y servicios en la nube para extraer y clasificar entidades clave a partir de textos; hacer análisis de contenidos y de opiniones. Al igual que Semantria, cuenta con una extensión para ser utilizada en combinación con Excel y otra para combinar con Google Sheet (https://text2data.com/).

La lista de este tipo de herramientas de *software* continúa con un sinnúmero de fabricantes y productos que realizan alguna o algunas de las tareas relativas a textos ya enunciadas: Rossette Text Analytics, Stratifyd, Luminoso, Bitext, NetOwl, Natural Language Toolkit, Aylien, Expert System, Smartlogic, Ascribe, Datumbox, Indico, RapidMiner Text Mining Extension, Keatext, Pingar, TextualETL, KH Coder, QDA Miner Lite, TAMS, Visual Text, Pentaho, etcétera.

Como puede observarse en la lista anterior, la cantidad de opciones es realmente amplia en este rubro. Como en otros casos, y dependiendo de si el proyecto de datos es de tipo textual o no, el responsable de la biblioteca debe comenzar estudiando el problema a resolver y las características más específicas de los datos a analizar para poder estar en posibilidad de seleccionar el conjunto de herramientas idóneas que utilizará para las tareas de extracción y análisis de datos, así como aquellas para visualización e interpretación de los mismos. Muchas de las herramientas mencionadas en las listas anteriores son de acceso abierto, por lo cual se vuelven excelentes opciones para que las bibliotecas comiencen a incursionar en el manejo de los datos masivos, la Inteligencia Artificial, el aprendizaje de máquina, etcétera, sin necesidad de efectuar grandes desembolsos de inicio.

Las herramientas de visualización, interpretación o presentación de resultados

> *[…] las estadísticas, como los pasteles, son buenas si se sabe quién las hizo y se está seguro de los ingredientes.*
>
> Lawrence Lowell,
> Decano de la Universidad
> de Harvard, 1909

El siguiente grupo de herramientas de *software* presentado en la lista consiste en la presentación y visualización de los datos; éste es un componente de gran importancia. De nada sirve colectar muchos datos si después estos no pueden ser interpretados

adecuadamente para la toma de decisiones. Los datos en sí mismos tienen poco valor; el conocimiento que se puede extraer e interpretar de ellos es el verdadero insumo valioso.

Grandes conjuntos de cifras en tablas y columnas o en interminables listados pueden representar muy poco en términos de información relevante, de ahí la importancia de la presentación y la visualización de los datos. Esto es particularmente importante en el ambiente de los datos masivos, donde pueden colectarse millones de datos de muy diversas dimensiones y significados. Poder abstraer todo eso en formas sencillas de visualización es en cierta forma técnica y en otra arte, ya que debe combinar el diseño sencillo con la lógica y los números, algo que no es fácil de conjuntar, pero que sin duda es un factor primordial de éxito en el manejo de datos masivos. De hecho, ya existe toda una subdisciplina alrededor de ello: la "alfabetización visual", también llamada "habilidades visuales" (*visual literacy)*. Orland-Barak y Mazkit (2017, 11) la definen como:

> […] la capacidad de interpretar, negociar y dar sentido a información presentada en forma de imágenes, ampliando el significado de 'alfabetización', que comúnmente significa la interpretación de un texto escrito o impreso. La alfabetización visual se basa en la idea de que las imágenes también pueden ser 'leídas' ya que su significado puede entenderse a través de un proceso de lectura". Al respecto de la disciplina existe un "Journal of Visual Literacy" y una "International Visual Literacy Association".

Esta disciplina ha evolucionado enormemente en los últimos años, precisamente por la necesidad de poder expresar información adecuada eficientemente en forma visual. Esto cobra especial importancia en estos tiempos plagados de información sesgada, equivocada, engañosa, infundada, tendenciosa, pobremente interpretada, mal intencionada, incompleta o de plano falsa. Para mantener la percepción de credibilidad, es indispensable que la información desplegada en las bibliotecas se mantenga con alta calidad, y por eso las habilidades visuales cobran especial importancia en estas organizaciones.

Edward Tufte, pionero en la materia y uno de los autores más reconocidos al respecto actualmente, recomienda el uso de ilustraciones ricas en datos a la vez que sencillas y resume el propósito de la visualización de datos de la siguiente forma: "[…] la tarea del diseñador no es la complicación de lo simple, sino dar acceso visual a lo sutil y a lo difícil; es decir, la revelación de lo complejo[…] la excelencia gráfica es aquella que da al lector el mayor número de ideas en el menor tiempo con la menor cantidad de tinta en el menor espacio" (Tufte 2001, 16-18). Este autor ha acuñado además algunos términos de amplio uso en la disciplina; por ejemplo, el de 'tablas-basura' o *junkcharts* para referirse a la inclusión de elementos inútiles, poco informativos u obstaculizadores en la información de reportes; también el concepto de la 'relación tinta-datos' o *data-ink ratio*, la cual se refiere al ornamento excesivo en las representaciones gráficas de información cuantitativa. Él estableció que "[…] en algunos casos, la decoración puede ayudar a hacer editoriales sobre el núcleo de la gráfica. Pero es un grave error abusar de las medidas de los datos para hacer un comentario editorial o encajar un esquema decorativo" (Tufte 2001, 59).

Esto no es nuevo: en su libro de 1954, *Cómo mentir con estadísticas*, Darrell Huff resaltaba ya la importancia de que el lector común aprenda a interpretar de forma correcta tanto lo que aparentan decir, como lo que realmente dicen ciertas estadísticas publicadas para poder distinguir y descartar aquellas incompletas o manipuladas. El autor resaltaba desde entonces de forma clara las técnicas frecuentes de abuso de las estadísticas y sus métodos de visualización –en especial en los medios de comunicación masiva– con fines de manipular, sesgar, distorsionar, minimizar o sensacionalizar ciertos datos. Su texto se volvió un clásico y sigue vigente después de seis décadas, y continúa siendo lectura obligada en los cursos actuales de estadística (Huff 1954).

Tufte retoma esos conceptos aplicados ahora a los recursos contemporáneos: la red mundial, Power Point, los datos masivos, etcétera. Básicamente, ambos autores subrayan por un lado el hecho de que debemos revisar siempre con mente crítica las presentaciones y conclusiones de otros que se extraigan de los datos o

visualizaciones, pues pueden estar afectadas por sesgos, errores u omisiones del emisor, así como intenciones sensacionalistas, ya sea deliberadamente o por error. Y por el otro lado, ellos resaltan la importancia de emitir resultados, tablas, gráficas, etcétera, con la mayor exactitud y calidad. Así, la adecuada selección de herramientas al respecto es fundamental. Como sucede con toda herramienta, si no se sabe cómo y cuándo usarla adecuadamente, se vuelve inútil, y en ciertos casos, hasta peligrosa. Por todo lo anterior, el conocimiento previo, la capacitación y dominio de este tipo de herramientas se vuelve fundamental en las bibliotecas.

El primer elemento que influye en ello es la adecuada selección de datos a visualizar. Ello de inicio parece una verdad evidente, pero en la práctica no lo es. Desde hace más de medio siglo hay un adagio en la ciencia de la computación que afirma: "basura entra, basura sale".[33] Éste se refiere a que en este campo, si los datos de entrada a un sistema informático son defectuosos, sesgados o sin sentido, el resultado siempre será algo defectuoso con poca o nula utilidad. A pesar del paso del tiempo, el principio sigue siendo totalmente válido. La selección de datos a visualizar debe ser informativa, constructiva, interesante y retadora, pero además debe parecer atractiva. Las bibliotecas han compilado y presentado sus datos de uso desde hace más de una centuria, lo cual es conveniente, pero el problema es que en la inmensa mayoría de las veces esos datos no han cambiado, siguen siendo los mismos. Típicamente informan de cuántos usuarios atendieron en un cierto periodo –mes, semestre año–; cuántas consultas respondieron, cuántos recursos documentales poseen, cuántos libros o revistas fueron consultados, y otros datos semejantes. Eso está bien, siguen siendo necesarios. El problema es que para los tiempos actuales, ofrecen datos demasiado básicos y elementales. Simplemente con la gran cantidad de datos que los sistemas de automatización de

33 Se le conoce como "principio GIGO", por sus siglas en inglés: *Garbage In, Garbage Out*. Fue consignado por primera vez por el Free Online Dictionary of Computing (FOLDOC). Se le atribuye a Wilf Hey en 1965.

bibliotecas captan en primera instancia, las estadísticas pueden ser más ricas y variadas, sin entrar siquiera a los datos masivos.

Phetteplace (2012, 95) presenta una reseña de cómo plantear esto en la biblioteca. Él sugiere que, partiendo de una lluvia de ideas, deben plantearse nuevas estrategias de colecta de datos. Tomando como base las colecciones, los usuarios y las interacciones entre ambos, pueden plantearse nuevas preguntas acerca de los reportes: ¿Cuáles cruces entre datos dispares pueden hacerse de una manera nueva, útil y atractiva? ¿Cuánta información adicional puede extraerse de los catálogos y de las obras en la biblioteca? ¿Pueden hacerse análisis de datos de sitios web de interés para la biblioteca? Si no pueden encontrarse datos utilizables en la actual estrategia de recopilación, entonces vale la pena reconsiderar por qué se están acumulando hechos abstractos.

Este autor agrega que una vez que se han compilado los datos, una buena comprensión de los tipos básicos de visualizaciones y despliegue permite seleccionar los más apropiados: los típicos gráficos lineales para datos temporales, mapas para datos geográficos, gráficos de barras o de pastel para comparaciones simples, tablas para conjuntos de datos, etcétera. Pero hoy en día hay muchas más opciones, cada una de ellas adecuada a formas particulares de datos: gráficas de área o *treemaps*, mapas mentales, diagramas de Venn, gráficos de burbuja, grafos para datos abiertos enlazados, por mencionar algunas. Y, por supuesto, existen diversas herramientas de *software* para elaborar cada una de esas visualizaciones, por lo que es necesario extender el conocimiento hacia esas herramientas para poder hacer una selección adecuada de la visualización para cada tipo de datos. He ahí la gran importancia de ese conocimiento en las bibliotecas. Por un lado, poder establecer cuáles nuevos conjuntos de datos es conveniente colectar, y por otro cómo presentarlos de forma atractiva e interesante a usuarios y personal de la biblioteca. La mayoría de los autores –al igual que Phetteplace y Tufte– coinciden en que la visualización de datos debe estar concebida para ilustrar y no para ofuscar. Si un cierto diseño añade más complejidad, va en contra del propósito mismo de la representación. A medida que se

añadan más capas y se revelen dimensiones adicionales, la estructura y el mensaje de los datos debiesen ser más fáciles de interpretar. Ben Shneiderman resumió todo esto espléndidamente: "[…] el propósito de la visualización es ofrecer reflexión, no imágenes".

Dentro de las principales herramientas informáticas en este aspecto se distinguen Tableau Public, Many Eyes de IBM y Google Data Studio, las cuales son herramientas de propósito general para elaborar de forma sencilla múltiples tipos de visualizaciones a partir de datos como tablas, gráficas, mapas, etcétera. Son fáciles de generar y utilizar como los gráficos de Excel, pero con muchas mayores capacidades.

Las herramientas para la Inteligencia Artificial

> *Nadie lo expresa así, pero creo que la Inteligencia Artificial es casi una disciplina de las humanidades. En realidad consiste en un intento de entender la inteligencia y la cognición humanas.*
>
> Sebastián Thrun

En términos generales, muchas de las categorías anteriores son herramientas de *software* que contienen buena parte de Inteligencia Artificial: el análisis de textos para la creación de índices y catálogos; la traducción; los sistemas expertos, etcétera. Todos ellos contienen muchos de sus componentes, por lo que es imposible hablar de herramientas de datos masivos "con" IA o "sin" IA; prácticamente todas contienen algún elemento de ella en mayor o menor grado.

Por este motivo, muchas bibliotecas ya están desarrollando o utilizando algún servicio o producto que incorpora ese componente, aunque con frecuencia esto pasa desapercibido dado que la herramienta o servicio no contiene el nombre "Inteligencia Artificial" explícitamente en ellos. Si están utilizando un catálogo expandido más allá de un simple OPAC, colectando datos de usuarios en redes sociales, optimizando el sistema de consulta o utilizando

algún sistema experto para una tarea, muy probablemente están usando algún componente de IA en los *software*s que los operan. Existen muchos productos y servicios en la biblioteca que ya contienen en parte esta tecnología; en suma, no necesariamente todo el *software* está basado en IA, pero sí alguna de sus partes. Por esta razón, no se ofrecen en el mercado como herramientas de IA, pero en realidad lo son. Por ello es necesario que el bibliotecario esté al tanto de estos desarrollos, para estar en posibilidad de reconocer esas partes de esta tecnología y utilizarlas en su favor. Desechando la estereotipada imagen de un robot recorriendo la biblioteca, en la vida real ya existen innumerables aplicaciones de la Inteligencia Artificial en una amplia diversidad de usos dentro de la biblioteca; muchos de ellos utilizados cotidianamente en numerosas bibliotecas a nivel mundial.

No obstante, en efecto existen algunas herramientas que se ofrecen expresamente asociadas con el nombre de "Inteligencia Artificial", por lo que conviene hacer una breve revisión de ellas. Estas herramientas de *software* pueden encontrarse tanto en forma de "paquetes", es decir, conjuntos amplios de aplicaciones y servicios basados en IA contenidos en un solo sistema ofrecidos por un proveedor, como también en forma de "aplicaciones puntuales" de IA que sirven para resolver un solo propósito específico. Al igual que con las categorías anteriores, pueden encontrarse en aplicaciones de acceso abierto y como productos comerciales. Por supuesto, existen en todas las posibles variantes de la IA: sistemas expertos, análisis de textos, análisis de voz, asistentes robóticos, etcétera. Muchas bibliotecas a nivel mundial han construido aplicaciones específicas para alguno de sus servicios basadas en IA, las cuales sirven de ejemplo para pensar en desarrollos específicos de este tipo.

Entre los "paquetes" comerciales, puede citarse como ejemplo Savannah (https://www.orangeboyinc.com/benefits-and-features/), una plataforma inteligente especializada en bibliotecas que ofrece en un solo conjunto almacén de datos, segmentación de usuarios, comunicaciones con ellos en forma distribuida, informes de rendimiento, retroalimentación de NPS y capacidades de mapeo

de GIS.[34] Muchas bibliotecas combinan sus bases de datos de usuarios con cartografía GIS para hacer mapas con sus datos geográficos y enviarles información dirigida y selectiva de acuerdo con ello.

Semejante al anterior, se encuentra Patron Point, plataforma de automatización con funciones para ofrecer a las bibliotecas la mejora en la comunicación digital hacia su comunidad. Básicamente, es un servicio de Diseminación Selectiva de Información con Inteligencia Artificial agregada. Combina los datos de usuarios con los sistemas de información de la biblioteca y con servicios de terceros: catálogos, recursos electrónicos, bases de datos, sistemas de registro de eventos y muchos otros elementos; segmenta con detalle a los usuarios para automatizar las comunicaciones hacia ellos de forma mucho más dirigida y personalizada.

Ciertas aplicaciones "puntuales" están construidas para resolver algún problema específico con ayuda de las técnicas de IA. Entre éstas, podemos citar como ejemplo representativo a Collection HQ (https://www.collectionhq.com/), herramienta desarrollada especialmente para bibliotecas públicas con el fin de analizar los datos de uso de sus colecciones para mejorar la utilización y rendimiento de ellas. Tiene componentes de IA que ayudan a optimizar la adquisición de materiales para las colecciones, su uso, administración y promoción.

Otro ejemplo representativo de estas herramientas puntuales de IA es Gale Analytics (https://www.gale.com/databases/gale-analytics), la cual integra componentes demográficos para conocer con detalle a la comunidad de una cierta biblioteca, optimizar el uso y promoción de sus materiales, y extraer información del Sistema Integrado de Gestión Bibliotecaria (ILS) para agregar valor

---

34 La encuesta Net Promoter Score (NPS) consiste en una sola pregunta que mide la probabilidad de que un cliente o usuario recomiende una institución o servicio a otros. Es un indicador de la experiencia del usuario, su satisfacción y la eventual futura lealtad. Un Sistema de Información Geográfica (GIS) es una técnica para la recopilación, gestión y análisis de datos acomodados geográficamente. Analiza la ubicación espacial y organiza capas de información en visualizaciones usando mapas y escenas 3D.

a los datos existentes, optimizando su alcance, promoción y uso. Permite captar nuevos usuarios, crear e impulsar nuevos servicios bibliotecarios y asignar recursos de manera más eficiente.

Algunas de estas herramientas han sido creadas por bibliotecas. Entre los sistemas expertos desarrollados por ellas, encontramos Plexus, desarrollado por la Biblioteca Británica y la Universidad de Londres para bibliotecas públicas. Consiste en una herramienta que realiza ciertas tareas de referencia típicas del bibliotecario de consulta: obtiene la descripción de una necesidad de información del usuario y, de ser necesario, la complementa infiriendo conceptos adicionales o pidiendo al usuario que responda a algunas preguntas aclaratorias y delimitadoras. El sistema elabora entonces una estrategia de búsqueda que puede aplicarse a los acervos y/o bases de datos de la biblioteca o a otras fuentes de referencia afines.

Un buen número de las aplicaciones y herramientas en este rubro desarrolladas por bibliotecas giran alrededor del Análisis del Aprendizaje (*Learning Analytics*), también denominado Analítica del Aprendizaje. Si bien este tema es de interés en general de las instituciones académicas por su cercanía con las bibliotecas, con frecuencia es desarrollado dentro de ellas o asociado a las mismas. 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". Esta definición procede de la primera Conferencia Internacional sobre Análisis del Aprendizaje y el Conocimiento del 2011 (First International Conference… 2011). Otras diez conferencias al respecto se han realizado desde entonces y dado que el tema ha despertado gran interés en años recientes, se creó una sociedad expresamente para la investigación del Análisis del Aprendizaje. El New Media Consortium (2013, 5) lo define como:

> […] El Análisis del Aprendizaje es el campo asociado con la determinación de las tendencias y patrones de los datos masivos en la educación, o grandes conjuntos de datos relativos a los estudiantes,

para avanzar aún más en el sistema personalizado de apoyo a la enseñanza superior.

Este tipo específico de análisis de datos se utiliza para:

- Predicción; por ejemplo, para identificar a los estudiantes "en riesgo" en lo tocante a deserción o fracaso escolar, o a la inversa, detectar estudiantes con habilidades o potencial arriba del promedio.
- Personalización y adaptación para proporcionar a los estudiantes métodos, canales, herramientas de aprendizaje y hasta materiales de evaluación personalizados.
- Asesoría para proporcionar a los docentes información pertinente para tutelar y apoyar a los estudiantes.
- Retroalimentación para evaluar el interés y satisfacción de cursos, materiales educativos, servicios de información, técnicas y modalidades de instrucción, etcétera.
- Visualización de la información, generalmente en forma de tableros de aprendizaje (*learning dashboards)* que proporcionan datos generales del aprendizaje a través de herramientas de visualización de datos.

Existe una variante adicional del Análisis del Aprendizaje, el "Análisis Académico" (*Academic Analysis*), el cual también utiliza la ciencia de los datos y la IA. Baepler y Murdoch (2010, 3) la definen *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".

Un ejemplo aplicado de estos conceptos puede verse en la Biblioteca de la Universidad de Washburn en Kansas. Young (2017) menciona que en ella se planteó un proyecto para establecer hasta dónde había una correlación entre los estudiantes que consumen más material de biblioteca y su éxito académico. Durante varios años, obtuvieron datos minuciosos acerca de cómo el uso de la biblioteca se compara con otras métricas del éxito académico. Como

resultado, se demostró cuantitativamente que, en efecto, un mayor uso de la biblioteca y sus recursos incidían significativamente en ese éxito. A partir de ello, la universidad realizó ciertos cambios, como trasladar el departamento de tutoría 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. A partir de estas modificaciones en la biblioteca, la retención escolar aumentó 12 por ciento. Derivados de esta experiencia, muchos encargados de bibliotecas universitarias, como por ejemplo las del estado de Georgia, en Estados Unidos, han desarrollado otros proyectos de Análisis del Aprendizaje para medir hasta dónde los estudiantes que utilizan más recursos de información de la biblioteca tienden a lograr mayor éxito académico. Dado que los resultados son típicamente favorables pero a la vez cuantitativos, inciden significativamente en la presencia, importancia y presupuesto de las bibliotecas dentro de sus universidades. Siempre se había intuido y afirmado que los alumnos que más utilizan la biblioteca son más exitosos académicamente, pero hasta el uso de los datos masivos y la IA en la biblioteca no se había podido demostrar cuantitativamente.

La Universidad Nacional Autónoma de México ya ha iniciado algunos proyectos alrededor de los principios y técnicas del Análisis Académico precisamente con el propósito de estudiar temas sumamente difíciles de abordar con técnicas "tradicionales" de datos para mejorar el aprendizaje y la eficiencia terminal de sus estudiantes. Por ejemplo, el desarrollo del proyecto AppUNAM, el cual pretende recabar por medio de una aplicación móvil datos relevantes provenientes de esa comunidad. Este proyecto forma parte de un esfuerzo global de varias universidades denominado "Flujo de trabajo de retención estudiantil" (*Student Retention Workflow*) (Salazar 2020, 96). También en la UNAM, en la Coordinación de Universidad Abierta, Innovación Educativa y Educación a Distancia (CUAIEED), ya se creó en su estructura orgánica una Coordinación de Inteligencia Artificial, Aprendizaje de Máquina y Analítica del Aprendizaje, en la cual ya se han iniciado algunos proyectos sobre esta temática. En América Latina, la *Revista Iberoamericana*

de Educación/Educaçao dedicó todo un número al tema del Análisis del Aprendizaje bajo un enfoque particular de esta región (*Revista Iberoamericana*… 2019).

La nube de Google ofrece acceso a súper procesadores llamados Tensor Process Unit (TPU) utilizados en su motor de búsqueda, traductor, identificador de fotos, Google Assistant y Gmail con el propósito de que sus clientes puedan desarrollar aplicaciones de IA en computadores de capacidad extendida. Amazon Web Services y Microsoft Azure también proporcionan acceso a súper procesadores llamados Unidad de Procesamiento Gráfico (GPU) para desarrollar este tipo de sistemas.

En Inteligencia Artificial, existen además numerosas herramientas de *software* que no están construidas como una solución para "instalar y usar", sino como un conjunto de lenguajes, rutinas, y librerías de programación que realizan propósitos específicos y cuyas partes se van armando como un modelo a escala para obtener la solución de una cierta necesidad específica. Este tipo de herramientas se conocen como "marcos de IA" (AI Frameworks). En este caso, se encuentran como ejemplos representativos:

- Python (https://www.python.org/). Lenguaje y librerías con numerosos elementos de IA usado para incontables aplicaciones; con él se han construido muchos de los repositorios académicos y de datos que vemos hoy en día; puede verse un buen tutorial de Python en español en https://www.learnpython.org/es/.
- Amazon Machine Learning (https://docs.aws.amazon.com/machine-learning/latest/dg/what-is-amazon-machine-learning.html). Amplia plataforma para aprendizaje automatizado utilizada por miles de instituciones de todo el mundo, la cual contiene una gran variedad de soluciones de IA listas para ensamblar y construir interesantes aplicaciones a partir de ellas.
- Microsoft Cognitive Toolkit (https://docs.microsoft.com/en-us/cognitive-toolkit/). Plataforma semejante a la anterior,

con la ventaja de ser una herramienta de acceso abierto con numerosas rutinas y elementos para IA.

- Accord (http://accord-framework.net/). Plataforma para el aprendizaje de máquina con librerías adicionales de procesamiento de audio e imagen. Es un marco de IA para la construcción de aplicaciones de visualización, procesamiento de estadísticas, análisis predictivo, etcétera.

Lo anterior es sólo una pequeña muestra representativa de todo lo que existe disponible sobre el tema de la IA aplicado a bibliotecas. Como puede verse, en realidad las aplicaciones de IA son muy variadas y van mucho más allá del robot asistente en la biblioteca. Como otros aspectos de las TIC, llevan ya buen tiempo en las bibliotecas apoyando en muy diversos quehaceres, y lo mejor de todo es que las posibilidades por delante son todavía más amplias. Muchas de ellas pasan ignoradas por los bibliotecarios o son utilizadas por ellos sin la consciencia de que en realidad son aplicaciones con IA. Al respecto, Wheatley y Hervieux (2019) realizaron un estudio para medir la presencia de los desarrollos de IA en universidades de Estados Unidos y Canadá, en especial dentro de sus bibliotecas, y llegaron a la conclusión de que los bibliotecarios de esas instituciones no estaban tan conscientes del tema ni tan atraídos al mismo comparado con otros campos de aplicación tecnológica, contrario a lo que se pudiera pensar en esas ubicaciones geográficas.

Ello tal vez sea derivado de la natural desconfianza que el tema de la Inteligencia Artificial ha despertado desde siempre en muchos profesionales –de todas las áreas– como amenaza a sus labores, lo cual causa que no se profundice más en él. Esto no es nuevo: muchos de los aspectos de TIC que por décadas han sido totalmente comunes en las bibliotecas fueron vistos en ellas en su momento y por largo tiempo con escepticismo y desconfianza, como la fabricación de tarjetas catalográficas, los catálogos automatizados o los índices electrónicos, por citar unos ejemplos. Esto es natural, pero lo cierto es que a pesar de ello y con el tiempo se volvieron parte integral de las herramientas de las bibliotecas,

se convirtieron en algo de lo más cotidiano, y al final sus encargados aceptaron su utilidad y beneficio, y las aprovecharon como lo que son: herramientas tecnológicas para el apoyo al quehacer de los bibliotecarios. Lo mismo sucede con aquellas de la IA: más allá de mitos y exageraciones, son herramientas útiles para auxilio de las tareas en la biblioteca. Ya desde 1995 se estableció:

> Ningún programa informático puede emular completamente el conocimiento de un bibliotecario obtenido a través de una sólida comprensión teórica de los procesos de información y la experiencia del mundo real. Pero los algoritmos construidos sobre años de experiencia pueden responder a muchas preguntas, liberando así al personal de la biblioteca para hacer tareas avanzadas únicas y laboriosas (Expert Systems… 2014, s.p.).

Como puede concluirse, esto sigue totalmente vigente después de cinco lustros.

La especialidad del "análisis de datos" (*data analytics*)

> *Hoy y en el futuro, las empresas tendrán más datos de los que puedan imaginar y dispondrán de medios para capturarlos y gestionarlos. Pero más que nunca es necesaria la capacidad de analizar los datos adecuados en el momento oportuno para tomar decisiones y emprender acciones.*
>
> Judith Hurwits,
> "The Big Data Paradox"

Como ha podido verse a lo largo de este capítulo, existen innumerables herramientas de *software* utilizadas en el campo de los datos, en especial los masivos; las hay para todo tipo de propósitos, organizaciones, proyectos y para todos los presupuestos. Empero, en tiempos recientes se ha estado creando un énfasis especial en

el campo y las herramientas para el "análisis de datos", también llamado "analítica de datos" (*data analytics*).

Ya se ha hecho una reseña panorámica de las herramientas para este propósito en apartados anteriores. Pero es necesario subrayar que ésta es toda una especialidad. Hablando en general, ésta contiene el núcleo de la ciencia de los datos; esto es, detectar tendencias y patrones para plantear soluciones; pero más puntualmente, se refiere en especial a la tarea de identificar cuáles variables de la organización pueden ser relacionadas con ciertos datos y establecer así correlaciones para el planteamiento de preguntas y la eventual obtención de respuestas a través de técnicas específicas. Estas técnicas forman ciertamente el núcleo del análisis de datos como especialidad. Varios autores han enfatizado la importancia del método de diseño y procedimiento para analizar datos. Básicamente, los estudios establecen que diferentes personas que analizan el mismo evento, en las mismas condiciones, realizando procedimientos diferentes, pueden encontrar el mismo valor respecto a un parámetro estadístico, pero la evaluación sobre la importancia de los datos obtenidos y las acciones derivadas de ello pueden ser diferentes dependiendo del procedimiento de análisis utilizado. A esto se le denomina la "paradoja de los datos" y resalta la importancia de su correcto análisis (Berdondini 2019).

El análisis de datos debe ser estudiado con cautela y comprendido en toda su amplitud. En tiempos recientes, puede verse un sinnúmero de páginas web, textos, artículos, herramientas de *software*, tutoriales, cursos, posgrados, asesorías, productos, certificaciones, etcétera, ofrecidos por una miríada de organizaciones y empresas a propósito del tema del "análisis de datos". Obviamente el núcleo de toda la ciencia y la gestión de los datos consiste en poder extraer información valiosa y útil de los conjuntos de ellos, pues de nada sirve colectarlos y procesarlos si no se puede obtener un beneficio tangible para la organización, y esto se vuelve realidad cuando se analizan. Ya se ha hablado del "Valor" como la quinta característica esencial de los datos masivos a partir de la teoría de las tres "V", y se ha mencionado el hecho de que múltiples autores la consideran la más valiosa de todas ellas. Esto es lógico, pero no absoluto.

No cabe duda de que el análisis de datos es la parte central y más importante de la ciencia y la gestión de los datos, en especial los masivos, ya que ahí es donde se extraen las soluciones y decisiones relevantes para una organización; por lo mismo, merece especial atención y estudio, y como consecuencia lógica existen más herramientas, textos y productos al respecto en comparación con otras áreas y herramientas del manejo de datos. Pero es esencial e indispensable contextualizar este campo junto con los demás: no hacerlo pone en riesgo el resultado y propósito ulterior de la gestión de datos.

De la observación de los sitios web, las ofertas de productos, capacitación y asesoría, los cursos y diplomados, las certificaciones, etcétera, pareciera a primera vista que esto es todo lo que tiene que ser atendido con respecto a la ciencia y la gestión de los datos: el análisis de ellos lo es todo. Si la organización resuelve el capítulo del análisis de los datos, todo el problema queda resuelto.

Ello se explica ante la grave carencia de expertos a nivel mundial mencionada anteriormente acerca de la ciencia de los datos y su gestión. Por supuesto es urgente formar y capacitar personal al respecto, pero es indispensable formarlo y entrenarlo en todas las áreas de la ciencia y la gestión de los datos. No puede desarrollarse solo uno de sus campos por más esencial que sea, pues ello introduce un desbalance nocivo y peligroso para el propósito final. Todo campo del conocimiento tiene áreas torales que ameritan estudio y formación especial, pero también áreas complementarias sin las cuales el campo central pierde su valor y su integridad.

Los campos del conocimiento humano tienen una parte dialéctica con teorías, bases, principios, fundamentos, conceptos, metodologías, normativa, así como una parte práctica con procedimientos, experiencias, estudios, recomendaciones, manuales, estándares, técnicas y, obviamente, herramientas al respecto. El campo de los datos no es la excepción: tiene una parte teórica –la ciencia de los datos– y una parte práctica –la gestión de los datos– ambas con todos sus subcampos y disciplinas afines. Como en todos los demás campos del conocimiento humano, nadie puede saber todo. Por tanto, hay personas que se van volviendo expertas

en alguno o algunos de los subcampos de la teoría y otras en alguno o alguno de los subcampos de la práctica; algunos pocos llegan a adquirir algo de experiencia en ambos. Las personas se van especializando de acuerdo con sus estudios y a su experiencia ante la imposibilidad de saber todo.

Como en todas las disciplinas aplicadas, lógicamente hay más personas dedicadas a la práctica que a la teoría, e igualmente habrá un mayor número de personas dedicadas a los subcampos más relevantes y con mayor demanda del mercado. Por tanto, suena lógico que en la gestión de datos –y en especial los masivos– haya más personas dedicadas al análisis de datos al ser una de las áreas más importantes. Pero no es conveniente en modo alguno para la disciplina o para las organizaciones que todos se formen en ese subcampo. Toda organización debe tener personal que sepa cómo analizar los datos y extraer soluciones de ellos, pero además debe tener personal que sepa cómo diseñarlos, colectarlos, normalizarlos, auditarlos, almacenarlos, accederlos, protegerlos, etcétera. Si la organización tiene un solo experto dedicado a los datos, esta persona debe saber todo acerca de ello de una manera balanceada y contextualizada. De nada sirve tener un experto en su análisis si todas las demás áreas han sido descuidadas. Si se tiene un departamento o área dedicado a los datos, cada una de las personas dentro del mismo debe estar formada en alguna o algunas de las áreas de la gestión de datos y en conjunto deben cubrir todos o al menos la mayoría de los aspectos de la misma.

Los conocimientos del personal de una cierta organización provienen por un lado de la formación profesional, y por otro de la capacitación, educación continua y entrenamiento. No puede afirmarse de manera absoluta que alguna de las dos vertientes sea más importante que la otra; ambas tienen un valor preponderante y una utilidad dentro de cada contexto. A algunas organizaciones les es conveniente tener personal profesional para ciertas tareas; a otras les es útil tener personal entrenado y experimentado en sus quehaceres prácticos; a algunas más les conviene tener ambos tipos de personal. En todos los casos, se debe cubrir la variedad más amplia de especialidades de acuerdo con sus necesidades. El punto

central de lo anterior es que a ninguna organización le es conveniente tener únicamente expertos en análisis de datos si no tiene a nadie que pueda contender con los demás aspectos de su gestión. Igualmente con las personas, no es conveniente formarse únicamente en análisis de datos; si no hay otros expertos en las otras áreas dentro de la organización, sus conocimientos y habilidades serán estériles para una gestión exitosa de los datos al no existir un contexto y un balance de los mismos dentro de la institución.

Los campos de la gestión de los datos son como las patas de una mesa: se requiere de todas para estar firme, sostenerse y ser funcional. Si se retira una de ellas, el conjunto puede sostenerse aparentemente, pero ante cualquier carga o uso se derrumbará. Si se omiten dos, será muy difícil siquiera aparentar que funciona, y si se omiten tres ni siquiera puede afirmarse que lo restante es una mesa. El análisis de los datos es simplemente una de las patas de la mesa: sin los demás aspectos simplemente no funciona. He ahí la importancia del contexto y del balance del conjunto. Lorica (2014, 4-5) estableció al respecto:

> […] A juzgar por la prensa popular, el término "científico de datos" ha venido a referirse cada vez más a alguien que se especializa en el análisis de datos: estadísticas, aprendizaje a máquina, etcétera [...] Lejos de limitarse al análisis de datos, un flujo de trabajo típico de la ciencia de los datos significa saltar incesantemente en una serie de tareas interdependientes. Los científicos de datos tienden a utilizar una gran variedad de herramientas [….] Los flujos de trabajo que implican muchas herramientas diferentes requieren mucho intercambio de contextos […].

Como conclusión a todo lo anterior, puede afirmarse que las organizaciones y las personas deben considerar al análisis de datos como una de las múltiples opciones de capacitación y entrenamiento, pero nunca como la única opción a desarrollar a nivel corporativo o personal. Siempre debe ponderarse el todo y con especial cautela con quién se obtendrá la capacitación. Recuérdese que en la actualidad existe un enorme modelo comercial

alrededor de la gestión de datos, y en especial en la capacitación, asesoría y certificaciones al respecto. Ello se ha vuelto un negocio millonario por parte de múltiples empresas ofertantes; obviamente esto es válido en buena medida, pero por otra parte se ha ido gestando una enorme expectativa alrededor del campo más atractivo del análisis de datos, donde buena parte de ello no atiende al contexto general de lo que las organizaciones o las personas realmente necesitan, y donde no todos los ofertantes son serios. Muchos venden porque está de moda y es buen negocio aunque lo que ofrezcan no sea de calidad o realmente útil. He ahí la capital importancia de ponderar el todo y el con quién.

El punto central de todo lo anterior que no debe perderse de vista es que para solucionar un problema de datos, su análisis no es suficiente; se requiere que concurran integral y equilibradamente todas las partes de la gestión de datos. Es conveniente resaltar en este punto que el proceso de los datos comprende muchos pasos; si bien no todos ellos se encuentran en cada uno de los problemas relacionados, en efecto se requieren varias etapas para su adecuada gestión: diseño, modelado y tipificación de datos, captura, colecta o minería de ellos, su codificación y depuración; normalización y estructuración, inclusión de metadatos; transformación; agregación o desagregación de datos, validación; almacenamiento, gobernanza y preservación; visualización y despliegue de los datos, y sí, finalmente, su análisis e interpretación. Puede concluirse de ello que una organización requiere de más personal especializado que solo un analista de datos, y además, que un buen analista de datos no tan sólo sabe y puede analizarlos; también debe ser capaz de participar integralmente en todas las demás etapas.

Hecha la reflexión anterior, puede entrarse en materia de lo que implica puntualmente el análisis de datos como especialidad: consiste en un subconjunto de la gestión de los datos que se refiere a la tarea de identificar aquellas variables de la organización que pueden ser relacionadas con ciertos datos y establecer así correlaciones para el planteamiento de preguntas y la eventual extracción de respuestas a través de técnicas y herramientas específicas con el propósito final de plantear decisiones y líneas

de acción. Los autores dividen el análisis de datos en cuatro grandes áreas o subdivisiones:

a) Análisis Prescriptivo. También llamado "Analítica Prescriptiva" (*Prescriptive Analytics*): consiste en un tipo de análisis de los datos en bruto; específicamente, maneja los distintos escenarios posibles de la organización, sus recursos disponibles, comportamiento pasado y actual, recomendaciones y acciones anteriores, y a partir de ello se puede sugerir un curso de acción o estrategia a corto o largo plazo (Segal 2019).
b) Análisis Diagnóstico, también llamado "Analítica Diagnóstica" (*Diagnostic Analytics*): consiste en un tipo de análisis que investiga a profundidad dentro de los datos y hechos del pasado para tratar de explicar las causas y efectos de ciertos eventos ocurridos; es un paso ulterior para descubrir el razonamiento detrás de ciertos resultados. Utiliza técnicas como el descubrimiento de datos, el dragado y minería de datos, las correlaciones, etcétera, para determinar las mejores fuentes para resolver cierto problema (SISENSE s.f.).
c) Análisis Predictivo, también llamado "Analítica Predictiva" (*Predictive Analytics*): semejante al anterior, pero se basa en estudiar los datos y hechos recientes para tratar de predecir ciertos eventos, tendencias o comportamientos. Utiliza técnicas avanzadas de análisis que aprovechan los datos para descubrir hechos y situaciones en tiempo real y predecir acontecimientos futuros. Es una nueva mezcla de análisis estadístico clásico apoyado en herramientas de Inteligencia Artificial (IBM s.d.).
d) Análisis Descriptivo, también llamado "Analítica Descriptiva" (*Descriptive Analytics*): analiza las bases de datos existentes en la organización para tratar de describir y explicar cierto estado de la cuestión o patrón de un cierto conjunto de datos dentro de ella. Tiene por objeto proporcionar descripciones o resúmenes de hechos y cifras en formatos comprensibles tanto para informar, como para preparar los datos para análisis posteriores. Se basa en la extracción y

> la agregación de datos, así como en técnicas y herramientas de visualización (Dataversity 2017).

Algunos productores de herramientas y servicios para estas tareas pretenden hacer una diferencia entre el simple "análisis de datos" (*data analysis*) y la "analítica de datos" *(data analytics),* señalando que esta última es una versión mucho más sofisticada que el primero, y que las diferencias son radicales. Obviamente sus productos buscan presentarse como versiones más completas y adelantadas. En realidad, de la lectura de los textos al respecto se desprende que solo es una diferencia semántica introducida artificialmente con fines de mercadotecnia, y que no son dos conceptos diferentes: son simplemente dos maneras de denominar lo mismo enfatizando ciertos elementos.

Lo que sí es un hecho es que en efecto existen técnicas y herramientas específicas para el análisis de datos masivos, las cuales hacen énfasis en las características de este tipo de datos definidas por las "V": Volumen, Velocidad, Variedad, Veracidad y Valor. Dichas características conllevan a que en efecto los métodos estadísticos que funcionan bien para datos de volúmenes reducidos no se escalan adecuadamente a datos masivos. Del mismo modo, muchas técnicas computacionales que funcionan adecuadamente para datos de volúmenes moderados se enfrentan a importantes problemas en el análisis de datos masivos. Todo ello introduce nuevos desafíos: almacenamiento y análisis de datos masivos; descubrimiento de conocimientos; complejidades computacionales; escalabilidad de datos, visualización de los mismos, y seguridad de la información. Por lo anterior es muy conveniente tener en mente que en efecto hay consideraciones diferentes del análisis de datos para conjuntos de menores dimensiones con respecto al análisis de datos masivos. Este último conlleva técnicas y herramientas diferenciadas o adecuadas, tales como el análisis estadístico, el aprendizaje de máquina, la minería de datos extendida, el análisis inteligente, los datos con buena gobernanza, la computación en la nube, la computación cuántica y el procesamiento de flujos de datos (Acharjya y Ahmed 2016, 512-517).

En tiempos recientes, se ha ido desarrollando una nueva variante del análisis de datos denominada "analítica profunda" (*deep analytics)*, la cual consiste en procesos de aplicación de la minería de datos combinada con otras técnicas de procesamiento de datos para analizar, extraer y organizar cantidades de datos sumamente masivas en una forma aceptable, útil y beneficiosa para una organización con el fin de encontrar nuevos usos y/o aplicaciones de esos datos. Este tipo de análisis por lo general se realiza extrayendo ciertos datos selectos de inmensos conjuntos de ellos –Petabytes o hasta Exabytes– distribuidos a lo largo de arquitecturas complejas y dispersas en múltiples almacenes de datos, y por lo mismo requiere de recursos muy especializados para poder realizar este proceso: computación distribuida en varios servidores o nodos informáticos, computación en la nube, búsqueda distribuida y análisis profundo de metadatos, etcétera. El análisis profundo a menudo se combina con aplicaciones de IA, aprendizaje de máquina, minería de datos, correlaciones complejas, etcétera, para poder así lograr su cometido y extraer de esos inmensos conjuntos de datos información útil y especializada.

Entrando al campo de las bibliotecas, conviene revisar los temas y usos del análisis de datos específicamente dentro de estas organizaciones. Diversos autores, como Showers y colegas (2015) han señalado ya algunas de las aplicaciones prácticas de ello: desarrollar colecciones; diseñar nuevos servicios; conocimiento de los usuarios, sus preferencias y necesidades; demostrar cuantitativamente el valor e impacto de la biblioteca dentro de su entorno; mejorar la experiencia educativa e informativa de los usuarios; el ya mencionado "análisis del aprendizaje", y en general, para mejorar la toma de decisiones. Para ello puede utilizarse sin duda el análisis de datos, tanto masivos como en menor escala. Pero ambos conllevan un perfecto entendimiento del significado del análisis de datos para no caer en errores o imprecisiones.

Como ya se mencionó al principio de este apartado, el verdadero núcleo del análisis de datos se refiere puntualmente a la tarea de identificar cuáles variables de la organización pueden ser relacionadas con ciertos datos y establecer así correlaciones para el

planteamiento de preguntas y la eventual obtención de respuestas a través de técnicas específicas. Esto significa que antes de analizar datos, la biblioteca debe asegurarse de seleccionar aquellos que en efecto representan ciertos hechos o realidades, para poder así establecer y obtener las correlaciones correctas que conducen a preguntas y respuestas reales y pertinentes.

Uno de los errores que se cometen con más frecuencia consiste en utilizar automáticamente los datos "tradicionales" que la biblioteca ya colecta para sus estadísticas; dice el adagio popular que "el que solo sabe usar un martillo, a todo le ve forma de clavo". Otro error común consiste en colectar los datos que se pueden obtener más fácilmente, simplemente por esa razón: porque son fáciles de obtener. El análisis de datos expresamente indica que la selección de los datos debe tener de inicio una estrecha correlación con los hechos a representar; de otra forma, los resultados serán sesgados o irrelevantes. Aunque suene obvio, esto no es fácil; requiere por un lado de un profundo conocimiento de las causas y efectos de la función o tarea de la biblioteca que se desea representar, y por el otro, de un amplio dominio acerca de cómo plantear y diseñar correlaciones específicas de datos. Ello implica dos especialidades: el bibliotecario experto en esa función o tarea de la biblioteca y el analista de datos. Lo ideal es que fuesen la misma persona; esto es, un bibliotecario con experiencia en su área y además con buen conocimiento en el análisis de datos; no obstante, una dupla de dos expertos en estrecha sincronía también hará bien la tarea. Sin esta función dual correctamente aplicada, es fácil caer en la obtención y el uso de datos demasiado simples para la tarea, o lo contrario: pretender obtener inmensas cantidades de datos de los cuales la mayoría son irrelevantes o poco representativos de la función, que a la vez dificultan la tarea. Debe planearse desde el inicio con exactitud qué se va medir, por qué y cómo. He ahí la capital importancia de esa dupla de expertos, o del experto con dos especialidades.

Las áreas de la biblioteca donde estos análisis pueden aplicarse son numerosas, como ya ha sido mencionado; los posibles puntos de extracción y colecta de datos lo son todavía más: las búsquedas en catálogos y colecciones, las consultas, el préstamo de libros,

las vistas y descargas de documentos, el acceso con clave a ciertos servicios, las redes sociales entre la biblioteca y sus usuarios, por citar algunos. Entre tantas posibilidades, es fácil perderse en el mar de datos a obtener y por ello es imperativo desde el principio hacer un correcto diseño de ellos y sus hechos asociados para de ahí partir hacia un análisis pertinente. No obstante, el esfuerzo vale la pena; Farmer y Safer (2016, 6) lo expresan así:

> […] Los beneficios del análisis de datos cuantitativos para la mejora de las bibliotecas se traduce en una mejor recuperación de la inversión […] El análisis de datos da como resultado general el máximo aprovechamiento de los recursos, dado que el rendimiento y la productividad mejoran, y la calidad se controla, por lo que no solo se reducen los gastos, sino que también aumenta la satisfacción de los usuarios y del personal.

Ésta es una de las varias obras que la American Library Association (ALA) ha editado al respecto del análisis de datos en las bibliotecas, la cual consigna muchas de las habilidades y los conocimientos que los bibliotecarios deben adquirir al respecto: fundamentos de conceptos estadísticos; conocimiento de las fuentes de datos recomendadas para diversas funciones y procesos de biblioteca, así como orientación acerca de cómo usarlas; técnicas para limpieza de datos; técnicas para encontrar las correlaciones de los datos con los métodos apropiados de análisis al respecto, y cómo visualizar los resultados. Consciente de la importancia del análisis de datos para mejora en la biblioteca, la ALA ha ido produciendo una serie de textos a este respecto, cuyo conjunto cubre diversos temas específicos; por ejemplo, Farney (2018) trata la extracción de datos de los usuarios a partir de sus actividades de navegación en la red, los catálogos de la biblioteca, los descubridores de información y los repositorios. En este texto, la autora explica cómo establecer conjuntos de datos adecuados y coherentes a partir de esta actividad, analizando e identificando las áreas que se ajustan a las prioridades de la biblioteca; cómo limpiar los datos y combinar diversas fuentes de ellos, y cómo realizar los análisis correspondientes para ulteriormente aplicar los resultados en mejorar las colecciones y la

satisfacción de los usuarios. De la misma serie de la ALA, y en conjunto con la Asociación de Bibliotecas Universitarias y de Investigación (ACRL) y la Asociación de Bibliotecas Públicas (PLA), Hernon y colegas (2015) explican cómo extraer datos significativos y coherentes para la gestión de una biblioteca y mejorar su rendición de cuentas a través de un mejor uso de las colecciones, la evaluación comparativa y otras prácticas óptimas.

Como puede verse de los textos anteriores, que son solo una muestra, existe ya un fuerte interés de las asociaciones bibliotecarias por estudiar, difundir y capacitar acerca del tema del análisis de datos a los profesionales de la disciplina. De hecho, en muchos de los textos al respecto puede observarse que algunos autores ya empiezan a denominar al análisis de datos aplicado en bibliotecas más puntualmente como "análisis bibliotecario" (*library analysis)*, para señalar con esta acepción específicamente a la actividad relacionada a los datos en este entorno y hacer evidente que ya es una especialidad dentro de las bibliotecas.

La gobernanza de los datos masivos

La calidad nunca es una feliz coincidencia. Siempre proviene de un esfuerzo inteligente.

JOHN RUSKIN

Los datos cobran cada vez mayor importancia en una gran cantidad de negocios y organizaciones de todo tipo y tamaño. Por su naturaleza de volumen, velocidad de generación y variedad, deben ser administrados adecuadamente; de lo contrario, los datos pueden causar más problemas que beneficios. He ahí la importancia de la gobernanza de los datos, en especial los masivos. Existen numerosas definiciones de este concepto, las cuales hacen énfasis en uno u otro aspecto del mismo; por ejemplo, la Asociación Internacional de Gestión de Datos (Data Management Association o DAMA) define la gobernanza de datos como "[…] la planificación, supervisión y control de la gestión de los datos y la utilización de los datos y las fuentes relacionadas con ellos". Techopedia (2020) define este concepto como "[…] una gestión global de la calidad, la facilidad de uso, la disponibilidad, la seguridad y la coherencia de los datos de una organización". Ya se hizo mención anteriormente de la definición del Instituto de Gobernanza de Datos. Como resultante de muchas de estas definiciones y en resumen, puede establecerse que la gobernanza de datos tiene como objetivo garantizar la calidad y seguridad de los datos utilizados en una organización, estableciendo y supervisando un

conjunto de políticas, procesos, puestos, normas, métricas y responsabilidades que aseguren el uso eficaz, eficiente, seguro y coherente de esos datos.

Es importante establecer claramente la diferencia entre la gobernanza de los datos y la gestión de ellos. Olavsrud (2020) hace una diferencia entre ambos conceptos: el primero es parte del segundo; la gobernanza de los datos –siendo una parte muy importante– es sólo un subconjunto del concepto más general de gestión de datos. La gobernanza de los datos trata las funciones, procesos, estándares, métricas y responsabilidades personales para establecer con claridad la propiedad y el acceso de los activos de datos al interior de la organización; su uso protegido, su coherencia, así como la rendición de cuentas. La gestión de datos es un término mucho más amplio que describe todos los procesos utilizados para planificar, especificar, habilitar, crear, adquirir, mantener, utilizar, archivar, acceder, controlar, depurar y gobernar los datos.

La gobernanza de los datos crea un marco de referencia de los datos dentro de una organización; la gestión de los datos conlleva la ejecución práctica de ese marco. Esencialmente, la gobernanza de los datos busca que, en lo tocante a ellos, las personas adecuadas tengan las responsabilidades correctas por medio de procesos, normas y métricas pertinentes.

Por su naturaleza, la gobernanza de los datos conduce a ventajas que por lo general no son obtenibles solo a través de su gestión:

- Proporciona una visión y comprensión generalizada y coherente, así como una terminología común para los datos dentro de la organización.
- Permite optimizar procesos, procedimientos, funciones y puestos dentro de la organización.
- Mejora la calidad de los datos y garantiza la exactitud, integridad y consistencia de los mismos.
- Optimiza la ubicación de todos los datos relacionados con las áreas fundamentales de la organización.
- Define claramente a las personas dentro de la organización que manejan datos y sus responsabilidades.

- Permite el cumplimiento cabal de los requisitos legales mínimos exigibles, en especial aquellos relacionados con la privacidad y la protección de datos personales.
- Previene conflictos y/o traslapes entre los distintos conjuntos de datos y las secciones de la organización.
- Ahorra costos de operación.

Básicamente, la gobernanza de datos se diseña a partir del establecimiento de una política organizacional creada específicamente al respecto. Si bien es algo que suena deseable desde el principio, en la medida en que las organizaciones van captando cada vez más datos, su gobernanza se hará cada vez más imperativa: la simple administración de ellos ya no será suficiente. Para su diseño e implementación, existen ya varias listas o enunciados de principios básicos. Techopedia (2020) consigna una lista de seis de ellos:

1) Reconocer los datos como un activo: De inicio debe reconocerse la importancia y el valor de los datos como activos para una organización. De ello se derivan las acciones para poder definirlos, controlarlos y acceder a ellos de forma cuidadosa y orientada a los procesos. En consecuencia, la administración puede confiar en la exactitud y la utilidad de los datos.
2) Establecer propiedad y responsabilidad de los datos: Es indispensable definir claramente al interior de la organización la propiedad y responsabilidad de los datos. En toda organización debe quedar muy claro quién colecta, accede y procesa qué datos, y solo debe hacerse bajo procesos definidos y autorizados. No pertenecen y no son sólo responsabilidad del área de informática. La participación en la gobernanza de los datos debe provenir de todos los departamentos para que su gestión sea productiva y exitosa.
3) Adoptar políticas, normas y regulaciones estandarizadas para la seguridad: Todo proceso de gestión de datos debe seguir estándares y reglamentos reconocidos y normalizados para evitar riesgos y fallas en lo tocante a la definición de los datos confidenciales, su acceso, sus políticas de

privacidad y normas de seguridad. Los procesos normalizados deben cumplirse estrictamente a fin de evitar problemas con los datos.

4) Gestionar la calidad de los datos de forma consistente: La calidad de los datos debe ser definida, creada y mantenida de manera consistente desde el principio. Los datos de la organización deben ser comprobados periódicamente contra los estándares de calidad preestablecidos.
5) Gestionar el cambio: El proceso de gestión de datos debe definir las actividades de gestión del cambio dentro de la organización de manera proactiva. Por este motivo, es muy importante hacer un seguimiento a lo largo del tiempo de los cambios en los datos y las consecuencias derivadas.
6) Auditar los datos: Una auditoría de datos debe ser un proceso estándar en las organizaciones. Por lo tanto, el proceso de gobernanza de los datos debe apoyar una política de auditoría transparente. La gobernanza de los datos es un proceso continuo y debe gestionarse adecuadamente a lo largo de los años. Los principios básicos de esta gobernanza deben mantenerse sencillos y comprensibles para todos los niveles de la organización.

Algunas empresas ofrecen plataformas y herramientas informáticas de gobernanza de datos, como muchos otros productos y servicios relacionados con ellos. Como una muestra de ellas se encuentran:

- Erwin Data Governance. Herramienta de "*Software* como Servicio" en la nube o SaaS, la cual trata de administrar los procesos definidos como parte de la gobernanza de datos, tales como el modelado de los procesos, el modelado de datos o la biblioteca de datos.
- Data Governance de Adobe Experience Platform. Ofrece clasificar, gestionar y reforzar la forma de uso de los datos en toda la organización para obtener el máximo provecho de ellos.

- Talend Data Fabric. Consiste en un conjunto de aplicaciones para colectar, gobernar, transformar y compartir datos entre la nube y el entorno local de la organización.
- Onna. Aplicación que se opera en la nube y que se encarga de centralizar la información que está dispersa con el fin de conservar tanto información histórica como actual en tiempo real.
- ManageEngine ADAudit Plus. Herramienta de *software* que permite hacer auditoría sobre los datos.

La lista anterior se presenta solamente con fines enunciativos. Todos los productos de este tipo deben verse con máxima cautela. Cabe resaltar en este punto que la gobernanza de los datos consiste fundamentalmente en crear un marco de referencia coordinado de los datos dentro de una organización; ello se logra primordialmente con la construcción de una política al respecto de la cual se derivan procedimientos, procesos, etcétera: definitivamente no consiste en la utilización de herramientas informáticas.

La gobernanza es una conceptualización teórica que se escribe, no un programa de computadora que se ejecuta. Obviamente es conveniente usar herramientas computacionales para poner en marcha algunas de sus partes, como la auditoría de datos, la creación de tablas de datos maestros o la seguridad informática; también para verificar el cumplimiento de ellas, pero no debe confundirse por ningún motivo la utilización de aplicaciones destinadas a ejecutar o supervisar alguna de sus partes con la gobernanza como un todo. Como ya se ha establecido, la gobernanza contempla muy diversos aspectos de cohesión, calidad, seguridad, responsabilidad de los datos, etcétera. A la fecha, no existe ninguna aplicación informática que permita integrar universalmente en un solo lugar todos esos aspectos. Las herramientas de este tipo en el mercado pueden facilitar la tarea atendiendo a una o varias de sus partes, pero definitivamente no sustituyen el proceso intelectual de concepción y diseño de una estrategia coherente de los datos, ni tampoco el de su supervisión.

Para comprender esto mejor, podemos utilizar la lista o "Marco de Referencia Funcional de Gobernanza de Datos" elaborado por la Asociación Internacional de Gestión de Datos (DAMA), en donde señala diez áreas de atención, acción y desarrollo en lo relativo a este tema (DAMA s.d., 5). He aquí un resumen:

1) Arquitectura de datos: reconocer los datos como activos con valor, así como los recursos relacionados con ellos como parte integral de la organización; establecer la estructura general de datos para ella.
2) Modelado y diseño de datos: analizar, diseñar, construir, probar y dar mantenimiento a los datos.
3) Almacenamiento de datos y operaciones: administrar las estructuras y los dispositivos para el almacenamiento de datos y su acceso.
4) Seguridad de datos: establecer medidas para garantizar la privacidad, la confidencialidad y el acceso adecuado.
5) Integración e interoperabilidad de datos: diseñar, colectar o extraer, transformar, transferir, replicar y mapear datos buscando datos integrados e interoperables a lo largo de procesos y operaciones.
6) Documentos y contenido: transformar, indizar y optimizar el acceso a los datos encontrados en fuentes no estructuradas para su integración e interoperabilidad con los datos estructurados.
7) Datos de referencia y datos maestros: definir datos compartidos y normalizados entre áreas de la organización para reducir redundancias y garantizar mejor calidad de ellos; crear conjuntos de "datos maestros".
8) Inteligencia empresarial: generalizar el procesamiento y análisis de los datos para optimizar la toma de decisiones en toda la organización.
9) Metadatos: diseñar, estructurar, recopilar, clasificar, mantener, integrar, controlar, gestionar y entregar adecuados metadatos.

10) Calidad de los datos: definir, vigilar y supervisar la coherencia e integridad de los datos para mejorar la calidad de los mismos.

Esta lista, que es bastante exhaustiva, ayuda a comprender mejor lo expuesto anteriormente. De su análisis se desprende que los puntos fundamentales de la gobernanza de datos son todos conceptos a definir y establecer, de los cuales se desprenderán acciones y procesos. En ningún elemento de la lista se destaca la importancia de adquirir y utilizar una buena herramienta informática. Obviamente, ningún proyecto que gestione grandes volúmenes de datos podrá hacerlo sin recursos adecuados en ese aspecto, y por tanto la gestión de datos en efecto requiere de adecuadas herramientas informáticas. Pero se reitera el hecho de que la gobernanza de datos no es lo mismo que su gestión, y en la gobernanza –a diferencia de la gestión– la herramienta no es un punto clave. Igualmente, en los productos ofrecidos en el mercado para ello se observa que no existe ninguno que cubra todos los puntos de atención enumerados en la lista para la adecuada gobernanza de los datos.

Como en muchos otros aspectos de la administración en las organizaciones, los datos requieren que una parte de su gestión especifique los detalles fundamentales para garantizar su calidad, cohesión y seguridad dentro de ella, asegurando el uso eficaz, eficiente, seguro y rentable de esos datos. Esa parte de la gestión es sin duda la gobernanza de los datos, y como en toda organización, conviene incorporarla dentro de los datos de las bibliotecas.

Los bibliotecarios y los datos masivos

Los bibliotecarios pueden ser los guías expertos en la súper-carretera de los datos masivos.

AMY AFFELT,
The Accidental Data Scientist

Como ya ha sido establecido, la gestión de datos, en especial los masivos, es una disciplina multidisciplinaria y compete a una gran diversidad de profesionales de la información; por supuesto entre ellos están los bibliotecarios debido a su formación y experiencia usuales. Como todas las especializaciones, requiere de conjuntos nuevos de conocimientos, habilidades, actitudes y experiencia.

En términos generales, los bibliotecarios necesitan conocer hoy en día los fundamentos de la ciencia de los datos, su gestión y curaduría; las fuentes de información al respecto del tema; la bibliominería, cómo y cuándo se utilizan los datos masivos en la biblioteca, y dónde pueden encontrarse; el análisis y visualización de datos, así como la selección de herramientas informáticas para su explotación, todo ello como parte de una formación integral en su disciplina. A nivel mundial, hay una gran escasez de profesionales de los datos que posean un adecuado conocimiento de cómo diseñar, definir, colectar, depurar, transformar, analizar y presentar proyectos y estructuras de datos. Si cualquiera de estos pasos de la gestión de datos no está bien conceptualizado y no se tiene idea clara del valor, el procesamiento y la utilización de los

mismos, estos serán desaprovechados y las instituciones se embarcarán en proyectos inciertos o se quedarán estáticas sin iniciar ninguno debido a la falta de capacidades técnicas. El personal calificado, y no las herramientas, es la principal clave del éxito en estas iniciativas.

Más allá de esta formación integral de la disciplina, y en función del tipo de biblioteca y/o organización en la que trabajen, los bibliotecarios deben ir adquiriendo y profundizando en ese conocimiento para poder ayudar en su entorno en muy diversos campos de aplicación. He aquí un resumen de esas actividades recomendadas por múltiples autores; entre ellos Bieraugel (2013), Showers (2015), Lyon y Mattern (2016):

- Los bibliotecarios de instituciones de investigación deben comprender cómo los datos se relacionan ahora con la investigación académica, y cómo pueden ser difundidos y reaprovechados.
- Los bibliotecarios de instituciones de enseñanza superior deben comprender cómo los datos provenientes de profesores y alumnos pueden ser utilizados para mejorar la eficiencia educativa de la institución, los planes de estudio, la orientación vocacional, las experiencias del aprendizaje, la retención en la carrera, etcétera. También pueden capacitar a los alumnos en el mejor uso y explotación de los datos.
- Los bibliotecarios de empresas en general necesitan saber cómo éstas pueden aprovechar los datos masivos, la minería de datos, su análisis, etcétera, para incorporar ventajas competitivas al negocio.
- Los bibliotecarios de empresas de Servicios de Información Bibliotecarios (LIS) deben estar preparados acerca de las cambiantes necesidades y características de esos servicios y sus usuarios para poder aprovechar los datos masivos y su análisis con el fin de diseñar y crear nuevos productos y servicios al respecto, así como optimizar los existentes.
- Los bibliotecarios de ciencias sociales y humanidades deben estar conscientes de que los datos y su análisis también

son cada vez más comunes en sus disciplinas, y que cada vez más las "humanidades digitales" requieren de estas herramientas y a sus especialistas.

- Los bibliotecarios que se dedican a la organización y el registro de la información deben estudiar y diseñar nuevas taxonomías, esquemas de metadatos y estructuras para hacer a los grandes conjuntos de datos más visibles, accesibles y útiles.
- Los bibliotecarios que se dedican a la sistematización de métodos de recuperación documental deben diseñar nuevos mecanismos y herramientas para lograr mejores resultados de la búsqueda de información y descubridores.
- Los bibliotecarios que se dedican a la bibliometría necesitan diseñar nuevos mecanismos y herramientas para extraer mejores resultados del análisis de citas, referencias, palabras clave, uso y análisis de textos completos.
- Los administradores bibliotecarios deben asimilar cómo el uso de datos masivos puede serles útil para rediseñar múltiples aspectos administrativos de la biblioteca: adquisición y uso de sus recursos de información; decisiones de aplicación y/o redireccionamiento del presupuesto; horarios de operación, selección y descarte documental; reducción de costos ocultos; diseño, usabilidad, rendimiento y percepción de nuevos servicios.
- Los bibliotecarios que se dedican a la curaduría de datos pueden asesorar a los miembros de sus comunidades acerca de la colecta, el almacenamiento y la accesibilidad de grandes conjuntos de datos, construcción de repositorios, y especialmente cómo crear datos estructurados.
- Por todo lo anterior, los bibliotecarios que se dedican a la docencia deben comenzar a formar profesionales de la bibliotecología que puedan comprender el alcance dentro de la disciplina de las actividades enunciadas y vayan adquiriendo los conocimientos y habilidades básicas durante su formación profesional.

- Además de todas las disciplinas y especialidades directamente relacionadas con los datos, es importante resaltar que en la actualidad existen muchas otras que tienen ya estrecha relación con ellas y que son de interés para la bibliotecología, tales como las humanidades digitales, las ciencias sociales digitales, la computación social, etcétera.

Como puede verse, cada uno de los enunciados anteriores abre una amplia posibilidad de proyectos de datos en las bibliotecas y organizaciones afines; indudablemente el campo de acción en ellas es sumamente amplio. Partiendo de conocimientos generales y comunes, en cada una de esas eventuales aplicaciones se requieren profesionales con conocimientos, habilidades, actitudes, experiencias y dominio de herramientas específicas.

Claro que suena imposible que una sola biblioteca tenga especialistas en todos y cada uno de esos campos, pero también debe quedar claro que no puede carecer totalmente de algún personal especializado en ello. Por tanto, la biblioteca puede y debe ir formando sus especialistas de datos en los campos de aplicación propios de su interés y contexto. En primer lugar, porque eso le permite entrar en dimensiones acordes a las necesidades y circunstancias actuales del mundo de la información para poder seguir siendo competitivas e interesantes para sus comunidades y sus financiadores. En segundo lugar, porque ello permite formar profesionales altamente demandados en los tiempos actuales y de los cuales hay una escasez a nivel mundial, y esto significa nuevos y mejores puestos de trabajo para los bibliotecarios profesionales actuales y los estudiantes de la carrera.

Lyon, Mattern, Acker y Langmead (2015), y Lyon y Mattern (2016) realizaron dos estudios en el campo de la ciencia de los datos en los cuales identificaron un grupo de seis roles o funciones principales en esta ciencia a partir de necesidades del mundo real para puestos de trabajo reales. Dentro de los seis roles se encuentran los "tradicionales": analista de datos, ingeniero de datos y periodista de datos, pero además ellos encontraron otros tres roles adicionales relacionados estrechamente con la disciplina de

nuestro interés: bibliotecario de datos, archivista de datos y gestor/curador de datos. Ellos también establecieron que estos profesionales de la información debían desarrollar cinco aspectos en relación a este tema (Lyon y Mattern 2015, 3):

1) Formación académica.
2) Experiencia práctica.
3) Conocimiento, familiaridad y comprensión de los temas.
4) Habilidades de ejecución.
5) Competencias y dominio de herramientas y tecnologías.

Como se ha mencionado a lo largo de este texto, la intención del mismo ha sido ilustrar en el conocimiento, familiaridad y comprensión de los temas y, en parte, en las habilidades y competencias.

Al margen de la inmensa oferta de cursos especialidades, certificaciones, etcétera, ofrecidos comercialmente por innumerables empresas con respecto al tema de los datos, existe además una cantidad pequeña pero aceptable de cursos, textos y capacitaciones sobre el tema ofrecidos sin fines de lucro por diversas organizaciones, los cuales son ideales para que los bibliotecarios puedan iniciarse en este campo. Como ya se mencionó, debe tenerse mucha cautela al seleccionarlos, pues numerosas empresas y organizaciones ofrecen cursos supuestamente gratuitos a manera de gancho para captar a los eventuales alumnos, pero tras un análisis más meticuloso, se observa que hay costos ocultos subyacentes, algunos de ellos muy altos; la gran mayoría de ellos derivados de la emisión de una certificación del curso. Esta recomendación cautelar incluye a cursos interesantes en universidades de prestigio, como Harvard o el MIT; si no se busca la certificación, las posibilidades de cursos gratuitos aumentan sustancialmente.

Como en todo aprendizaje, se recomienda ir de lo general a lo particular, y de lo básico a lo especializado. Por ejemplo, no es conveniente iniciar el aprendizaje con el análisis de datos sin tener nociones generales de la ciencia y la gestión de los datos, ni es recomendable iniciarse en programación Python si no se dominan los principios más sencillos de la programación y consulta

en bases de datos con SQL. Por ello, se recomienda usar la información desarrollada a lo largo de este texto para elaborar un plan personal de capacitación y aprendizaje en datos con una evolución y profundización gradual. Como en muchas otras disciplinas emergentes, la oferta es infinitamente mayor en idioma inglés que en español, por lo que debe considerase primariamente la existencia de cursos y textos en ese idioma.

A guisa de ejemplos, Bernard Marr (2020) señala una selección de la revista *Forbes* acerca de nueve cursos gratuitos sobre diversos temas en ciencia de los datos. Puede verse también una serie de pequeños cursos y textos introductorios al respecto de los datos en el sitio de YouTube –algunos de ellos bastante buenos– buscando bajo las entradas *"data science"*, *"data management"* y *"data analysis"*, así como sus equivalentes en español: "ciencia de los datos", "gestión de datos" y "análisis de datos". Se hace énfasis aquí en la misma recomendación de cautela al respecto de los cursos "gancho".

El punto central de todo lo anterior es que hoy en día es altamente recomendable que los bibliotecarios profesionales comiencen a buscar un cierto adiestramiento y capacitación en la ciencia de los datos, su gestión, su análisis, sus usos y aplicaciones, etcétera. Pero no necesariamente tienen que considerarse para ello posgrados, diplomados o largas tandas de cursos formales escolarizados. Todas éstas son buenas opciones de capacitación, y si se tienen a la mano, indudablemente deben ser aprovechadas, pero no todos los profesionales están en capacidad de dedicarle el tiempo y dinero que ello implica. Las alternativas señaladas en este apartado son opciones válidas para lograr esa capacitación sin invertir tantos recursos, a un ritmo que cada quien establezca y en el horario que mejor le convenga. Cabe resaltar aquí nuevamente que en este entorno, las certificaciones son buenas, pero no indispensables, y su costo es alto. En los mercados laborales de las disciplinas que tienen una demanda equiparable o menor a la oferta, las certificaciones son un factor crucial para determinar diferencias entre los aspirantes. Pero éste no es el caso: dada la gran escasez universal de expertos en el tema de los datos, lo que más

pesa al final en este mercado es demostrar en la práctica que se es capaz de realizar una tarea solicitada y no tanto mostrar un papel que lo afirme. La experiencia aquí es un factor de gran peso para ello, por lo que es deseable que el personal interesado en el tema se involucre en cierta forma en proyectos que su institución emprenda al respecto para poder dominar esa práctica.

Resumen y conclusiones

Lentamente, estamos entrando en una era en la que los datos masivos son el punto de inicio y no el final.

PEARL ZHU,
"The Digital Master"

La segunda mitad del siglo XX se caracterizó por un enorme crecimiento de la información producida en el mundo. Dentro de este fenómeno, hay un segmento muy significativo que son los datos. En lo que va del presente siglo, este sector ha crecido en proporción todavía mucho más en comparación con la información. Actualmente el mundo produce y consume una inmensa cantidad de datos; estos se han convertido en sí mismos tanto en una fuente de información como en un nuevo insumo. Ellos se agregan ahora a los productos de información "tradicionales" o "terminados"; es decir, el resultado concluido del análisis y síntesis de ciertos datos por parte de personas o grupos en forma de publicaciones: libros, revistas académicas, textos periodísticos, manuales, tesis, compendios, patentes, etcétera. Por su propia naturaleza, los datos requieren de un tratamiento específico, lo cual dio origen a una serie de teorías, principios, modalidades, métodos, herramientas y tecnologías para su tratamiento y uso. Se creó así la gestión de datos, y con ella sus subcampos o especialidades: ingeniería de datos, análisis de datos, minería de datos, procesamiento de datos, entre otras; la suma de todas ellas ha ido conformando la ciencia de los datos: el estudio de datos organizados para identificar aquellos que son

importantes para la toma de decisiones en el contexto de un problema específico o un cierto modelo de negocio, además del desarrollo de modelos y algoritmos para la resolución de problemas a gran escala en las organizaciones.

El análisis de datos para la solución de problemas data de tiempos inmemoriales; básicamente, cuando el ser humano aprendió a recopilarlos y analizarlos de forma sistematizada y rigurosa dio forma a la ciencia moderna, y gracias a ésta, su uso y aprovechamiento fue creciendo cada vez más, así como las teorías, principios, herramientas, etcétera, para hacerlo. A lo largo del siglo XX se crearon la gran mayoría de elementos teóricos para su estudio y tratamiento formal. Además, el advenimiento y desarrollo del procesamiento electrónico de datos, y en especial de la capacidad de almacenar inmensas cantidades de datos en estos dispositivos durante ese siglo, sentó las bases para un crecimiento inédito al respecto. A esto hay que agregar el inusitado crecimiento de la red mundial y de las telecomunicaciones en la última década del siglo pasado. Todo lo anterior dio un nuevo y mayúsculo impulso a la producción de información en su forma digital: billones de piezas de información y de datos se crearon así en esta modalidad, sumándose a lo ya existente, multiplicando exponencialmente la cantidad de información acumulada. En especial, algunos sectores específicos de la red mundial han crecido enormemente en los últimos años, como las redes sociales, el "Internet de las Cosas" y los dispositivos personales que se conectan a la Internet e intercambian datos con otros dispositivos y/o sistemas de forma automática. Todo ello ha tenido como consecuencia una todavía mayor producción de datos creando un flujo inusitado de ellos. Cuando a todo esto se le agrega que alrededor de los datos se ha construido una industria multimillonaria, resultó como consecuencia una incesante y creciente producción, colecta, almacenamiento y uso de inmensas cantidades de datos en la actualidad.

Debido al auge del fenómeno de los datos, numerosos autores lo han señalado y estudiado en las últimas décadas. Varias organizaciones de renombre mundial se dedican sistemáticamente al conteo y tratamiento de la producción de datos, su valor económico y

su gestión. Innumerables empresas se han creado para la venta y el aprovechamiento de productos y servicios relacionados con los datos. De su estudio se creó todo un cúmulo de teorías, principios, metodología y herramientas. A comienzos de este siglo, todos estos factores, necesidades, tecnologías, etcétera, aunados al inmenso volumen fueron conformando a su vez un nuevo fenómeno más complejo, conocido como *"*datos masivos*" (big data)*. Este concepto consiste en el tratamiento y análisis de conjuntos de datos tan grandes, variados, complejos y dispares, producidos a una velocidad tan rápida y provenientes de muy diversas fuentes, que los equipos, programas y procedimientos "tradicionales" de procesamiento de información: servidores, bases de datos, buscadores, algoritmos, etcétera, no son suficientes y por tanto requieren métodos, equipos y programas mucho más poderosos, sofisticados y especializados para compilarlos, analizarlos y correlacionarlos. Todo con el fin de poder extraer rápidamente de esos datos patrones, tendencias y asociaciones, principalmente del comportamiento y las interacciones humanas, y a partir de ello estar en posibilidad de tomar decisiones fundamentadas que ayuden a las organizaciones, lo cual otorga a los datos un enorme valor agregado.

Los cúmulos de datos se usan hoy en día para múltiples negocios y organizaciones de todo tipo de sectores: banca y finanzas, comunicaciones y transportes, industria y comercio, salud, entretenimiento, gobierno y educación. Dentro de este último sector, se encuentran las bibliotecas, que también son susceptibles de beneficiarse de este desarrollo; especialmente en las universidades y los centros de investigación, en donde se han convertido en un nuevo y valioso insumo, por lo que se ha tenido que buscar dónde alojarlos y resguardarlos de forma adecuada y sistemática. Con frecuencia, las bibliotecas de esas organizaciones y su personal han sido designados para ello. Aunque esto pareciera a primera vista que es simplemente una pequeña actividad más agregada a las bibliotecas, no es así: representa a la vez un enorme reto y una gran oportunidad. Por un lado, implica que el personal bibliotecario debe adquirir nuevos conocimientos y habilidades para su correcto manejo, y por el otro representa nuevas oportunidades

para reposicionar a la biblioteca dentro de las responsabilidades y quehaceres y contemporáneos de su comunidad.

Esta tarea ha ido creciendo en la última década de forma notable, pero no es un simple y pequeño agregado casual: requiere de estructura organizacional y personal calificado para realizarla adecuadamente. Todas las nuevas necesidades, conceptos y soluciones derivados de esta tarea han dado origen a una nueva especialidad en el mundo de la información, la "Gestión de Datos de Investigación" o Research Data Management (RDM). Es toda una nueva responsabilidad dedicada a la gestión, el depósito y la distribución de información dentro de sus organizaciones específicamente en forma de datos. Destacados autores, como Witt y Horstmann (2016, 251), ya han enumerado las principales tareas requeridas a los bibliotecarios a este respecto: 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, y 5) alojar y mantener repositorios en sus acervos.

Grandes organizaciones bibliotecarias multiinstitucionales a lo largo de todo el mundo (IFLA, ALA, ARL, JISC, etcétera) se han concientizado de esta creciente importancia de los datos dentro de su ambiente y han creado grupos de interés y estudio acerca del tema, con la conclusión de que es insoslayable que las bibliotecas formen parte proactiva de este fenómeno. Concordantemente, el número de publicaciones acerca de los datos y su relación con las bibliotecas, el número de congresos y ponencias al respecto, las revistas, manuales y guías especializadas, así como el número de proyectos desarrollados por bibliotecas a este respecto son evidencia palpable de un interés y una importancia a todas luces creciente en este sector.

La custodia de datos de investigación en bibliotecas es sin duda la parte más visible de este fenómeno, pero de ninguna manera es el único uso posible de los datos en ellas. También están generando y usando datos propios en una amplia variedad de frentes

y tareas: para estudios de usuarios, para análisis predictivo de las colecciones y los servicios, para desarrollo de nuevas y sofisticadas herramientas de búsqueda y recuperación, para estudios de aprendizaje profundo, en sistemas expertos, traducción de textos, OCR e Inteligencia Artificial, por mencionar algunos temas.

Muchas vertientes específicas pueden encontrarse dentro de estas líneas generales del uso de datos en bibliotecas, y un buen número de ellas van más allá: caen en el campo de los datos masivos. Existen ya muchos ejemplos de ello: de inicio, en la creación y el diseño de nuevas taxonomías de información y de esquemas de metadatos. Para poder explotar información y datos –de cualquier volumen–, se requieren metadatos adecuados; sin ellos, los conjuntos de datos, en especial los masivos, son una masa amorfa con poca o nula utilidad. Hoy en día, la forma más rápida de construir todo tipo de taxonomías acerca de un campo de conocimiento de forma exhaustiva es a través del manejo y análisis coherente de grandes cantidades de datos de elementos vinculados: vocabularios con definiciones en lenguaje natural; taxonomías simples, diccionarios de datos y jerarquías; tesauros con términos relacionados, y ontologías o taxonomías completas: modelos relacionales con atributos, restricciones, relaciones; modelos de requisitos funcionales, etcétera.

A este respecto, se mencionaron como ejemplos destacados los modelos conceptuales subyacentes de las RDA, el estándar de catalogación para la formulación de registros bibliográficos para bibliotecas, archivos, museos, etcétera, tales como: los Requisitos Funcionales para Registros Bibliográficos (FRBR), los Requisitos Funcionales para Datos de Autoridades (FRAD), los Requisitos Funcionales para Datos de Autoridades de Temas (FRSAD) y la ontología PRESS, todos ellos avalados por la IFLA y compatibles con el Modelo de Referencia de Bibliotecas (*Library Reference Model).*

Se hizo referencia también a los ejemplos muy representativos de grandes sistemas bibliotecarios y organizaciones gestoras de información documental que extraen datos de sus respectivas colecciones con muchos millones de ítems para analizar y modelar desde sus registros las interrelaciones existentes entre personas,

eventos, lugares, etcétera, allí contenidos. Este tipo de proyectos está modificando radicalmente la forma de construir catálogos, buscadores y descubridores de información a raíz de estudios que han expuesto la insatisfacción de los usuarios al respecto.

Todos estos proyectos de datos están estableciendo sin duda dimensiones inéditas de gran utilidad y aceptación en el ámbito de las bibliotecas. Derivado de ellos, muchas bibliotecas ya agregan grandes conjuntos de datos adicionales a sus catálogos optimizando con ello sus buscadores: tablas de contenido, índices, glosarios, temarios, etcétera, están siendo asociados a los registros catalográficos originales, lo que potencia enormemente la búsqueda y descubrimiento de información, dado que el buscador no dispone ya solamente de unas pocas palabras del autor, título o tema, sino muchas palabras contenidas en el índice o glosario de cada libro. Algunas bibliotecas especializadas en literatura extraen y agregan todos los personajes, lugares, épocas y eventos consignados en novelas, obras de teatro y otras similares. Las bibliotecas de química agregan fórmulas, sustancias, compuestos, procesos industriales, etcétera, adaptándose al contexto y características de esa disciplina. Lo mismo puede hacerse en los demás campos de conocimiento.

Los datos masivos se encuentran ahora también en la biblioteca en los estudios métricos de la información documental, en todas sus especialidades: bibliometría, informetría y bibliotecometría, así como en otras asociadas: cienciometría, webmetría, altmetría y la emergente archivometría. 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, y se han convertido ya en otro ejemplo clásico de la minería de datos aplicada.

Y no tan solo en los estudios de metría: con la utilización de la minería de datos, técnicas lingüísticas, estadísticas, de aprendizaje de máquina, recuperación de información, comprensión del lenguaje natural, razonamiento basado en casos y otras más, los estudios de análisis de textos permiten a organizaciones y personas 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. Las aplicaciones de todo ello son muy variadas: identificación de textos y su correspondiente extracción de elementos; categorización y/o taxonomía de textos, extracción de ideas, temas, conceptos, entidades, relaciones y eventos; traducción de textos, reconocimiento óptico de caracteres, por citar algunas.

Un buen número de usos actuales de los datos masivos en bibliotecas y en toda la industria relacionada con Bibliotecas y Servicios de Información (Library and Information Services o LIS) caen dentro del campo de la Inteligencia Artificial o IA, con variadas aplicaciones en diversos subcampos dentro de ella. En primer lugar, se encuentra el "aprendizaje de máquina" para los más variados propósitos: indización, catalogación, clasificación, recuperación de información en línea, elaboración de resúmenes, servicios de referencia, tablas de contenido, etcétera. Muchas de estas organizaciones –bibliotecas, editores y afines– ya han empezado a construir aplicaciones prácticas de aprendizaje de máquina en muy diversas vertientes; por ejemplo, el análisis y síntesis de documentos. Consiste en programas que pueden "leer" un cierto texto y extraer información a partir de él. Como ya se mencionó, los sistemas de este tipo se construyen para documentos muy específicos –textos, imágenes, partituras, etcétera–, pues no existe todavía el "sistema interpretador" universal para todo tipo de ellos.

Igualmente, no existe a la fecha un sistema que pueda leer libros y construir sus fichas catalográficas completas a partir de ellos de forma sistemática y confiable, pero sí existen los que pueden extraer suficiente información coherente para proporcionar elementos valiosos para las personas, como los catalogadores, o para los sistemas, como los "descubridores de biblioteca" o *library discoverers*. Ésta se ha convertido en una de las aplicaciones que amerita mayor reflexión en las bibliotecas: se debate actualmente

si los catálogos deben seguirse construyendo al estilo "tradicional" o es necesario ya efectuar un cambio hacia nuevas estructuras de ordenamiento y recuperación documental soportadas por estos elementos (Bourg 2017).

Otro ejemplo práctico de programas de aprendizaje de máquina muy conocido en las bibliotecas son los sistemas de Reconocimiento Óptico de Caracteres (Optical Character Recognition u OCR), utilizados para interpretar texto que ha sido digitalizado en forma de imagen para convertirlo en formatos de texto interpretables por computadora: doc, odt, txt, pdf, rtf, y otros. Este tipo de programas pertenecen al campo de la IA, ya que su tarea consiste en leer e interpretar letras a partir de una forma gráfica, de la misma forma que lo realizan los seres humanos, y caen en el subcampo del "aprendizaje de máquina" debido a que estos programas pueden "aprender" lo que las personas les van indicando, como errores de interpretación, manchas en papel, fuentes tipográficas obsoletas y descontinuadas, así como las correcciones pertinentes.

Otro de los subcampos de la IA utilizado desde hace tiempo y con regularidad en las bibliotecas son los denominados "sistemas expertos". Ya despertaban el interés de los bibliotecarios desde los años ochenta, como puede verse en los numerosos textos al respecto desde esa época, los cuales trataban desde entonces la indización basada en el conocimiento, el procesamiento de lenguaje natural, la catalogación, la recuperación de información y la consulta, entre otros temas. Los sistemas expertos son programas informáticos que utilizan principios y métodos de la Inteligencia Artificial para resolver problemas dentro de un campo especializado, los cuales usualmente requieren de la experiencia de personal experto; de ahí el nombre. Incorporan los conocimientos técnicos acumulados por personas expertas en un tema y se diseñan para funcionar lo más parecido a ellas.

En general, contienen una *base de conocimientos* de hechos y relaciones representados en forma de datos y vínculos, y por supuesto tienen la capacidad de hacer inferencias basadas en ellos. Los diseñadores de estos sistemas utilizan diversas técnicas para la creación de esa base de conocimientos, como el análisis de

protocolos y procedimientos escritos, la descripción verbal de tareas realizadas por una persona, los cuestionarios, encuestas y entrevistas, la observación de procesos y su simulación, así como el descubrimiento y la documentación del conocimiento tácito dentro de la organización. Este último es un filón poco explotado y no obstante de gran valor en las bibliotecas, pues mucho del conocimiento de los bibliotecarios acerca de la gestión y explotación de información cae dentro de esta posibilidad. El conocimiento tácito o interno de los bibliotecarios es su conocimiento acumulado, generado por su experiencia e inherente al personal de biblioteca y que por lo general ha sido materializado mediante diferentes procesos al interior de ella.

Muchas otras aplicaciones prácticas en las bibliotecas actuales usan sistemas expertos y/o aprendizaje de máquina provenientes de la IA:

- Se mencionaron algunos ejemplos acerca de cómo algunas bibliotecas aprovechan y/o realizan estudios acerca de las formas en que los usuarios buscan y recuperan información para aprender más acerca de la lógica y las maneras que utilizan para acceder a la información, todo con el fin ulterior de mejorar sus catálogos internos, sus "descubridores" de información, sus OPAC, etcétera.
- Muchos de esos estudios tienen que ver específicamente con el "lenguaje natural" que los usuarios, como toda persona, utilizan para buscar con el fin de enseñar a los computadores a entender y descifrar ese lenguaje. Ahí se busca extraer los conceptos clave del lenguaje dentro de una pregunta y su posible solución a través del procesamiento por medio de la IA. Por su naturaleza, estas aplicaciones requieren de datos muy numerosos.
- Muchos programas informáticos de bibliotecas guardan información de búsquedas previas de los usuarios para "aprender" de éstas, con el fin de "personalizar" la página de cada uno de ellos, "recordando" lo que han buscado con anterioridad y estableciendo patrones. Con ella, el sistema

puede posteriormente hacer sugerencias al usuario como "las personas que consultaron este texto también consultaron estos otros…" o "este autor o tema se relaciona con este otro".

- Con el uso de estas técnicas de memorización y personalización, las páginas web de la biblioteca, se construyen de manera que se permite a cada usuario crear y guardar la forma, el aspecto y la distribución física de su página; los formatos para despliegue, y sus búsquedas anteriores, de manera que la personalización de la página de cada usuario pueda ser propia y distinta, a su gusto y conveniencia.
- Numerosas bibliotecas extraen ya datos de las redes sociales de sus usuarios interconectadas a los servicios de la biblioteca para detectar temas de tendencia (*trend topics*), contar "me gusta" y otros eventos similares acerca de sus servicios o informaciones; recibir sugerencias de adquisición de obras; verificar eficacia y dar seguimiento de sus servicios; medir "usabilidad" de nuevos servicios y opciones; detectar fallas o problemas; diseñar nuevos tutoriales, y muchos otros usos más.

Todos estos son solo algunos ejemplos para ilustrar las numerosas aplicaciones que las bibliotecas ya están utilizando por medio de la IA y sus subcampos de sistemas expertos, aprendizaje de máquina, etcétera; todas ellas basadas en el uso y la explotación constante y sólida de sus propios datos.

De todo lo anterior se concluye que son evidentes las innumerables ventajas y beneficios que las bibliotecas pueden obtener del manejo de datos, en especial los masivos. No obstante, como todas las tecnologías creadas por el ser humano, conlleva una cara negativa que es necesario que los bibliotecarios conozcan profundamente para neutralizarla. En primera instancia, los datos masivos son de difícil manejo, en parte debido a su propia naturaleza de inmenso volumen, velocidad, y variedad, y por otra parte debido a que hay un gran desconocimiento generalizado acerca de cómo manejarlos adecuadamente debido a la novedad del tema

y la gran escasez mundial de expertos en ellos. Ambos factores ocasionan frecuentemente un mal planteamiento de objetivos y técnicas, datos duplicados, inconsistencia o sesgo de ellos; pobre selección de herramientas de análisis, interpretaciones erróneas, etcétera, con las subsecuentes consecuencias negativas. A esto debe agregarse que el manejo de datos masivos presenta grandes retos técnicos y requiere de cierto presupuesto para sus proyectos. Obviamente todo esto es un inconveniente, pero como fue subrayado, el aspecto más negativo del uso de los datos masivos y el que merece mayor cuidado consiste en que introducen el riesgo de un posible abuso en la privacidad y la confidencialidad de datos personales.

Desafortunadamente, su uso implica con frecuencia colecta de datos personales, lo cual conlleva enormes riesgos a la privacidad. Cuando se compilan y almacenan grandes cantidades de datos –si se incluyen datos personales entre ellos–, se crea el riesgo de que sean usados para propósitos diferentes a los diseñados originalmente. Siempre existe la posibilidad de fugas de información del equipo de la organización, deliberadas o por error, que resulten en la extracción de datos por parte de terceros con intenciones aviesas.

Como se estableció, la privacidad de datos en un computador no es un simple problema de seguridad informática: conlleva además principios y cuestiones éticas, legislaciones y normas al respecto, gobernanza de los datos, responsabilidades técnicas y administrativas, y rendición de cuentas. La protección de datos es en esencia un problema técnico que implica asegurar los datos contra accesos no autorizados: quién y cómo los cuida. La privacidad de los datos va más allá: es una cuestión ética y legal que implica aspectos todavía más profundos en la organización: quién puede tener datos personales; durante cuánto tiempo, quién define a aquellos que los pueden acceder, quién puede accederlos autorizadamente y en qué circunstancias, quién los puede modificar, a quién y cómo pueden ser transferidos, etcétera. Por tanto, la protección de los datos es un requisito necesario mas no suficiente para lograr un fin mayor: la privacidad de los datos.

La privacidad no son sólo buenos deseos, obedece a principios internacionales derivados de la Declaración Universal de los Derechos Humanos. En años recientes, la Asamblea General de las Naciones Unidas adoptó varias resoluciones sobre el "Derecho a la privacidad en la era digital," conminando a todos los países a que "respeten y protejan el derecho a la privacidad". A partir de esto, muchos países y regiones han adoptado medidas al respecto, siendo la más avanzada e importante de ellas y modelo a seguir la "Regulación de Protección de Datos Personales" de la Unión Europea o GDPR de 2018. Este modelo es también uno de los aliados más valiosos para las bibliotecas en esta tarea, ya que incluye todas las regulaciones esenciales hacia los proveedores, las cuales fueron estudiadas con detalle en el apartado correspondiente. Se recomienda a las bibliotecas que se encuentran fuera de esa región que revisen siempre las condiciones de servicio de sus proveedores para verificar hasta qué grado ellos cumplen con este marco regulatorio. Además de ello, hoy en día más de cien países poseen algún nivel de legislación y regulaciones acerca de privacidad y protección de datos. México cuenta con una amplia legislación al respecto por la cual las bibliotecas son sujetos obligados.

En el campo de las bibliotecas, la ALA estableció claramente desde hace muchas décadas su postura al respecto, incluida en su "Interpretación de la Carta de los derechos a la privacidad". La IFLA también se pronunció al respecto desde hace muchos años en su "Declaración de IFLA sobre el Acceso a la información de Identificación Personal en los Registros Históricos", que toma entre sus principios la libertad de acceso a la información y la libertad de expresión en las que, de forma muy especial y explícita, se considera a la privacidad como parte indispensable de la salvaguarda de esos derechos.

Los principios mencionados hablan de privacidad y confidencialidad de los datos personales. Se aclaró la diferencia para su uso dentro de las bibliotecas: surgieron desde hace mucho tiempo con la privacidad, la cual significa que en una biblioteca el usuario tiene el derecho a leer y consultar lo que desee sin que los textos de su interés sean materia de escrutinio de terceros. La

confidencialidad se gesta cuando una biblioteca entra en posesión de datos personales que hagan identificable al usuario y, por tanto, debe tomar las medidas para evitar su acceso no autorizado. Por tanto, la confidencialidad es un proceso que protege, entre otras cosas, a la privacidad, y es un derecho de todo usuario. La confidencialidad es la obligación ética y legal que tiene la biblioteca de proteger ese derecho.

Estos principios que datan de hace casi un siglo, han servido como guía para innumerables bibliotecas en el mundo para construir políticas y mecanismos de privacidad dentro de sus servicios bibliotecarios y de información, así como las medidas pertinentes para mantener la confidencialidad de los usuarios. Tarea que fue relativamente sencilla con los procesos "tradicionales" en la biblioteca que colectaban datos personales, ya que estos eran destruidos a corto plazo. La consulta de catálogos, índices y otros impresos no dejaban huella asociada del usuario y sus intereses en esos tiempos.

Todas las bibliotecas que compilaban y usaban datos para estadísticas de uso de las colecciones lo hicieron siempre de forma anónima, también bajo principios de privacidad. Desafortunadamente, la globalización de datos en el mundo digital trajo la invasión masiva a la privacidad y a la confidencialidad de datos personales. Debido al auge de la recopilación a gran escala de este tipo de datos, la vigilancia electrónica y la intercepción de comunicaciones digitales, la libertad de información, la privacidad y la confidencialidad se han visto seriamente amenazadas en las últimas décadas. Esto es especialmente sensible en los datos masivos, y de ahí el aspecto negativo de su uso en las bibliotecas. Es evidente que este no es un problema exclusivo de ellas, pero sin duda las afecta en mayor grado y por lo mismo requiere de su cuidado y atención.

La colecta, transferencia y venta de datos personales se ha convertido en un gran negocio y esto ha introducido en las bibliotecas y sus usuarios serios riegos al respecto de la privacidad, pues implica la recopilación de datos de ellos por parte de proveedores de servicios de información, de servicios en la nube, de aplicaciones para teléfonos, lo cual sucede a menudo de forma subrepticia o

disfrazada. La adquisición por parte de la biblioteca con publicaciones y servicios de información digitales y en red: consulta de obras, descubridores, servicios de búsqueda y documentación, tablas de contenido, etcétera, provenientes de proveedores comerciales externos a la biblioteca, introduce por tanto serios riesgos a la privacidad, ya que con demasiada frecuencia ellos pretenden que sus requerimientos de captación de datos personales sean aplicados en las bibliotecas. Obviamente esto es del todo inaceptable, pues contradice totalmente los antiguos principios bibliotecarios acerca de la privacidad de sus usuarios.

Sin ser una tarea fácil, tampoco es imposible, y existe solución para ello. En efecto, es un problema agravado por la tecnología, pero su solución no se basa en ella: al igual que muchos otros problemas en la biblioteca, es cuestión mayormente de método y procedimiento. Como ya fue establecido, el problema de la protección de datos –la parte tecnológica– debe ser atendido, pero es sólo un componente menor. La parte sustantiva de ello debe incluir aspectos todavía más amplios, contemplados en las técnicas y principios de la "gobernanza de datos", y que fueron analizados en el apartado correspondiente. A partir de ellos, cada biblioteca puede y debe construir sus "políticas de privacidad y protección de datos" propias de su contexto y entorno, las cuales permiten construir una sólida base para la planificación y las acciones destinadas a proteger la privacidad y los datos personales de sus usuarios. Estas políticas deben incluir las cuestiones éticas y legales y ser consistentes con las leyes locales para conformarse como el marco de referencia de la organización, y deben establecer los rubros generales acerca de las personas que son y serán responsables de definir las cuestiones acerca de privacidad y datos personales al interior de ella, redactar y actualizar los planes y programas institucionales al respecto; supervisarlos, acceder a los datos, transferirlos y resguardarlos. Las políticas abarcan a toda la organización integralmente y son documentos a nivel teórico, y por eso tenderán a ser mucho más estables en el tiempo.

Partiendo de las políticas como base, la biblioteca debe desarrollar procedimientos, guías, buenas prácticas, estándares, etcétera.

Estas son las versiones prácticas que instrumentan los conceptos esbozados de manera teórica y general en las políticas, y detallan acciones preestablecidas y secuenciales que abarcan toda la variedad de procesos y departamentos de la biblioteca. Las políticas establecen el por qué, el qué y el quién; los procedimientos y guías establecen el cómo, cuándo, dónde y, en su caso, detallan los quiénes. La experiencia acumulada irá generando las "buenas prácticas" y los estándares. Es conveniente también diseñar y realizar "auditorías" de todos los servicios que ofrece la biblioteca, tanto internas como a través de proveedores, para asegurarse de que todos los puntos han sido cubiertos; especialmente en proyectos de datos masivos. En el capítulo correspondiente se detallaron esos puntos.

Se enfatizó en el hecho de que siempre que sea posible, la biblioteca debe anonimizar al máximo los datos usados por ella para sus proyectos, estadística o retroalimentación: los datos que no se tienen no pueden fugarse ni ser extraídos. En todo servicio o aplicación que diseñe o construya la biblioteca, debe siempre reflexionarse previamente acerca de cuáles datos son indispensables para el mismo y cuáles no: en la gran mayoría de casos sucede que los proyectos funcionan sin incluir datos personales o con un mínimo de ellos. No debe recabarse este tipo de datos para ningún proyecto o servicio si no es indispensable. En la gran mayoría de procesos de la biblioteca que recaban datos para análisis –como es el caso de las estadísticas de uso de las colecciones de una biblioteca– es totalmente factible compilarlos sin registrar datos sensibles de los usuarios.

Como fue visto a lo largo del apartado correspondiente, al minimizar en la biblioteca el número de puntos donde se manejan datos personales, la cantidad de lugares a proteger se reduce sensiblemente, lo cual facilita enormemente la tarea a los responsables de ella. Como ha sido reseñado, existen metodologías relativamente sencillas para reducir e inclusive eliminar el uso de datos personales a lo largo de muchas de las aplicaciones y servicios de la biblioteca, sin duplicar registros de datos personales a lo largo de cada departamento, sección, servicio o proyecto en la biblioteca.

Como es evidente, el conjunto de políticas, procedimientos, guías, estándares, auditorías y metodologías aquí consignadas confirman el principio enunciado inicialmente de que la mayor parte de todo ello consiste en método y procedimiento, y solo una mínima parte son elementos tecnológicos. Lo más importante sigue siendo *cómo* se hacen las cosas, y no *con qué* tecnología se hacen.

No obstante, las herramientas tecnológicas son indispensables para el manejo de los datos, especialmente los masivos. Por este motivo se desarrolló un capítulo completo con varios apartados dedicado al estudio de ellas: sistemas, programas y aplicaciones informáticos, algoritmos, metodologías, etcétera, tanto comerciales como de acceso abierto. Dado que no existe el gran sistema informático universal que abarque integralmente todas las necesidades al respecto, existen numerosas herramientas puntuales de alta especialidad para cada tipo de propósito, producidas además por una infinidad de fabricantes. Como los proyectos de datos no son iguales entre sí y cada uno tiene su contexto y características propias, es necesario integrar un cierto conjunto de herramientas metodológicas e informáticas para su desarrollo y solución en cada uno de ellos. De ello se desprende que uno de los conocimientos primordiales al iniciarse en el campo de los datos masivos consiste en adquirir una buena idea general de todas las variedades y capacidades de esas herramientas, para así estar en capacidad de poder seleccionar la adecuada para cada proyecto y necesidad.

Por su amplitud en funciones generales, algunas herramientas se han vuelto de uso común para ciertos sectores: comercial, transportes, de la salud, educativo y, por supuesto, también las hay para uso en bibliotecas y otras organizaciones de información. Debido a su alto número, ninguna persona conoce o domina todas las herramientas existentes, pero no obstante es necesario hoy en día que todo profesional de la biblioteca conozca en términos generales la oferta, las posibilidades y capacidades de aquellas, para así poder aproximarse a una eventual selección de alguna de ellas, de la misma forma que todo bibliotecario lo hace al seleccionar un sistema para la automatización o la gestión de su biblioteca (ILS), un buscador o un descubridor

especializado, un catálogo automatizado u OPAC, y otros semejantes. Actualmente, se requiere que el responsable de la biblioteca adquiera un conocimiento panorámico y general de las herramientas para el tratamiento de datos.

Debido al crecimiento inusitado de la gestión de datos a nivel mundial en prácticamente todos los sectores, se producen innumerables productos, sistemas, servicios y aplicaciones alrededor de los datos que ya están siendo desarrolladas por grandes empresas y organizaciones. En particular, varias importantes instituciones bibliotecarias ya han desarrollado y puesto a disposición de bibliotecas más pequeñas una serie de herramientas que pueden aprovechar en su beneficio con mínima o nula inversión. Se mencionaron los importantes ejemplos del Servicio de Datos Enlazados de la Biblioteca del Congreso de Estados Unidos, el de la Biblioteca Británica; la Red de la Biblioteca Nacional de Medicina (NNLM) también de ese país; las guías de las Bibliotecas del Instituto Smithsoniano para la creación y depósito de datos en repositorios, etcétera. Esto se ha ido convirtiendo gradualmente en una importante fuente para la adquisición de herramientas en este sentido. Además de lo anterior, muchos de los proveedores usuales de las bibliotecas también están haciendo desarrollos al respecto de los datos masivos, con los cuales las bibliotecas pueden hacer alianzas estratégicas para su aprovechamiento e impulso, o simplemente adquirir estos productos en condiciones preferenciales. Se mencionaron al respecto los estudios textuales de OCLC o el proyecto SN SciGraph de datos abiertos enlazados de la división de Ciencias Naturales de Springer Nature.

Existen también múltiples opciones de herramientas específicamente construidas para bibliotecas, provenientes de numerosos proveedores comerciales dedicados a alguna de las facetas del proceso y análisis de los datos. Se analizó el suministro de ello en forma de servicios bajo el esquema comercial conocido como la nube, del cual se presentó un resumen de sus principales modalidades para su conocimiento básico: "*Software* como servicio" (SaaS), "Plataforma como servicio" (PaaS), "Infraestructura como servicio" (IaaS), y por supuesto, "Datos como Servicio" (DaaS) y

"Datos y Plataforma como Servicio" (DaPaas). Se destacó el hecho de que todas las grandes plataformas tecnológicas comerciales existentes hoy en día –Google, Amazon, Microsoft, Apple, Facebook, IBM, Oracle, etcétera– ofrecen algún conjunto de servicios relacionados con los datos masivos; unos gratuitos y otros de paga. Esta segunda vertiente de servicios en la nube es cada vez más utilizada, ya que permite a los usuarios adquirir grandes recursos con relativamente menores inversiones; no obstante, debe usarse con suma cautela por las razones de seguridad y privacidad ya enunciadas. Las bibliotecas han aprovechado muchas de estas herramientas y servicios desde hace ya algunos años, y las han incorporado a sus quehaceres con infinitas posibilidades.

Para su revisión, las herramientas fueron divididas arbitrariamente en los siguientes tipos:

- Los manejadores de bases de datos, tanto SQL como NoSQL.
- Los manejadores de datos documentales.
- Las herramientas de "normalización" y mapeo de datos.
- Las herramientas para el análisis de datos masivos, con el fin de extraer patrones o tendencias de ellos.
- Las herramientas de visualización, interpretación o presentación de resultados.
- Las herramientas de Inteligencia Artificial.
- Herramientas de aplicación específica, como las mencionadas para mezclas, etcétera.

Dado que en tiempos recientes se ha estado creando un especial énfasis en el campo y las herramientas para el análisis o analítica de datos (*data analytics),* se dedicó un apartado especialmente a ello. Esta especialidad contiene el núcleo de gestión de los datos; esto es, detectar tendencias y patrones en ellos para plantear soluciones, pero más puntualmente, se refiere a la tarea de identificar cuáles variables de la organización pueden ser relacionadas con ciertos datos y establecer así correlaciones para el planteamiento de preguntas y la eventual obtención de respuestas a través de técnicas específicas. Como fue mencionado, este tema debe ser estudiado con cautela y

comprendido en toda su amplitud. Obviamente el propósito principal de toda la ciencia y la gestión de los datos consiste en poder extraer información valiosa y útil de sus conjuntos, pues su simple colecta y proceso tienen poco valor en sí mismos: su beneficio se materializa cuando se analizan. Por tanto, no hay duda de que el análisis de datos es la parte central y más importante de la ciencia y la gestión de los datos, en especial los masivos, ya que ahí es donde se extraen las soluciones y decisiones relevantes para una organización, y por ello merece especial atención y estudio.

En consecuencia, existen más herramientas, textos y productos al respecto en comparación con otras áreas y herramientas del manejo de datos. Pero es indispensable contextualizar y balancear este campo junto con los demás: no hacerlo pone en riesgo el resultado y propósito ulterior de la gestión de datos. De la observación de las innumerables empresas, sitios web, las ofertas de herramientas de *software*, capacitación y asesoría, cursos, posgrados y diplomados, certificaciones, etcétera, queda en el lector la percepción de que esto es lo que debe atenderse con respecto a la ciencia y la gestión de los datos: si la organización resuelve el capítulo del análisis de los datos, todo su problema queda resuelto.

Ello tiene su explicación en la ya mencionada grave escasez de expertos a nivel mundial acerca de la ciencia de los datos y su gestión. Claro que las organizaciones requieren formar y capacitar personal al respecto, pero es indispensable formarlo y entrenarlo equilibradamente en todas las áreas de la ciencia y la gestión de los datos. No es conveniente desarrollar solo uno de sus campos por más central que sea, pues ello introduce un desbalance nocivo y peligroso para el propósito final. A ninguna organización le es conveniente tener únicamente expertos en análisis de datos si carece de personal que pueda contender con los demás aspectos de su gestión. Igualmente a nivel personal, no se recomienda formarse únicamente en análisis de datos; si no se cuenta con otros expertos en las demás áreas de la institución, sus conocimientos y habilidades serán fútiles para una gestión exitosa de los datos al no existir un contexto y un balance de los mismos dentro de la organización.

Como conclusión a lo anterior, puede afirmarse que las organizaciones y las personas deben considerar al análisis de datos como una de las múltiples opciones de capacitación y entrenamiento, pero de ninguna manera como la única opción a desarrollar a nivel corporativo o personal: siempre debe ponderarse toda la oferta en contexto. Recuérdese que en la actualidad existe un gran modelo comercial alrededor de la gestión de datos, y en especial en la capacitación, asesoría y certificaciones al respecto. Se ha vuelto un negocio millonario por parte de múltiples empresas ofertantes; no es que ello sea malo en sí mismo, pero derivado de ello se ha ido introduciendo artificialmente una enorme expectativa alrededor del campo más atractivo del análisis de datos, en donde buena parte de ello no atiende al contexto general de lo que las organizaciones o las personas realmente necesitan, y no todos los ofertantes son serios. Muchos venden porque está de moda y deja buenos dividendos aunque lo que ofrezcan no sea de calidad o realmente útil. De ahí la capital importancia de ponderar toda la oferta y con quién se obtendrá la capacitación.

El punto central de todo lo anterior sigue siendo que no debe nunca perderse de vista que para solucionar un problema de datos el análisis de ellos no es suficiente; se requiere que concurran integral y equilibradamente todas las partes de la gestión de datos. El proceso de los datos comprende muchos pasos; si bien no todos ellos se encuentran en cada uno de los problemas relacionados, siempre se requieren varias etapas para su adecuada gestión: diseño, modelado y tipificación de datos, captura, colecta o minería de ellos, su codificación y depuración; normalización y estructuración, inclusión de metadatos; transformación; agregación o desagregación de datos, validación; almacenamiento y preservación; visualización y despliegue de los datos, y sí, finalmente su análisis e interpretación. Por tanto, un buen analista de datos no tan solo sabe y puede analizarlos; también debe ser capaz de participar integralmente en todas las demás etapas.

Más específicamente, se revisaron los temas y usos del análisis de datos dentro de las bibliotecas. Se reseñó lo compilado por diversos autores al respecto: desarrollo de colecciones; diseño de

nuevos servicios; conocimiento de los usuarios, así como sus preferencias y necesidades; demostrar cuantitativamente el valor e impacto de la biblioteca dentro de su entorno; mejorar la experiencia educativa e informativa de los usuarios; el análisis del aprendizaje, descubrimiento del conocimiento y demás usos para optimizar la toma de decisiones, tanto a nivel de datos masivos como en menor escala. Se destacó la importancia de que antes de analizar datos, la biblioteca debe asegurarse de seleccionar aquellos que en efecto representan ciertos hechos o realidades para poder establecer y obtener las correlaciones correctas que conducen a preguntas y respuestas reales y pertinentes.

Uno de los errores que se cometen con más frecuencia consiste en utilizar los datos que se pueden obtener más fácilmente, simplemente porque son fáciles de obtener, o utilizar forzosamente los datos típicos o tradicionales de la biblioteca. El buen análisis de datos implica que su selección debe tener de inicio una estrecha correlación con los hechos a representar; de otra forma, los resultados serán sesgados o irrelevantes. Esto no es fácil aunque parezca obvio; requiere por un lado de un profundo conocimiento de las causas y efectos de la función o tarea de la biblioteca que se desea representar, y por el otro, de un amplio dominio acerca de cómo plantear y diseñar correlaciones específicas de datos. Ello implica dos especialidades: el bibliotecario experto en esa función o tarea de la biblioteca y el analista experto en datos; la situación ideal se produce cuando estas dos especialidades las tiene una misma persona: un bibliotecario con experiencia en su área y además con buen conocimiento en el análisis de datos; no obstante, ello no es indispensable. Sin esta función dual correctamente aplicada, es fácil caer en la obtención y uso de datos demasiado simplistas para la tarea, o perderse en un mar de datos de los cuales la mayoría son irrelevantes o poco representativos de la función. Por ello la importancia de planear de inicio y con exactitud qué se va medir, por qué y cómo. Las áreas de la biblioteca donde el análisis de datos puede aplicarse son numerosas y debido a ello los posibles puntos de extracción y colecta de datos lo son todavía más: búsquedas en catálogos y colecciones, consultas, préstamo de libros, vistas y descargas de

documentos, acceso con clave a ciertos servicios, redes sociales entre la biblioteca y sus usuarios, por mencionar algunos. He ahí por qué es indispensable de inicio hacer un correcto diseño de ellos y sus hechos asociados para estar en posibilidad de hacer un análisis pertinente. El beneficio de todo ello en las bibliotecas ha sido tratado y descrito por numerosos autores, una inmensa mayoría de ellos editados por la American Library Association (ALA), lo cual da cuenta de la importancia actual del tema. En ellas se consigna también con detalle las habilidades y conocimientos que los bibliotecarios deben adquirir al respecto: fundamentos de conceptos estadísticos; fuentes de datos recomendadas para diversas funciones y procesos de biblioteca, así como orientación acerca de cómo usarlas; técnicas para limpieza de datos; cómo encontrar las correlaciones de los datos con los métodos apropiados de análisis al respecto, y cómo visualizar los resultados. Muchas otras asociaciones y organizaciones bibliotecarias han manifestado también su interés por estudiar, difundir y capacitar acerca del tema del análisis de datos para los profesionales de la disciplina. Muchos de los autores al respecto se refieren ya al análisis de datos aplicado en bibliotecas más puntualmente como "análisis bibliotecario" (*library analysis*) para nombrar a esta importante actividad y hacer evidente que ya es una especialidad dentro de las bibliotecas.

Dada la cada vez mayor importancia del manejo de datos en una gran cantidad de negocios y organizaciones, se destacó la importancia de que sean administrados adecuadamente: de lo contrario, los datos pueden causar más problemas que beneficios. He ahí la importancia del tema de la gobernanza de los datos, en especial los masivos. Se estableció al respecto que la gobernanza de datos tiene como objetivo garantizar la calidad y seguridad de los datos utilizados en una organización, estableciendo y supervisando un conjunto de políticas, procesos, puestos, normas, métricas y responsabilidades que aseguren el uso eficaz, eficiente, seguro y coherente de esos datos. Se destacó la importante diferencia entre la gobernanza de los datos y la gestión de ellos. La gestión de datos es un concepto mucho más amplio que describe todos los procesos utilizados para planificar, especificar, habilitar, crear, ad-

quirir, mantener, utilizar, archivar, acceder, controlar, depurar y gobernar los datos. La gobernanza de los datos es una parte de su gestión que trata las funciones, los procesos, los estándares, las métricas y las responsabilidades personales para establecer con claridad la propiedad y el acceso de los activos de datos al interior de la organización, su uso protegido, su coherencia, así como la rendición de cuentas.

La gobernanza de los datos crea un marco de referencia de los datos dentro de una organización; la gestión de los datos conlleva la ejecución práctica de ese marco. Esencialmente, la gobernanza de los datos busca que en lo tocante a ellos, las personas adecuadas tengan las responsabilidades correctas por medio de procesos, normas y métricas pertinentes. Por su misma naturaleza, la gobernanza de los datos conduce a ventajas que por lo general no son obtenibles solo a través de su gestión: proporcionar una visión, comprensión y terminología generalizada y coherente de los datos; optimizar procesos, procedimientos, funciones y puestos dentro de la organización; mejorar la calidad de los datos y garantizar su exactitud, integridad y consistencia, así como optimizar su ubicación en las áreas fundamentales; permitir el cumplimiento cabal de los requisitos legales mínimos exigibles, en especial aquellos relacionados con la privacidad y la protección de datos personales; prevenir conflictos y/o traslapes entre los distintos conjuntos de datos, y ahorrar costos de operación. Básicamente, la gobernanza de datos se diseña a partir del establecimiento de una política organizacional creada específicamente al respecto. Los puntos fundamentales de la gobernanza de datos son todos conceptos a definir y establecer, de los cuales se desprenderán acciones y procesos. Es algo deseable desde el principio, pero se volverá imperativo en la medida que las organizaciones vayan captando cada vez más datos: su simple administración ya no será suficiente.

Como ha podido percibirse a lo largo de todo este texto, uno de los puntos más importantes para el éxito de la gestión de datos tiene que ver con los recursos humanos dedicados al efecto. Su tratamiento –en especial los datos masivos– es una tarea multidisciplinaria, y por tanto involucra a una gran diversidad de

profesionales de la información, entre los cuales obviamente deben estar los bibliotecarios, dada su formación y experiencia tradicionales. Como ha sido establecido, el correcto tratamiento de los datos es una tarea altamente especializada que requiere de nuevos conjuntos de conocimientos, habilidades, experiencia y actitudes. En la actualidad existen relativamente pocos expertos en este tema a nivel mundial con capacidades adecuadas acerca de cómo diseñar, definir, colectar, depurar, transformar, analizar y presentar proyectos y estructuras de datos; la carencia de estas capacidades técnicas tiene como consecuencia proyectos de gestión de datos mal conceptualizados o la inacción total. Por tanto, es un campo emergente de desarrollo profesional de gran importancia en el gremio.

Como fue establecido, en términos generales los bibliotecarios necesitan conocer los fundamentos y principios de la ciencia de los datos, su gestión y curaduría; dónde se pueden encontrar los datos masivos y cómo aprovecharlos; la bibliominería, el análisis y visualización de datos, así como la selección y aplicación de herramientas informáticas para su explotación, todo ello como una nueva parte de su formación y experiencia integral en la disciplina.

Al margen de la eventual formación profesional, y en estrecha relación con el tipo de biblioteca u organización en la que laboren, es recomendable que los bibliotecarios vayan adquiriendo y profundizando el conocimiento y experiencia en el tratamiento de los datos, para estar así en posibilidades de desarrollar proyectos en los muy diversos campos de aplicación dentro de las bibliotecas que fueron analizados a lo largo del texto: reaprovechamiento de datos de investigación; su curaduría y repositorios; nuevos productos y servicios en las bibliotecas; mejora de eficiencia educativa escolar; estudio y diseño de nuevas taxonomías y esquemas de metadatos para búsqueda y recuperación; bibliometría y análisis de textos; incorporación de ventajas competitivas en las empresas en general; nuevos productos y servicios en empresas de Servicios de Información Bibliotecarios (LIS); la utilización de datos en humanidades digitales, ciencias sociales digitales, computación social; optimización de la administración y recursos bibliotecarios, y, obviamente, la do-

cencia y educación continua acerca de los datos. Se resalta el hecho de que la anterior es una lista indicativa y no exhaustiva de los campos de aplicación de los datos en las bibliotecas.

Como pudo percibirse a lo largo del texto, los capítulos y apartados presentaron una amplia posibilidad de proyectos de datos en las bibliotecas y organizaciones afines; indudablemente el campo de acción en ellas es sumamente amplio. El punto central consiste en que en cada una de esas eventuales aplicaciones reseñadas se requiere de profesionales con conocimientos, habilidades, actitudes, experiencias y dominio de herramientas específicas relativas a los datos. Es evidente que ninguna biblioteca podrá tener especialistas en todos y cada uno de los posibles campos de acción, pero también debe quedar claro que no puede carecer totalmente de personal especializado en ello. Por tanto, las bibliotecas pueden y deben ir formando sus especialistas de datos en los campos de aplicación propios de su contexto e interés. De inicio porque eso les permite mantenerse en las necesidades y circunstancias actuales del mundo de la información para poder seguir siendo competitivas e interesantes para sus comunidades y sus financiadores. Además, porque ello permite formar profesionales altamente demandados en los tiempos actuales, los cuales son escasos a nivel mundial, y eso significa nuevos y mejores puestos de trabajo para los bibliotecarios profesionales actuales y los estudiantes de la carrera.

Se reseñaron con detalle lo que los principales autores han señalado como los principales "roles" o funciones principales en esta tarea a partir de necesidades del mundo real para puestos de trabajo reales. Tres funciones "tradicionales": analista de datos, ingeniero de datos y periodista de datos, y además otras tres relacionadas estrechamente con las disciplinas de nuestro interés: bibliotecario de datos, archivista de datos y gestor/curador de datos. Los autores también establecieron los cinco aspectos fundamentales que estos profesionales de la información deben desarrollar al respecto: formación académica; experiencia práctica; conocimiento, familiaridad y comprensión de los temas; habilidades de ejecución, y competencias de dominio de herramientas y tecnologías. Como

fue mencionado a lo largo de este texto, su intención fue ilustrar en el conocimiento, la familiaridad y la comprensión de los temas, e indicar las habilidades y competencias.

Dentro de la inmensa oferta de cursos, especialidades, certificaciones, etcétera, ofrecidos comercialmente por innumerables empresas con respecto al tema de los datos, existe una cantidad pequeña pero aceptable de cursos, textos y capacitaciones sobre el tema ofrecidos sin fines de lucro por diversas organizaciones, los cuales son ideales para que los bibliotecarios puedan iniciarse en este campo. Se insiste en el hecho de que debe tenerse mucha cautela a la hora de seleccionarlos, pues numerosas empresas y organizaciones –muchas de ellas de gran prestigio– ofrecen cursos supuestamente gratuitos para "enganchar" a los eventuales alumnos, pero un análisis ulterior descubre que hay costos ocultos subyacentes, algunos de ellos muy altos; la mayor parte derivados de la emisión de certificaciones de los cursos. Como en todo aprendizaje, se recomienda ir de lo general a lo particular, y de lo básico a lo especializado. Por ello se sugiere elaborar un plan institucional y/o personal de capacitación y aprendizaje en datos, con evolución y profundización gradual. La oferta es indudablemente mayor en idioma inglés que en español, por lo que deben ponderarse con cuidado las opciones de cursos y textos en uno u otro idioma.

La gestión de datos, en especial los masivos, ha dejado ya de ser un tema tecnológico "de moda" para convertirse en toda una realidad emergente. Si bien siguen existiendo enormes mitos y exageraciones acerca de su utilidad, es indudable ya que en efecto pueden ser utilizados de manera sistemática para beneficio de las organizaciones, entre ellas las bibliotecas. Las principales organizaciones bibliotecarias del mundo ya han señalado esta importancia y han hecho numerosos estudios y recomendaciones al respecto. Es un hecho que los campos de aplicación de datos dentro de las bibliotecas son muy variados. Debido a su creciente auge e importancia dentro del mundo de la información, es un tema que no puede ni debe ser soslayado por las bibliotecas, manteniéndose en una zona de confort al margen de ellos. En efecto significa nuevos esfuerzos y reacomodos, gastos y molestias; pero

de ninguna manera es un pequeño agregado casual o un capricho que eventualmente puede ser adoptado por las bibliotecas como curiosidad técnica; representa a la vez un enorme reto y una gran oportunidad: por un lado implica que su personal bibliotecario debe adquirir nuevos conocimientos, habilidades, experiencia y actitudes para su correcto manejo, pero por el otro representa nuevas e inmensas oportunidades para reposicionar a la biblioteca dentro de las responsabilidades y quehaceres contemporáneos de su comunidad. Por esta razón requiere de estructura organizacional y personal calificado para realizar la tarea adecuadamente, como muchas de las otras tareas sustantivas de la biblioteca.

Finalmente, se resalta de nuevo que en los proyectos de datos la principal clave del éxito no está en las herramientas informáticas –por muy importantes que estas sean–, sino en el personal calificado. Por esta razón, hoy en día es altamente recomendable que los bibliotecarios profesionales comiencen a buscar una cierta formación, adiestramiento y experiencia en la ciencia de los datos, su gestión, su análisis, sus usos y aplicaciones, su curaduría, etcétera. No es simplemente una curiosidad académica; es en su propio beneficio y el de su institución. El mundo de la información sigue evolucionando, y por eso es indispensable que las bibliotecas y su personal lo hagan también. Como en muchos otros aspectos de tecnología y bibliotecas, el tratamiento sistemático y profesional de los datos ha superado ya la etapa inicial de auge y moda, y se ha convertido en una realidad que sin duda alguna debe ser estudiada, incluida y aprovechada seriamente en el medio. Las bibliotecas no deben ni pueden esperar a ser un campo añejo de desarrollo para considerar su inclusión en ellas, como ha sucedido con otros desarrollos tecnológicos.

Jeffrey Stanton hizo una reflexión desde 2012 sobre el tema de datos y bibliotecas que sigue totalmente vigente al día de hoy, y es toda una editorial al respecto:

> […] Un bibliotecario no necesita convertirse en programador, pero todo bibliotecario interesado en la creación de conocimiento debería tener una familiaridad esencial con la forma en que las

> diversas herramientas de *software* pueden transformar los datos. Un bibliotecario no necesita ser un ingeniero de bases de datos, pero todo bibliotecario debe entender los fundamentos de las herramientas de recuperación de información. Un bibliotecario no necesita ser un estadístico, pero todo bibliotecario debe tener una clara comprensión de la forma en que los resúmenes descriptivos y las pruebas básicas de los datos numéricos pueden ser utilizados y mal utilizados. Por último, un bibliotecario no necesita ser diseñador gráfico, pero todo bibliotecario debe reconocer las características de las presentaciones de datos eficaces. En resumen, para cumplir sus misiones, los bibliotecarios pueden ejercer una variedad de sofisticadas habilidades que ocupan el terreno central entre la comprensión de las necesidades del usuario de la información en un extremo y la conservación de los datos en el otro (Stanton 2012).

Jesse Shera (1973, en Wright 2013, 47) estableció desde hace muchos años que "la ciencia de la información se ocupa exclusivamente de la transmisión de señales, mientras que la bibliotecología se basa en las interacciones humanas y se ocupa de las ideas y el conocimiento, además de la información". De acuerdo con esta visión, los bibliotecarios profesionales contemporáneos pueden ofrecer grandes e importantes aportaciones a la ciencia de los datos desde su muy particular perspectiva por sus capacidades técnicas y su experiencia profesional, pero sobre todo, por su formación humanística: una oportunidad que no debe ser desaprovechada.

Referencias bibliográficas

Todas las referencias electrónicas han sido verificadas como exactas y existentes al 31 de marzo de 2021.

Acharjya, D.P.; Ahmed, Kauser. 2016. "A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools". *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 7, núm. 2: 511-518 https://www.researchgate.net/publication/296550027_A_Survey_on_Big_Data_Analytics_Challenges_Open_Research_Issues_and_Tools.

Affelt, Amy. 2015. *The Accidental Data Scientist*. Medford, N.J.: Information Today. 223.

American Libraries Association (ALA). 1939. *Bill of Rights*. Adoptado por el Consejo de la ALA junio 19, 1939; enmendado octubre 14, 1944; junio 18, 1948; febrero 2, 1961; junio 27, 1967; enero 23, 1980; enero 29, 2019. http://www.ala.org/advocacy/intfreedom/librarybill.

———. 2002. *Privacy: An Interpretation of the Library Bill of Rights*. Junio 2002. http://www.ala.org/advocacy/intfreedom/librarybill/interpretations/privacy.

———. 2007. ALA Privacy Tool Kit,: 8-9. http://www.ala.org/advocacy/privacy/toolkit.

Alvares, Lilian; Araújo Jr., Rogério. 2010. "Marcos históricos da ciência da informação: Breve cronologia dos pioneiros, das obras clássicas e dos eventos fundamentais". *TransInformação*, Campinas, Brasil, vol. 22, num. 3, set./dez. 2010: 195-205 http://www.scielo.br/pdf/tinf/v22n3/a01v22n3.pdf.

Alvaro, Elsa; Brooks, Heather y Ham, Monica. 2011. "E-science librarianship: Field undefined". *Issues in Science and Technology Librarianship*, num. 66 (verano). http://www.istl.org/11-summer/index.html.

Andaur, Gabriela. 2016. "Panorama de la Gestión de Datos de Investigación en América Latina y el Caribe". Blog del proyecto Learn. Entrada del 16 de mayo de 2016 http://learn-rdm.eu/es/gestion-de-datos-de-investigacion-en-america-latina/.

Artificial Intelligence and Machine Learning in Libraries. 2018. Jason Griffey (Ed.). *Library Technology Reports*, ALA Techsource, vol. 55, núm. 1. DOI: https://doi.org/10.5860/ltr.55n1.

Asociación Española de Usuarios de las Telecomunicaciones y la Sociedad de la Información. *Manual de Ciudades Digitales*. 2012. https://nuevasciudadesdigitales.wordpress.com/manual-ciudades-digitales/.

Austin, Sidley. 2016. "Top Ten Data Protection and Privacy Issues to Watch in 2016". *Lexology*, enero 11, 2016. http://www.lexology.com/library/detail.aspx?g=6f3d4d67-ba04-42a9-9131-28bbd4a13f9f.

Ávila, Eder. 2020-1. "Integración de los principios de 'linked data' en el registro bibliográfico": 75-94. *El manejo de datos. Aproximación desde los estudios de la información*. 2020. Araceli Torres (Coord.) México: IIBI-UNAM. http://ru.iibi.unam.mx/jspui/handle/IIBI_UNAM/L225.

——— . 2020-2. *Los Datos enlazados y su uso en bibliotecas*. México: UNAM. Instituto de Investigaciones Bibliotecológicas y de la Información http://ru.iibi.unam.mx/jspui/bitstream/IIBI_UNAM/56/3/datos_enlazados.pdf

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

Bailey, Charles. 1991. "Intelligent Library Systems: Artificial Intelligence Technology and Library Automation Systems". *Advances in Library Automation and Networking*, vol. 4, pp.1-23 http://eprints.rclis.org/4891/.

Baker, Ryan. 2015. *Big Data and Education*. Nueva York: Teachers College, Columbia University. http://www.columbia.edu/~rsb2162/bigdataeducation.html.

Ball, Rafael. 2019. "Big Data and Their Impact on Libraries". *American Journal of Information Science and Technology,* vol. 3, num. 1: 1-9. DOI: 10.11648/j.ajist.20190301.11.

Bell, Arthur. 1947. *Christian Huygens & the Development of Science in Seventeenth Century*. Londres: E. Arnold & Co.

Bell, Steven. 2013. "Promise and Problems of Big Data". *Library Journal*. March 13, 2013. http://lj.libraryjournal.com/2013/03/opinion/steven-bell/promise-and-pro.

Berdondini, Andrea. 2019. "The Information Paradox". Blog Towards Data Science, entrada del 9 de julio de 2019. https://towardsdatascience.com/the-information-paradox-38a411517f15.

Berendt, Bettina; Littlejohn, Allison; Kern, Philippe; Mitros, Piotr; Shacklock, Xanthe y Blakemore, Michael. 2017. "Big data for monitoring educational systems". Publications Office of the European Union, Luxembourg. https://doi.org/10.2766/38557.

Bieraugel, Mark. 2013. "Keeping up with… Big Data". *American Library Association – Association of College and Research Libraries*. Sitio de la AL. http://www.ala.org/acrl/publications/keeping_up_with/big_data.

Blummer, Barbara; Kentin, Jeffrey. 2018. "Big Data and Libraries: Identifying Themes in the Literature". *Internet Reference Services Quarterly*, vol. 23, núm. 1-2, 15-40, DOI: 10.1080/10875301.2018.1524337.

Borgman, Christine 2007. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: MIT Press.

Bourg, Chris. 2017. "What happens to libraries and librarians when machines can read all the books?" *Feral Librarian* (blog). Entrada del 16 de marzo, 2017 https://chrisbourg.wordpress.com/2017/03/16/what-happens-to-libraries-and-librarians-when-machines-can-read-all-the-books/.

Brinton, Willard. 1914. *Graphic Methods for presenting facts.* Nueva York: The Engineering Magazine Company https://archive.org/details/graphicmethodsfo00brinrich/page/n13/mode/2up.

British Library Data Model -Books. S.f. https://www.bl.uk/bibliographic/pdfs/bldatamodelbook.pdf.

Calhoun, Karen. 2006. *The Changing Nature of the Catalog and its Integration with Other Discovery Tool.;* Final Report, Prepared for the Library of Congress. March 17, 2006 http://www.loc.gov/catdir/calhoun-report-final.pdf.

Cao, Longbing. 2017. "Data science: A comprehensive overview". ACM *Computing Surveys*, vol. 50, num. 3, article 43: 1-42. DOI: http://dx.doi.org/10.1145/3076253.

Carlson, Scott. 2006. "Lost in a Sea of Science Data". *The Chronicle of Higher Education,* entrada del 23 de junio de 2006. https://chronicle.com/article/Lost-in-a-Sea-of-Science-Data/9136.

Codd, Edgar F. 1970. "A relational model of data for large shared data banks". *Communications of the* ACM, vol. 13, núm. 6, junio 1970,: 377-387. https://doi.org/10.1145/362384.362685.

Columbus, Louis. 2017. "IBM predicts demand for data scientists will soar 28% by 2020". *Forbes Magazine*. Entrada del 13 de mayo de 2017. https://www.forbes.com/sites/louiscolumbus/2017/05/13/ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020/#232bc4927e3b.

Commission Nationale de l'Informatique et des Libertés (CNIL). 2019. *Map of data protection around the world*. https://www.cnil.fr/en/data-protection-around-the-world.

Cox, Andrew; Kennan, Mary Anne; Lyon, Liz; y Pinfield, Stephen. 2017. "Developments in Research Data Management in Academic Libraries: Towards an Understanding of Research Data Service Maturity". *Journal of the Association for Information Science and Technology,* vol. 68, núm. 9: 2182-2200.

Dans, Enrique. 2011. "Big Data: Una pequeña introducción". Blog del autor, entrada del 19 de octubre de 2011 https://www.enriquedans.com/2011/10/big-data-una-pequena-introduccion.html.

Dartmouth Summer Research Project on Artificial Intelligence (DSRPAI). John McCarthy & Marvin Minsky, July 1956. Hanover, New Hampshire. AI *Magazine* 2006, vol. 27, núm. 4.

Data Management Association (DAMA) (s.f.). "Data Governance Functional Reference Framework". *Data Governance Part* III*: Frameworks–Structure for Organizing Complexity*. Nascio Governance Series, 5 https://www.nascio.org/wp-content/uploads/2019/11/NASCIO-DataGovernancePTIII.pdf.

Dataversity. 2017. *Descriptive Analytics.* https://www.dataversity.net/fundamentals-descriptive-analytics/.

Davies, Roy. 1989. "The Creation of New Knowledge by Information Retrieval and Classification". *Journal of Documentation*, vol. 45, núm. 4: 273-301.

Defining Data Governance. Data Governance Institute. http://datagovernance.com/defining-data-governance/

Devens, Richard Miller. 1865. *Cyclopaedia of Commercial and Business Anecdotes*. Nueva York: D. Appleton and company, 210. https://archive.org/details/cyclopaediacomm00devegoog.

Diebold, Francis. 2013. *On the Origin(s) and Development of "Big Data": The Phenomenon, the Term, and the Discipline*. University of Pennsylvania, Draft. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2202843.

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

Duy, Joanna; Vaughan, Liwen. 2006. "Can electronic journal usage data replace citation data as a measure of journal use? An empirical examination". *The Journal of Academic Librarianship*, núm. 32: 512-517. https://doi.org/10.1016/j.acalib.2006.05.005.

———. 2005. "Are citation data a valid measure of journal use? An empirical examination in an academic context". *Proceedings of the 10th International Conference of the International Society for Scientometrics & Infometrics:* 390-397. Estocolmo: Karolinska University Press. https://spectrum.library.concordia.ca/6414/1/ISSIfall2005_v9.pdf.

———. 2003. "Usage data for electronic resources: a comparison between locally-collected and vendor-provided statistics". *The Journal of Academic Librarianship*, vol. 29, núm. 1: 16-22. https://eric.ed.gov/?id=EJ673398.

Engard, Nicole (Ed.). (2009). *Library Mashups: Exploring New Ways to Deliver Library Data*. Medford, N.J.: Information Today.

———. (Ed.). 2015. More *Library Mashups: Exploring New Ways to Deliver Library Data*. Medford, N.J.: Information Today.

Eur-Lex. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32002L0058. S.f.

Expert systems: HTML, the WWW, and the librarian. 2014 *The Free Library*. https://www.thefreelibrary.com/Expert+systems%3a+HTML%2c+the+WWW%2c+and+the+librarian.-a016880355

Farmer, Lesley; Safer, Alan. 2016. *Library Improvement through Data Analytics*. Chicago: ALA Neal-Schuman.

Farney, Tabatha. 2018. *Using Digital Analytics for Smart Assessment*. Chicago: ALA.

Fayyad, Usama; Piatetsky-Shapiro, Gregory y Smyth, Padhraic. 1996. "From Data Mining to Knowledge Discovery in Databases". AI *Magazine*, vol. 17, núm. 3: 37-54. https://ojs.aaai.org//index.php/aimagazine/article/view/1230 Citado por: Press, Gil (2013). "A Very Short History of Data Science". *Forbes* https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#64a5f94455cf.

Feigenbaum, Edward. 1989. "Toward the Library of the Future". *Long Range Planning*, vol. 22, núm. 1: 18-123. https://doi.org/10.1016/0024-6301(89)90059-9.

Feria, Lourdes. 2020. "Minería de texto aplicada a un diagnóstico de usuarios en Ciencia y Tecnología: Aprendizajes para fortalecer la investigación bibliotecológica", 31-42. *El manejo de datos. Aproximación desde los estudios de la información*. Araceli Torres (Coord.) México: UNAM. Instituto de Investigaciones Bibliotecológicas y de la Información. http://ru.iibi.unam.mx/jspui/handle/IIBI_UNAM/L225.

First International Conference on Learning Analytics and Knowledge – Proceedings of LAK'11. 2011. Nueva York: Association for Computing Machinery. https://dl.acm.org/doi/proceedings/10.1145/2090116.

Forbes. 2014. "12 Big Data Definitions: What's Yours?". https://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/?sh=4c4ede2213ae.

Gantz, John F.; Reinsel, David. 2007. *The Expanding Digital Universe.* An IDC White paper. (International Data Corporation). Sponsored by EMC Corporation. Marzo 2007. https://web.archive.org/web/20090612013506/http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf.

Gantz, John F.; Reinsel, David. 2012. *The Digital Universe in 2020.* IDC. (International Data Corporation). Sponsored by EMC Corporation. December 2012. https://www.slideshare.net/arms8586/the-digital-universe-in-2020.

Gartner Glossary. 2005. "Definition of Big Data. Sitio web del Grupo Gartner". Entrada por: Big Data https://www.gartner.com/en/information-technology/glossary/big-data.

——— 2005. Definition of EIM. Sitio web del Grupo Gartner. Entrada por: Enterprise Information Management https://www.gartner.com/en/information-technology/glossary/enterprise-information-management-eim.

General Data Protection Regulation (GDPR). https://gdpr.eu/tag/gdpr/. s.f.

Geertz, Clifford. 1973. "Thick Description: Towards an Interpretative theory of culture". *The Interpretation of Cultures*, Nueva York: Basic Books: 3-31. https://chairoflogicphiloscult.files.wordpress.com/2013/02/clifford-geertz-the-interpretation-of-cultures.pdf.

Gibbons, Paul. 2015. *The Science of Successful Organizational Change.* Boger.

Gorbea, Salvador; Piña, Maricela. 2013. "Propuesta de un indicador para medir el comportamiento del desarrollo disciplinar de las Ciencias Bibliotecológica y de la Información en instituciones académicas". *Investigación Bibliotecológica: Archivonomía, bibliotecología e información,* vol. 27, núm. 60: 153-180 http://ru.iibi.unam.mx/jspui/bitstream/IIBI_UNAM/A111/1/art60-7.pdf.

Gorman, Michael. 2000. *Our Enduring Values Revisited: Librarianship in the Twentieth Century*. Chicago: ALA http://publish.illinois.edu/whylibraries/files/2019/10/gorman-2015-enduring.pdf.

Halevi, Gali; Nicolas, Barnaby; Bar-Ilan, Judit. 2016. "The Complexity of Measuring the Impact of Books". *Publishing Research Quarterly,* vol. 32: 187-200. https://doi.org/10.1007/s12109-016-9464-5.

———— . 2014. *Bibliometric Big Data and its Uses* http://bdigital.unal.edu.co/12475/7/biliometricsbigdata.pdf.

Halevi, Gali; Moed, Henk. 2012. "The evolution of big data as a research and scientific topic: Overview of the literature". *Research Trends*, núm. 30: 3-6. https://www.researchgate.net/publication/285119834_The_evolution_of_big_data_as_a_research_and_scientific_topic_Overview_of_the_literature.

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

Han, Jiawei; Kamber; Micheline, and Pei, Jian. 2011. *Data Mining: Concepts and Techniques*, 3a ed. Morgan Kaufmann

Harada, Takishi. 2019. *Robotics and artificial intelligence technology in Japanese libraries*. IFLA WLIC, 21-22 August 2019. http://library.ifla.org/2695/1/s08-2019-harada-en.pdf.

Hernon, Peter; Dugan, Robert; Matthews, Joseph. 2015. *Managing with Data: Using* ACRLM*etrics and* PLA*metrics*. Chicago. ALA.

Hayashi, Chikio *et al.* (Eds.). 1996. "Data Science, Classification, and Related Methods". *Proceedings of the Fifth Conference of the International Federation of Classification Societies* (IFCS-96), Kobe, Japón, Marzo 27-30, 1996. Springer. Índice en: https://d-nb.info/955715512/04.

Hayashi, Chikio. 1996. "What is data science?" *Data science, classification, and related methods*: 40-51. Springer: Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japón, March 1996. https://www.springer.com/gp/book/9784431702085.

Heery, Rachel. 2004. "Metadata Futures: Steps toward Semantic Interoperability". *Metadata in Practice*. D. I. Hillman and E. L. Westbrooks (Eds.): 257-71. Chicago: American Library Association.

Hey, Tony; Hey, Jessie. 2006. "E-Science and its Implications for the Library Community". *Library Hi Tech*, vol. 24, núm. 4: 515-528. http://www.emeraldinsight.com/doi/pdfplus/10.1108/07378830610715383.

Hey, Tony; Tansley, Stewart; Tolle Kristin (Eds.) 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, Wa.: Microsoft Research. 252. https://digital.library.unt.edu/ark:/67531/metadc31516/m2/1/high_res_d/4th_paradigm_book_complete_lr.pdf.

Huang, Wenyi; Wu, Zhaohui; Liang, Chen; Mitra, Prasenjit y Giles, Lee. 2015. "A Neural Probabilistic Model for Context based Citation Recommendation". AAAI *2015: Proceedings of the 29th* AAAI *Conference on Artificial Intelligence*: 2404-2410. https://clgiles.ist.psu.edu/pubs/AAAI2015-neural-probabilistic.pdf.

Huff, Darrell. 1954. *How to Lie with Statistics*. Nueva York: W. W. Norton & Co. Versiones en castellano: *Cómo mentir con estadísticas*. Barcelona: Sagitario, 1965 y Barcelona: Crítica, 2011.

IBM (s.d.). *Predictive Analytics*. https://www.ibm.com/analytics/predictive-analytics.

International Federation of Library Associations and Institutions. Big Data Special Interest Group. 2014. https://www.ifla.org/units/big-data/.

———. 2012. IFLA *Code of Ethics for librarians and other information workers*. Approved August 2012. https://www.ifla.org/files/assets/faife/codesofethics/englishcodeofethicsfull.pdf.

———. 2015. IFLA *Statement on Privacy in the Library Environment*. Endorsed by IFLA Governing Board, August 4, 2015. https://www.ifla.org/node/9803.

———. 2017. *Linked Data for Libraries*. https://www.ifla.org/best-practice-for-national-bibliographic-agencies-in-a-digital-age/node/8549.

———. 2018. International Federation of Library Associations and Institutions.

———. S.f. Resources. https://www.ifla.org/publications/node/1691.

———. S.f.-2. "Library Theory and Research Projects". https://www.ifla.org/library-theory-and-research/projects.

IFLA Journal, vol. 44, num. 3, octubre 2018. https://www.ifla.org/files/assets/hq/publications/ifla-journal/ifla-journal-44-3_2018.pdf.

Internet Archive. "About the Internet Archive". https://archive.org/about/.

Ipsos. Global Citizens and Data Privacy. 2019. Presentación. https://www.ipsos.com/sites/default/files/globalcitizens-data-privacy.pdf.

Jacknis, Norman. 2017. "The AI-Enhanced Library". Blog del autor. Entrada del 21 de junio de 2017. https://jacknis.com/tag/library/.

Jharotia, Anil. 2016. "Big Data Technology: Big Opportunity for Librarians". *Librarianship in ICT Age.* Agra: Y. K. Publishers: 1-9. https://www.researchgate.net/publication/326972552_Big_Data_Technology_Big_Opportunity_for_Librarians.

Joyanes, Luis. 2013. *Big Data: Análisis de grandes volúmenes de datos en organizaciones.* México: Alfaomega.

Kalantari, Ali; Kamsin, Amirrudi; Kamaruddin, Halim; Ebrahim *et al.* 2017. "A bibliometric approach to tracking big data research trends". *Journal of Big Data,* vol. 4, núm. 30. https://doi.org/10.1186/s40537-017-0088-1.

Kenwright, David. 1999. "Automation or interaction: what's best for big data?". *Proceedings Visualization '99*, San Francisco, California, 1999: 491-495, DOI: 10.1109/VISUAL.1999.809940

King, David Lee. 2018. "Big Data and Libraries". Blog del autor. Entrada del 13 de Diciembre 2018. https://davidleeking.com/big-data-and-libraries/.

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

Lang, Charles; Siemens, George; Wise, Alyssa y Gasevic, Dragan (Eds.). 2017. *Handbook of Learning Analytics. Society for Learning Analytics Research* (solar). https://solaresearch.org/wp-content/uploads/2017/05/hla17.pdf.

Lawlor, Bonnie. 2016. "An overview of the NFAIS 2016 Annual Conference: Data sparks discovery of tomorrow's global knowledge". *Information Services & Use*, vol. 36, núm. 1/2: 3-21. DOI: 10.3233/ISU-160807.

Lewis, Martin. 2010. "Libraries and the management of research data". McKnight, S. (ed.) *Envisioning Future Academic Library Services: Initiatives, Ideas and Challenges*. London: Facet: 145–168. https://eprints.whiterose.ac.uk/11171/1/LEWIS_Chapter_v10.pdf.

Lexico Dictionary. 2014. Entrada por: Big Data. (Lexico By Oxford) https://quizlet.com/203476200/business-intelligence-cis-chapter-13-flash-cards/.

Library of Congress, s.f. https://id.loc.gov/.

Liu, Shan; Shen, Xiao-Lang. 2018. "Library management and innovation in the Big Data Era". *Library Hi Tech*, vol. 36, núm. 3: 374-377. https://doi.org/10.1108/LHT-09-2018-272.

Lohr, Steve. 2013. The 'Origins of Big Data': an etymological detective story. New York Times blog. https://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/

Lomotey, Richard; Deters, Ralph. 2014. "Towards knowledge discovery in Big Data". *Proceedings of the 8th International Symposium on Service-oriented System Engineering*. IEEE Computer Society: 181-191. DOI: 10.1109/SOSE.2014.25.

López-Yepes, José (Ed.). 2004. *Diccionario Enciclopédico de Ciencias de la Documentación*. Madrid: Síntesis, 2 vol.

Lorica, Ben. 2014. *Big Data Now*. Sebastopol, California, EUA: O'Reilly Media. https://www.oreilly.com/data/free/files/bigdatanow2013.pdf.

Lyman, Peter; Varian, Hal. 2000. "How Much Information?". *Journal of Electronic Publishing*. December 2000, vol. 6, núm. 2. http://www.press.umich.edu/jep/06-02/lyman.html.

Lyon, Liz; Mattern, Eleanor; Acker, Amelia y Langmead, Alison. 2015. "Applying translational principles to data science curriculum development". *iPres Conference Proceedings*. Chapel Hill, November 2015. http://d-scholarship.pitt.edu/27159/1/Applying_Translational_Principles_to_Dat.pdf.

Lyon, Liz; Mattern, Eleanor. 2016. "Education for Real-World Data Science Roles (Part 2): A Translational Approach to Curriculum Development". *International Journal of Digital Curation*, vol. 11, núm. 2 DOI:10.2218/ijdc.v11i2.417.

Markey, Karen. 2007. "The Online Library Catalog: Paradise Lost and Paradise Regained?" *D-Lib Magazine*, vol. 13, núms. 1-2, Jan./Feb. 2007. http://www.dlib.org/dlib/january07/markey/01markey.html.

Marr, Bernard. 2015. "Why only one of the 5 V's of big data really matters". IBM *Big Data & Analytics Hub*. Entrada del 19 de marzo de 2015. https://www.ibmbigdatahub.com/blog/why-only-one-5-vs-big-data-really-matters .

———— . 2020. *Revista Forbes*. Entrada del 24 de Febrero de 2020. https://www.forbes.com/sites/bernzardmarr/2020/02/24/the-9-best-free-online-data-science-courses-in-2020/?sh=6fabfd7d2bbf.

Martínez Musiño, Celso. 2020. "Big Data - Análisis informétrico de documentos indexados en Scopus y Web of Science". *Investigación Bibliotecológica: Archivonomía, bibliotecología e información,* vol. 34, núm. 82: 87-102. http://dx.doi.org/10.22201/iibi.24488321xe.2020.82.58035.

Mashey, John. 1999. "Big Data… and the Next Wave of Infrastress". *USENIX Annual Conference*, Monterey, California, junio 6-11, 1999. https://www.usenix.org/conference/1999-usenix-annual-technical-conference/big-data-and-next-wave-infrastress-problems.

Matusiak, Krystyna. 2019. "Research Data Management and Libraries: Opportunities and Challenges", *El manejo de datos. Aproximación desde los estudios de la información*. Araceli Torres (Coord.). México: UNAM. Instituto de Investigaciones Bibliotecológicas y de la Información. http://ru.iibi.unam.mx/jspui/handle/IIBI_UNAM/L225.

De Mauro, Andrea; Greco, Marco y Grimaldi, Michele. 2016. "A Formal Definition of Big Data Based on Its Essential Features". *Library Review*, vol. 65, núm. 3: 122-135. DOI:10.1108/LR-06-2015-0061.

McJones, Paul (Ed.). 1995. *The 1995* SQL *Reunion: People, Projects, and Politics*. Recorded May 29, 1995. Computer History Museum. CHM Reference number: X7466.2015.2015. 52. http://archive.computerhistory.org/resources/access/text/2015/07/102740133-05-01-acc.pdf.

Van der Meulen, Rob. 2018. *What Edge computing means for infrastructure and operations leaders*. Gartner Research Group. https://www.gartner.com/smarterwithgartner/what-edge-computing-means-for-infrastructure-and-operations-leaders.

México: *Ley General de Protección de Datos Personales en Posesión de Sujetos Obligados*. (DOF–26-01-2017). https://www.dof.gob.mx/nota_detalle.php?codigo=5469949&fecha=26/01/2017.

Milton, Simon. 1998. "Top-Level Ontology: The problem with Naturalism". N. Guarino (Ed.), *Formal Ontology in Information Systems:* 85-94. Ámsterdam, Países Bajos: IOS Press.

Moed, Henk F. 2012. "The use of datasets in bibliometric research". *Research Trends*, 30, septiembre 2012. https://www.researchtrends.com/issue-30-september-2012/the-use-of-big-datasets-in-bibliometric-research/.

Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishan, P., Qazvinian, V. *et al*. 2009. "Using citations to generate surveys of scientific paradigms". *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics:* 584-592.

Morris, Robert; Truskowski, B.J. 2003. "The evolution of storage systems". *IBM Systems Journal*, vol. 42, núm. 2: 205-217. DOI: 10.1147/sj.422.0205.

Murtaugh, Fionn; Devlin, Keith. 2018. "The Development of Data Science: Implications for Education, Employment, Research, and the Data Revolution for Sustainable Development". *Big Data and Cognitive Computing*, 2018, vol. XX, num. 1; https://web.stanford.edu/~kdevlin/Papers/Murtagh-Devlin_2018.pdf.

Naur, Peter. 1974. *Concise Survey of Computer Methods*. Studentlitteratur: Lund, Sweden: 397 Citado por: Press, Gil (2013). *A Very Short History of Data Science*. *Revista Forbes* https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#64a5f94455cf.

Network of the National Library of Medicine (NNLM). https://nnlm.gov/data/thesaurus/data-repository. S.f.

New media Consortium. 2013. *The* NMC *Horizon Report: 2013 Higher Education Edition*: 44. https://files.eric.ed.gov/fulltext/ED559358.pdf.

New Zealand National Library. 2017. "Integrated Library Systems (ILS) checklist". *Choosing an Integrated Library System (ILS)*. Sitio web de la biblioteca. https://natlib.govt.nz/schools/school-libraries/library-systems-and-operations/your-library-catalogue/choosing-an-integrated-library-system-ils.

Nicholson, Shawn; Bennett, Terrence. 2016. "Dissemination and discovery of diverse data: Do libraries promote their unique research data collections?". *International Information & Library Review*, vol. 48, núm. 2: 85-93. https://www.tandfonline.com/doi/full/10.1080/10572317.2016.1176448.

OCLC. *OCLC and linked data*. S.f. https://www.oclc.org/en/worldcat/oclc-and-linked-data.html.

ODC. International Open Data Charter. *Carta internacional de datos abiertos*. https://opendatacharter.net/principles-es/. S.f.

Ohsumi, Noboru. 2000. "From data analysis to data science". *Data Analysis, Classification, and Related Methods*. Kiers, Rasson, Groenen y Schader (Eds.). Heidelberg: Springer: 329-334.

Open Archives. 2014. *The Open Archives Initiative Protocol for Metadata Harvesting*. https://www.openarchives.org/OAI/openarchivesprotocol.html.

Olavsrud, Thor. 2012. "Big Data Causes Concern and Big Confusion". CIO *Blog,* entrada del 24 de febrero de 2012. http://www.cio.com/article/700804/Big_Data_Causes_Concern_and_Big_Confusion?page=2&taxonomyId=3002.

———. 2020. "What is Data Governance?". CIO *Blog*, entrada del 11 de febrero de 2020. https://www.cio.com/article/3521011/what-is-data-governance-a-best-practices-framework-for-managing-data-assets.html.

Olendorf, Robert; Wang, Yan. 2017. "Big Data in Libraries". Suh S. and Anthony T. (eds.) *Big Data and Visual Analytics*. Springer, Cham: 191-202. DOI: https://doi.org/10.1007/978-3-319-63917-8_11.

Olivares Marín, Susana. 2020. Proceso de gestión del conocimiento tácito en bibliotecas digitales universitarias. Tesis de Doctorado en Bibliotecología y Estudios de la Información. México: UNAM.

Oracle (s.f.). *Qué es una base de datos relacional*. https://www.oracle.com/mx/database/what-is-a-relational-database/.

Organización de las Naciones Unidas (ONU). 2013. "Resolución 68/167 sobre el derecho a la privacidad en la era digital". 18 de diciembre 2013. https://www.un.org/ga/search/view_doc.asp?symbol=A/C.3/68/L.45/Rev.1&Lang=S.

Orland-Barak, Lily; Mazkit, Ditza. 2017. *Methodologies of Mediation in Professional Learning*. Springer Intl.

Oshkosh Daily Northwestern Newspaper. 1960. (Wisconsin), 24 Mayo 1960: 19 https://newspaperarchive.com/oshkosh-daily-northwestern-may-24-1960-p-19/.

Oxford English Dictionary. Entrada por "Information", subentrada "Information explosion". https://www.oed.com/viewdictionaryentry/Entry/95568#eid112206197.

Parry, Marc. 2018. "Big Data on Campus", *The New York Times*, Entrada del 18 de julio 2012. https://www.nytimes.com/2012/07/22/education/edlife/colleges-awakening-to-the-opportunities-of-data-mining.html.

Perales, Alicia. 1962. "La documentación". *Anuario de Biblioteconomía y Archivonomía*. Año II. México: Facultad de Filosofía y Letras, UNAM: 9-34.

Phetteplace, Eric. 2012. "Effectively Visualizing Library Data". *Reference & User Services Quarterly* (ALA), vol. 52, núm. 2, invierno 2012: 93-97. https://www.jstor.org/stable/refuseserq.52.2.93.

Piganiol, Pierre. 1971. "Prefacio del Reporte OCDE" *Information for a Changing Society*. París: OCDE: 13. https://files.eric.ed.gov/fulltext/ED057307.pdf.

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

Plale, Beth. 2013. "Big data opportunities and challenges for IR, text mining and NLP". *Proceedings of the 2013 International Workshop on Mining Unstructured Big Data Using Natural Language Processing* (UnstructureNLP'13). Nueva York, ACM: 1-2. https://dl.acm.org/doi/10.1145/2513549.2514739.

———. 2016. *Notes from the School of Information Sciences*, University of Illinois Urbana-Champaign. Entrada del 23 de febrero, 2016. https://ischool.illinois.edu/news-events/news/2016/02/project-will-help-researchers-explore-big-data-hathitrust-digitized.

Plunkett, Tom; McDonald, Brian; Nelson, Bruce. 2013. *Oracle Big Data Handbook*. McGraw-Hill Osborne Media.

Pollock, Rufus. 2013. *What do we mean by Small Data*. Open Knowledge Foundation Blog. Entrada del 26 de abril de 2013. https://blog.okfn.org/2013/04/26/what-do-we-mean-by-small-data/.

Press, Gil. 2013. "A Very Short History of Data Science". *Revista Forbes.* https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-datascience/#64a5f94455cf.

Qazvinian, V., Radev, D. R. y Özgür, A. 2010. "Citation summarization through keyphrase extraction". *Proceedings of the 23rd International Conference on Computational Linguistics,* COLING *'10*. (Stroudsburg, PA: Association for Computational Linguistics): 895-903.

Qin, Jian; Norton, Jay (Eds.). 1999. "Knowledge Discovery in Bibliographic Databases". *Library Trends*, vol., 48, núm. 1, verano 1999. https://aquila.usm.edufac_pubs/4791/.

R3Data. https://www.re3data.org/ .S.f.

Rees, Alan; Saracevic, Tefko. 1967. "Education for Information Science and its relation to librarianship". *Special Libraries Association Annual Conference,* Nueva York, mayo 29, 1967: 2. Citados por: Shera, Jesse. 1968. Sobre Bibliotecología, Documentación y Ciencia de la Información. En: *Boletín* UNESCO *de Bibliotecas*, vol. XXII, núm. 2, marzo-abril 1968: 62-70.

Research Trends. 2012. https://www.researchtrends.com/wp-content/uploads/2012/09/Research_Trends_Issue30.pdf.

Revista Iberoamericana de Educación / Educaçao. 2019. *Analítica del aprendizaje y la educación (Learning Analytics and education): clasificación, descripción y predicción del aprendizaje de los estudiantes*, vol. 80, núm.1. https://rieoei.org/RIE/issue/view/Learning%20Analytics.

De Rosa, Cathy. 2006. *Perceptions of Libraries and Information Resources: A Report to the* OCLC. Columbus, OH.: OCLC.

Ryle, Gilbert. 1949. *Concept of the mind.* Nueva York: Hutchinson & Co.

Salazar, Javier. 2020. "Plan para el desarrollo de la Ciencia de Datos y Big Data (PDCDBD) en la UNAM con fines académicos y administrativos": 94-113. *El manejo de datos. Aproximación desde los estudios de la información*. Araceli Torres (Coord.). México: UNAM. Instituto de Investigaciones Bibliotecológicas y de la Información. http://ru.iibi.unam.mx/jspui/handle/IIBI_UNAM/L225.

Saracevic, Tefko. 1992. "Information Science: Origin, Evolution and Relations". Vakkari, Pertti y Cronin, Blaise (Eds.). *Conceptions of Library and Information Science. Historical, Empirical and Theoretical Perspectives*. London: Taylor Graham: 5-27.

Scielo Data. https://data.scielo.org/. S.f.

Shera, Jesse. 1973. "Toward a Theory of Librarianship and Information Science". *Knowing books and men; knowing computers, too*. Libraries Unlimited: 363.

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

Schmarzo, Bill. 2018. "Importance of Metadata in a Big Data World". *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.

Schwartz, Meredith. 2013. "What Governmental Big Data May Mean For Libraries". *Library Journal*. mayo 30, 2013. http://lj.libraryjournal.com/2013/05/oa/what-governmental-big-data-may-m.

Segal, Troy. 2019. *Investopedia. Prescriptive Analytics*. https://www.investopedia.com/terms/p/prescriptive-analytics.asp.

Showers, Ben (Ed). 2015. *Library Analytics and Metrics: Using Data to Drive Decisions and Services*. Londres: Facet. 176.

Sicense (s.d). "Data Analytics Glossary. Diagnostic Analytics". https://www.sisense.com/glossary/diagnostic-analytics/.

Small, Henry; Klavans, Richard. 2011. "Identifying Scientific Breakthroughs by Combining Co-citation Analysis and Citation Context". *Proceedings of 13th International Conference of the International Society for Scientometrics and Informetrics* (ISSI 2011).

Smithsonian Libraries. Data Repositories. https://library.si.edu/research/data-repositories. S.f.

Souza, Renato; Tudhope, Douglas y Almeida, Mauricio. 2011. *Towards a taxonomy of* KOS*: Dimensions for classifying Knowledge Organization Systems.* International Society for Knowledge Organization (ISKO). 11th isko International Conference. Rome (Italy): 23-26. Febrero 2010.

Springer Nature, s.f. https://www.springernature.com/gp/researchers/scigraph.

Staff, Frank. 1993. *The Penny Post 1680-1918*. The Lutterworth Press.

Stanton, Jeffrey. 2012. "Data Science: What's in it for the New Librarian?", *Infospace. The Official Blog of the Syracuse University iSchool*. Entrada del 16 de julio de 2012. https://ischool.syr.edu/infospace/2012/07/16/data-science-whats-in-it-for-the-new-librarian/.

Stephens, Owen. 2011. "Mashups and open data in libraries". *Serials*, vol. 24, núm. 3: 245–250. DOI: http://doi.org/10.1629/24245.

Streitfeld, David. 1989. "Infomania". *The Washington Post*, February 3, 1989. https://www.washingtonpost.com/archive/lifestyle/1989/02/03/infomania/54d862a2-ba33-4ffe-a810-f96a0c2ca3a3/.

Swan, Alma; Brown, Sheridan. 2008. *The Skills, Role and Career Structure of Data Scientists and Curators: An Assessment of Current Practice and Future Needs.* Report to the JISC. Truro, UK: Key Perspectives. 34. http://www.jisc.ac.uk/publications/documents/dataskillscareersfinalreport.aspx.

Techopedia. 2020. *What are some core principles of data governance?* https://www.techopedia.com/7/32187/enterprise/databases/what-are-some-core-principles-of-data-governance.

Tenopir Carol; Birch Ben y Allard, Suzie. 2012. "Academic libraries and research data services: Current practices and plans for the future". Association of College and Research Libraries. 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?". *Journal of eScience Librarianship*, vol. 4, núm. 2. https://escholarship.umassmed.edu/jeslib/vol4/iss2/4/.

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

Torres Vargas, G. Araceli. 2020. *El manejo de datos. Aproximación desde los estudios de la información*. México: UNAM. Instituto de Investigaciones Bibliotecológicas y de la Información. http://ru.iibi.unam.mx/jspui/handle/IIBI_UNAM/L225.

Tufte, Edward. 2001. *The visual display of quantitative information*. 2nd ed. Connecticut: Graphics Press.

Tukey, John W. 1962. "The Future of Data Analysis". *The Annals of Mathematical Statistics*, vol. 33, núm. 1: 1-67. DOI: 10.1214/aoms/1177704711 https://projecteuclid.org/euclid.aoms/1177704711. Citado por: Press, Gil 2013. *A Very Short History of Data Science. Revista Forbes*. https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#64a5f94455cf.

Uschold, Michael; Grüninger, Michael. 1996. "Ontologies: Principles, Methods, and Applications". *Knowledge Engineering Review*, vol. 11, núm. 2, Cambridge Journals On-Line: 93–155. DOI: https://doi.org/10.1017/S0269888900007797.

Wang, Chunning; Xu, Shaochun; Chen, Lichao y Chen, Xuhui. 2016. "Exposing Library Data with Big Data Technology: A Review". 2016 IEEE/ACIS *15th International Conference on Computer and Information Science (*ICIS*)*. DOI: 10.1109/ICIS.2016.7550937.

Warden, Peter. 2011. *Glossary of Big Data*. O'Reilly Media.

We Are Social. 2020. https://wearesocial.com/digital-2020.

Wheatley, Amanda; Hervieux, Sandy. 2019. "Artificial Intelligence in Academic Libraries: An Environmental Scan". *Information Services & Use*, vol. 39, num. 4: 347-356. DOI: 10.3233/ISU-190065.

Whyte Angus; Tedds, Jonathan. 2011. "Making the case for research data management". DCC Briefing Papers. Edinburgh: Digital Curation Centre. https://www.dcc.ac.uk/guidance/briefing-papers/making-case-rdm.

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

Wright, Curtis. 2013. *Jesse Shera, Librarianship, and information Science*. Library Juice Press: 146.

World Wide Web Consortium (W3C). 2011. *Library Linked Data Incubator Group Final Report*. https://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/.

Xu, Zeshui; Yu, Dejian. 2019. "A Bibliometrics analysis on big data research (2009–2018)". *Journal of Data, Information, and Management,* vol. 1: 3-15. https://doi.org/10.1007/s42488-019-00001-2.

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

Zeng, Marcia Ley; Qin, Jian. 2008. *Metadata*. Nueva York: Neal-Schuman.

Zhan, Ming; Widén, Gunilla. 2017. "Understanding big data in librarianship". *Journal of Librarianship and Information Science*, vol. 51, núm. 2: 561-576. DOI: 10.1177/0961000617742451.

Big Data in Libraries

NOTE TO THE ENGLISH VERSION:

This text is a translation from an original written in Spanish for Spanish-speaking readers. Therefore, it contains references to texts in such language on certain topics considered relevant for those readers, and makes reference to some specific issues and examples in Latin America. It also contains some definitions and explanations of English terms that are not necessarily obvious in the Ibero-American library context.

Introduction

Facts are ontological, evidence is epistemological, data is rhetorical
DANIEL ROSENBERG,
2013. "Data before the fact"

The concept of "Big Data" is not the vertical evolution of a single concept or technology over the decades, but the simultaneous conjunction of multiple phenomena, needs, technologies, theories, tools and methods related to information, thus becoming something more complex when concurring at a certain point. It includes multiple backgrounds, aspects, and components, and can be analyzed from different approaches; therefore, in order to better understand the concept it is necessary to review the main ones in order to obtain a complete vision of its essence.

During the second half of the 20th century, numerous authors –such as Derek de Solla Price, F.W. Lancaster or Alvin Toffler– studied and referred to the phenomenon called "information explosion". As it is well known, this term basically refers to the massive and relentless growth of published information, and the challenges that such volume has posed for its proper management, as well as to the social, technical, economic and other, etc., effects which derive from this phenomenon. According to the Oxford English Dictionary (s.d.), in the "Lawton Constitution Newspaper", of Lawton, Oklahoma, November 30, 1941, this term was used for the first time to refer to the enormous growth of global information. By the 1960s, it began to become widespread: in

1960 the Oshkosh Northwestern Newspaper (1960, 19) of the state of Wisconsin, USA, already mentioned the concept in its full meaning: "the information explosion in which the field of science alone estimates that the amount of material available has doubled in the last 10 years". During that decade the term would increasingly appeared in technical and academic papers.

Around that time, the term "information science" was also introduced as a successor to the terms: Iinformation behavior", "user studies" and "documentation". Alvares & Araújo, (2010, 200) mentioned: "The International Conference on Scientific Information, held in Washington in 1958, marked the transformation of 'documentation' into 'information science'". Tefko Saracevic also placed in the sixties the change of these denominations towards "information science", in his historical review of this discipline – Saracevic (1992, 5-27). He also defined the term as early as 1967:

> [Information Science] is concerned with the properties, behavior and circulation of information. It encompasses the analysis of systems, the mesological aspects of information and communication, information media and linguistic analysis, information organization, human-system relations, etc. – Rees & Saracevic (1967, 2).

At some point in the second half of the 20th century and with the advent of the digital era, the concept of "information explosion" evolved into "information overload", basically to refer –in addition to the immense amount of information produced– to the situation faced by a system, be it computational, social, etc., when the information inputs exceed its information processing capabilities.[1] Towards the end of the century, the concepts of "information explosion", "information overload and "information science", were also extended to data; and subsequently and as a consequence all these terms would generate their own specialties and contexts.

---

1 The concept of "information overload" is derived from the social sciences, and is attributed to Bertram Gross in 1964 in his text *The managing of organizations: The administrative struggles*.

Data have always been the primary source of information; basically, data assembled in context become information for the further purpose of inferring knowledge. However, for centuries the emphasis was on information, being data only a raw material in its process. Portions of information were used as products mostly finishsed i.e., the final result of the analysis and synthesis of certain data were presented by a person or group in the form of a publication. And then these products took the form of books, papers in academic journals, press texts, handbooks, dissertations, compendiums, patents, and so on. In the second half of the 20th century, the absolute emphasis on information progressively decreased while data gained preponderance, as it became evident that data inherently had an additional added value and that organizations could generate and/or collect it for better decision making and not only for publication purposes.

Obviously, this shift in emphasis between information and data did not happen overnight: it occurred gradually throughout the last decades of the past century. However, an interesting turning point in this regard can be observed in a note appeared in the Washington Post in February 1989. It reads:

> [...] according to one estimate, more new information has been cranked out in the last three decades than in the previous five millennia. The total amount of printed knowledge doubles every eight years... The result? Information anxiety, described as 'the black hole between data and knowledge'. The difference between the two: data is the raw material, and is passive; information is active and, ideally at least, enlightening. As we thrash around in the over-abundance of the first, the second becomes ever more elusive (Streitfeld 1989).

It is noteworthy how by the time of this note, data are still "raw material and passive", while information "is active and enlightening". From this, it can be concluded that in such year it was already assumed there existed an overabundance of data and information, but the former had not yet reached the current levels of ponderation.

On the other hand, at the end of the first decade of this century, it was also widely concluded that data collected in the course of academic research should be stored in a systematized form after the end of the project so that they could be reused later, since a certain set of compiled data could be analyzed in multiple ways by different groups and new results could be extracted from them. Data was no longer just raw material for producing information; it was an object and a product in itself, with its own intrinsic value, and therefore it required specific treatment.

Such treatment would be included in what was already known as "data management": a series of theories, principles, modalities, methods, tools, technologies, etc., for its handling and use. Specialties also emerged within it, such as "data engineering", which deals with data organization and retrieval, and has to do with how intrinsically clean and structured the data is within a certain set of them. "Data analytics", which is the task of identifying which variables in the organization can be related to certain data in order to formulate questions and eventually obtain solutions through specific techniques. All this shaped "data science", which consists of the study of organized data to identify those being important in the context of a specific problem or a certain business model; it also has to do with the development of models and algorithms intended to solve large-scale problems in organizations. Information science should not be mistaken with data science. While the two complement each other and overlap, they are not synonymous. Basically, *information science* is an interdisciplinary science which studies the properties, behavior and flows of information; its body of knowledge encompasses the creation, collection, organization, storage, retrieval, dissemination, utilization and preservation of recorded information as organized documentary resources. *Data science* is an interdisciplinary science concerned with the discovery of meaningful knowledge and utilitarian information from data. In addition to the objects of study, many authors agree that a further difference is observed in the stages of development: information science is mostly a vertical evolution of documentation, which as a discipline gradually crossed new frontiers from other

areas of human cognition adding new knowledge from them until it became what it is now. In the other hand, Data science is the conjunction of multiple disciplines very different from each other which evolved in parallel until they joined together at some point giving rise to a totally new vision of data.

In the early years of this century, Microsoft researchers Jim Gray and his collaborators argued that contemporary science was reaching its fourth paradigm. It had been accepted for centuries that science was based on its two fundamental paradigms: theory and experimentation. With the arrival of John Von Neumann's, theories and computers in the second half of the 20th century, simulation and modeling had been integrated as the third paradigm. Gray established that a fourth paradigm had arrived for science, which complemented the other three: data. Science was now one based intensely on data. A new generation of concepts, methodologies, tools and experts was therefore required to deal with it. Hey *et al.* (2009) published in that year the first anthology on the subject, considered the cornerstone of this vision of modern science in relation to data. Hey was also one of the first to point out, since 2006 that the emerging facets of science at that time, –e-Science, open science, etc.– established a new relationship between science and libraries precisely because of data – Hey & Hey (2006, 525-526); something similar was also pointed out in that year by Carlson (2006). Libraries and data were now formally linked. From this particular approach to data derived from scientific research, López-Yepes (2004, 59, 411) defined it as follows:

> [...] data constitute the basic unit of analysis contained in this type of information, which represent the testimony or minimum expression of a measurable fact... it is the starting point, the undisputed facts and principles for a scientific research. It is the sensible experience in empiricism... a set of data makes up a piece of information.

Another major factor contributing to the development of data was the unprecedented growth of the global network and telecommunications since the 1990s, which generated a new and growing

development of information in its digital form. Millions of pieces of information and data were created in this modality, which were added to what already existed, multiplying exponentially the amount of accumulated information. During the last decade, the enormous expansion of various sectors of the Internet has been creating an unusual data flow, especially the growth of social networks. Originally, people's basic data who subscribed to them were simple sets with an economic value, which were sold to companies or organizations for advertising or marketing purposes. It did not take long to the owners of these networks to realize that the interaction of millions of people every day produces huge streams of data which could be analyzed to extract new and additional information, which in turn also had a new economic value and could be offered for sale too.[2] Even if a user does not explicitly provides any data about himself, using networks implies leaving a data footprint, such as his/her location, preferences, choices, purchases or sales, time spent, places visited, search history, and so on. In the second decade of this century, the phenomenon known as the "Internet of Things" or IoT was added to this, [it] which consists of numerous common objects –beyond computers and smartphones– connected to the Internet to exchange data automatically with other devices and/or systems without expressed instruction: parcels, household appliances, GPS, merchandise in a warehouse, thermostats, traffic lights, cameras, etc. A special type of these are *wearables*, i.e., personal devices also interconnected to the Internet; this is a specific category referring to digital devices for personal use which collect and exchange data within the network: pulse watches, medical implants, fitness trackers, people and/or pet locators, augmented reality lens, hearing aids, etc.

All these phenomena added volume to information, and especially to data. The unprecedented amounts of data derived from all the above, in addition to what already existed, contributed to the concept of "Big Data". This is a term referred to "extremely

2 As an example, the InternetLiveStats ? site, states that more than 200 billion tweets were sent in 2020, and Facebook currently has 2.5 billion users.

large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions" (Lexico 2014). As can be seen, the concept of big data is not the vertical evolution of a certain idea over the decades, but instead the conjunction of multiple phenomena, theories, methods, technologies, etc., which led to something more complex. This concept will be reviewed in detail below.

Like many other technologies, big data was presented and treated from its inception and for years as the great panacea of data management, and therefore created in people and organizations too many expectations. In Jakob Nielsen's words: "[...] the two classic errors in predicting the future of a technology shift are to overestimate its short-term impact and underestimate its long-term impact". Currently, the euphoria of big data is beginning to subside and to take on its real dimensions and perspectives, and therefore this is a good time to analyze it. The main reason for this is that –despite all the exaggerations in this regard– the big data management really does represent a new and valid tool for information analysis for assiting decision-making in organizations, including libraries and archives, and it is therefore advisable for the staff of these organizations dedicated to information management to be trained in these techniques, tools and procedures. This knowledge and skills should not be left to the IT staff alone: it is essential that librarians and archivists also master it, since it is an element of added value, both for the organization where they work and for all the staff dedicated to information management. All indicates that big data can power major innovations in the entire information services environment –obviously libraries included– to create and deliver unprecedented and personalized services while optimizing costs.

In addition, other phenomena related to information and closely related to big data are being developed in parallel, such as the concept of "intelligent city" or "digital city", which basically consists of

> [...] the virtual space of interaction between all the actors participating in the life of a city –citizens, companies, administrations,

> visitors, etc. – using electronic media and ict as support, offering to these actors access to an innovative medium of relationship and communication, through the channel of their choice, at any time and place. The main objective is the improvement of the relationship and services between the actors that interact in the city, both in existing and future services, promoting a sustainable economic and social development of the city" (Manual de Ciudades Digitales 2012).

Obviously, a concept of "digital city" cannot be conceived without considering libraries and information services for this purpose.

The famed *Forbes Magazine*, specialized in finance and economics, stated that in 2015 there were fewer than 20,000 specialists in data and its analysis in the whole world, and that by 2020 2.7 million of them would be required (Columbus 2017). According to this, there is a huge gap between the supply and the demand of specialized data management personnel.

It is therefore advisable for librarians and archivists to start mastering the principles and theories of big data, the structured, unstructured or semi-structured data models, as well as some of the existing hardware and software tools and technologies for their handling and analysis. More on this later.

Background

Millions of dollars yearly are spent on collection of data... Those who can afford punch card processing machines can process more data than ever before - 3,000 records per hour!
WILLARD BRINTON, 1914

While the current concept of big data dates back to the beginning of this century, the underlying idea of data analysis for complex problem solutions comes from much further back. Strictly speaking, any effort to apply science beyond theory to understand and solve complex problems is an antecedent of data science and data analysis. Some authors trace these efforts back to the 17th century with the "point problem" texts by Christiaan Huygens and Blaise Pascal, which laid the foundations for problem solving with early statistical methods (Bell 1947). Also mentioned in this regard is Charles Babbage's research on the cost of transporting and sorting mail, which gave rise to the "Penny Post" or universal postal tariff in England in 1840 (Staff 1993). In the business world, Richard Devens mentioned for the first time in 1865 the term "business intelligence" to describe how the banker Sir Henry Furnese made profits by receiving and acting on information about his environment before his competitors (Devens 1865, 210). Beyond these first attempts and with the advent of World War II, countries in conflict began to create multidisciplinary groups of scientists to solve the arising problems through the comprehensive analysis of data. Thus were

created the first theories, principles, methodologies and groups of what began to be called "Operations Research" or OR. One of the tools that the conflict itself brought about was the electronic computer, which fostered the analysis of more and more data in a shorter time.

"Operations Research" was basically defined as a discipline which applied advanced analytical methods to find optimal solutions to complex problems involving decision making, using specialized techniques such as mathematical modeling and optimization, as well as statistical analysis. Therefore, OR overlaps with other disciplines, methodologies and tools such as linear algebra, mathematical simulation, industrial engineering, queueing theory, Markov chains, tree theory, economic modeling, knowledge management, stochastic variables, etc. By its very nature of being based on data analysis, it was strongly intertwined from the very beginning with computer sciences, such as software engineering, expert systems, algorithm design, data structure, etc.

After the conflict ended, knowledge created by OR quickly moved into industry, commerce, economics, meteorology, finance, and so on. Its principles and techniques were used to solve complex real-life problems involving large numbers of variables in various productive or government sectors. The rapid and relentless development of computer hardware and software capabilities in the second half of the last century meant that problems posed and the amount of data handled could grow exponentially. The name "Operations Research" gradually disappeared in the second half of the 20th century as new approaches, theories, tools, etc., used for data analysis were developed, giving way to new specialties with other names; however, OR laid the foundations for these new models.

In the 1960s, the basic principles and theories that gave rise to today's data science, information management and data management were created. In 1962 John Tukey wrote:

> […] for a long time I thought I was a statistician, interested in inferences from the particular to the general. But as I have watched

> mathematical statistics evolve, I have had cause to wonder and doubt… I have come to feel that my central interest is in data analysis… Data analysis, and the parts of statistics which adhere to it, must… take on the characteristics of science rather than those of mathematics… data analysis is intrinsically an empirical science… How vital and how important… is the rise of the stored-program electronic computer? (Tukey 1962, 2-4).[3]

As the decade progressed the answer was given, as computers and programs were developed and the foundation for database management was laid. John Tukey is credited with expressing the first ideas that gave rise to data science.

The concept of "database" was thus created during that decade of the 1960s. Although the Oxford English Dictionary online cites a 1962 report by the "System Development Corporation of California" as the first use of the term "database" in its specific technical sense, several early pioneer accounts mention that in the late 1960s the development and concept were still quite incipient. Among others, Mike Blasgen mentions: "[...] around 1968, already San Jose was doing work in database. But it was not called that, then. It was called 'data management' or 'file systems'…" (McJones 1995, 7). From the study of the most relevant texts on the subject, it is inferred that the concept really took off in the first half of the 1970s, and its commercialization and use occurred in the second half of that decade. The "Relational Database Model" was perfected in 1973 at the "IBM Research Laboratory" in San Jose, California, called at that time "System R", and which would give rise to the famous SQL – "Structured Query Language". Oracle, –still under the name "Relational Software, Inc."– began marketing the concept in 1979.

Edgar Codd published in 1970 a text that laid the foundations for the development of the database concept, becoming a milestone in database management. Until then, data were handled in

---

3 John Tukey is also credited with coining two basic computer terms: "bit", as a contraction of "binary digit", and "software", as the intangible counterpart of "hardware".

computers with structures under programmers' free will. The author stated there for the first time:

> [...] Future users of large data banks must be protected from having to know how the data is organized in the machine, i.e. the internal representation. A prompting service which supplies such information is not a satisfactory solution. Activities of users at terminals and most application programs should remain unaffected when the internal representation of data is changed... (Codd 1970, 377).

It should be noted that the author uses the term "data banks" rather than "databases", a concept that he would explain precisely there. Prior to this point, data and their exploitation programs were intrinsically intertwined. The main importance of all of the above is that from the development of databases, by adding certain structures to the data, it would already be separated from its programs, thus initiating a whole new stage in the management and exploitation of data by means of computers, giving an enormous boost to its management.

The development of databases can be divided into three main periods based on the data models or structures: navigational or hierarchical, SQL or relational, and post-relational databases. The many theories of hierarchical databases and the first computer systems for their management, known as "Database Management Systems" or DBMS, were developed by various manufacturers. They would evolve into relational, object-oriented, distributed, etc., databases, with countless specialized products to date.

In 1974 Peter Naur published his "Concise Survey of Computer Methods", a treatise on the computational methods of data processing at the time. He defined then: "[data is] a representation of facts or ideas in a formalized manner capable of being communicated or manipulated by some process". Something very relevant in this work is the author mentioning in his preface that in a congress in 1968 a plan of action was presented, entitled "Datalogy, the science of data and data processes and their place in education", as well as in the text of such work the term "data sci-

ence" was already used freely. He also offered the oldest definition of "data science": "[...] the science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences" (Naur 1974). Nowadays data science covers even more fields: data collection and selection, data management and preservation, big data sources, application of data mining techniques, "data warehouses", trend detection in social networks, human-computer interaction, data analysis and visualization, data and information quality assessment, and even information policies.

"Data management" also started in the 1960s. At its inception it dealt with theories, tools and methods for data processing with computer systems. Donald Knuth, one of the greatest developers on this subject, stated at the time that the key to proper data management –in such original meaning– consisted fundamentally of programmable algorithms and systematized formal mathematical techniques. For this reason, he wrote a whole series of books on these two elements, which are now considered fundamental works of deep computer-aided data processing tools.[4] Gradually the concept evolved –as well as other associated ones– into broader, multidisciplinary concepts, taking its current form. Today, data management is a theory and a management practice which consists of collecting, validating, organizing, storing and utilizing data in a secure, efficient and cost-effective manner, to turn it into a valuable resource within an organization. The goal of data management is to assist people and organizations in making decisions which maximize the benefits to them. To be effective, data management requires a "data strategy" as well as pre-established and reliable methods for handling and accessing data: collection, standardization, organization, storage, security, etc., all of which will facilitate its proper analysis. A solid data management strategy is essential, since collecting large amounts of data without concert makes them useless and unwieldy in a short term; its true value

4 "The Art of Computing Programming" is a series of twelve books published by Knuth on the subject starting in 1970.

does not depend on its mere existence within an organization, but rather on what can be done with it. As can be seen, data management evolved from an exclusively computer-related term into a complex and multidisciplinary conceptual structure.

In recent years, "Data governance" has been added to the list of concepts. The "Data Governance Institute"[5] defines it as "[…] a system of decision rights and accountability for information-related processes, executed according to agreed-upon models describing who can take what actions with what information, when, under what circumstances, and by what methods" (Recovered from: http://www.datagovernance.com/defining-data-governance/). More on this later.

In 1989, before the World Wide Web, Roy Davies already pointed out some of the first aspects of knowledge extraction from library catalogs (Davies 1989). In the same year Edward Feigenbaum, pioneer of Artificial Intelligence, considered libraries of that time "[...] warehouses of passive objects, where books and magazines sit on shelves waiting for some person to use his intelligence to find them, interpret them, and make them eventually divulge the knowledge they hold". He also made envisions of a "library of the future" where books would interact and collaborate with the user through an intelligent computer system that would be able to interact with several users simultaneously (Feigenbaum 1989, 122). By the late 1980s, the concepts of "knowledge extraction", "knowledge discovery" and more specifically "knowledge discovery in databases" began to emerge, marking a new stage in information processing, but now including significant data processing components, both in hardware and software, techniques and methodologies, algorithms, etc. These concepts and their derived techniques strongly contributed to consolidate data to obtain its contemporary value and relevance.

In 1996, Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth published a very important work on the subject: "From

5 The Data Governance Institute. http://www.datagovernance.com/defining-data-governance/

Data Mining to Knowledge Discovery in Databases". They stated there:

> [...] Historically, the notion of finding useful patterns in data has been given a variety of names, including data mining, knowledge extraction, information discovery, information harvesting, data archeology, data pattern processing.... In our opinion, KDD - Knowledge Discovery in Databases refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from data... the additional steps in the kdd process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining, are essential to ensure that useful knowledge is derived from the data. Blind application of data mining methods (rightly criticized as data-dredging in the statistical literature) can be a dangerous activity, easily leading to the discovery of meaningless and invalid patterns (Fayyad, Piatetsky-Shapiro & Smyth 1996, 39).

The following year, in 1997, they would publish the journal "Data Mining and Knowledge Discovery". Interest in the subject would increase rapidly: in the summer of the 1999 issue of "Library Trends", and in the November 1999 issue of "Communications of the ACM", the central theme was precisely "Knowledge Discovery in bibliographic databases" (Qin & Norton 1999).

With regard to the term "data mining", it is considered one of the stages of the aforementioned Knowledge Discovery in Databases. It comes from the terms used in the 1960s by statisticians: "data fishing" and "data dredging". In the 1980s the term "database mining" began to be used, but when this term was patented as a commercial product it evolved to "data mining", taking its current meaning since the 1990s. As can be seen in the aforementioned publication by Fayyad and his subsequent journal, the concept was already quite common by the mid-1990s. Han *et al.* defined it as:

> [...] Data mining is an interdisciplinary subfield of computer science and statistics... it uses methods from machine learning, statistics, and database systems.... It consists of the process of discovering patterns in large data sets with intelligent methods for the general purpose of extracting information from those data sets and transforming it into understandable structures for further use (2011, 15-21).

A little more than a decade ago the term was introduced into librarians as "bibliomining", a concept that will be discussed later.

In the same year, Hayashi presented a paper entitled "Data science, classification and related methods". He defined there:

> [...] data science is not only a synthetic concept to unify statistics, data analysis, and their related methods; it also encompasses their results. Data science aims to analyze and understand real phenomena with 'data'. In other words, the goal of data science is to reveal the hidden characteristics or structure of complicated natural, human and social phenomena using data, from a different perspective than traditional established theories and methods (Hayashi 1996, 41).

This is one of the early definitions of data science. In 2002, a journal specialized in this subject was created under this very name: *Data Science Journal.*

In a text from year 2000, Noboru Ohsumi mentioned that since 1992, he had pointed out: "[…] it is important to emphasize that we agree on the need to develop, through practice, research on the theory and application of data analysis into a new 'data science'" (Ohsumi 2000).

To go deeper into early data science development two excellent overviews are recommended: Murtaugh & Devlin (2018), and Cao (2017). The latter (Cao (2017, 15) mentions that the first conference having data science as a subject was the 1996 "IFCS Conference on Data Science, Classification, and Related Methods" (Hayashi *et al.* 1996). He also presented a full, modern definition of "data science" summarizing different insights of the concept:

> […] From the disciplinary perspective, data science is a new interdisciplinary field that synthesizes and builds on statistics, informatics, computing, communication, management, and sociology to study data and its environments –including domains and other contextual aspects, such as organizational and social aspects– in order to transform data to insights and decisions by following a data-to-knowledge-to-wisdom thinking and methodology (Cao 2017, 8).

A further factor that undoubtedly contributed to all of the above was the immense and relentless growth of data storage capacities throughout the 20th century and so far this century. Collecting large amounts of data is impossible if there is no place to store it. For the same reason, the concept of "big" or "massive" with respect to data has been closely linked to the capacities of the storage devices of each era and has been changing with each advance. The starting point was Hollerith's 1890 invention of the U.S. census punch card. The previous one had taken almost ten years to be counted, and in view of the growth of the population –from 50 to 63 million– it was considered impossible to carry it out in only ten years. The 63 million punched cards were processed with mechanical tabulators in only three years. Since then and for decades, cardboard punched cards became the ideal medium for storing data which could be processed with the help of tabulating, sorting, accounting, and, finally, computing machines. Useful as they were, however, cards also set a physical limit: 1 Gigabyte of data on punched cardboard cards required 12.5 million cards completely filled, weighed 31.25 tons, and occupied a volume of just over 31 cubic meters. The next step was the invention of magnetic tapes in the late 1920s,[6] which were also widely used for several decades, and came to store in their best versions some 30 Megabytes per unit. They were succeeded by magnetic drums, with capacities from 50 to 100 Kilobytes; they were followed by magnetic disks, which started with about 4 Megabytes of capacity in the

6 Fritz Pfleumer, an Austro-German engineer, invented in 1928 the way to store magnetic information on tape. His principles are still valid today.

1950s up to several Terabytes today, then by Linear-Tape Open or LTO tape cartridges, with current capacities of up to 24 Terabytes each, and later by solid state memories. In computer technology there is a well-known principle called "Moore's Law"[7] which basically states that the capacity of a computer's central processor doubles every 18 months. Mark Kryder –former Chief Technology Officer of the large hard disk drive manufacturer Seagate– similarly stated that the amount of data storage which can be placed on a certain area of magnetic media also doubles every 18 months. Although this is no longer true today, the storage capacity increases are still staggering.

In addition, costs have dropped to unprecedented proportions: 1 Gigabyte of magnetic storage on disk did cost over a hundred thousand dollars in 1980; the same capacity costs just under three dollar cents in 2020. On tape it can go as low as 1 cent per Gigabyte. In this regard, Morris and Truskowsky (2003, 206) stated that since 1996 the inflexion point was reached, where electronic storage became more cost-effective than paper.

7 Empirical principle established by Gordon E Moore, co-founder of Intel, in April 1965.

Concept

Aim for simplicity in Data Science. Real creativity won't make things more complex. Instead, it will simplify them.

DAMIAN DUFFY MINGLE

The term "data" comes etymologically from the Latin noun *"datum",* which translates as something "given" or "established". In simple terms, a "datum" is a symbolic representation of the attributes of an entity, fact or event, taking the form of a quantitative or a qualitative variable; it is the minimum expression of content about a subject. When data is considered and analyzed as a whole and in context, they constitute information; therefore, data is collected and grouped. When data was processed manually, obviously the amount that could be analyzed was very limited. With the advent of electromechanical machines from Hollerith onwards, data processing increased substantially, and with the advent of computers, the amount of data grew even more in function of the capacity of those devices up to the handling of immense quantities of the last few years: the "big data".

In this regard, Lohr (2013) mentions: "[...] the term big data, which encompasses computer science, statistics and econometrics, was first used in conferences given by John Mashey of Silicon Graphics in the mid-1990s" (Mashey, 1999). Kenwright (1999) presented a paper at the "Visualization '99" Conference which already included this term. Diebold (2013) states that "big data" should be

understood as a phenomenon, a term and a discipline at the same time. He mentions that he found in the literature some few references prior to the year 2000 in this regard, both academic and non-academic, where the term is used but the phenomenon is not thoroughly known, and adds that –on the contrary– in those years some academics were aware of the emerging phenomenon but did not use the term. Obviously the discipline would be created later.

Currently it is accepted as the starting point of the concept a note published in February 2001 by Doug Laney, an analyst specialized in information of the Meta Group, part of the Gartner Group, where he outlined the fundamental characteristics of this concept, widely used over time – Laney (2001). The Gartner Group's definition of big data has since been widely accepted for many years; it already included the three fundamental characteristics of this type of data established by Laney, known as the three "V's": "[…] big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation" (Gartner Glossary 2005).

Volume refers to the vast production and accumulation of data worldwide in unprecedented and ever-increasing quantities. *Velocity* refers to the rapid rate at which data is created daily: millions of web pages, messages, social networks, news, emails, just to name a few. *Variety* has to do with all imaginable types of data produced every day from numerous sources and formats: the innumerable ways of representing data create a serious problem in their interpretation. These three concepts will be discussed further below.

Continuing with definitions, Oracle, the highly specialized data management company, defines:

> "[...] Big data are increasingly large and complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software cannot handle them. But these massive volumes of data can be used to solve business problems that previously could not have been addressed" (Recovered from: https://www.oracle.com/big-data/what-is-big-data.html).

Dans (2011) states: "[...] big data refers to the processing and analysis of huge data repositories, so disproportionately large that it is impossible to treat them with conventional database and analytical tools".

Gantz & Reinsel (2012), experts from IDC Corporation defined:

> [...] big data is a new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data, enabling its capture, discovery and/or analysis at high velocity.

Martínez Musiño (2020, 96) elaborated a definition of big data from the perspective of information sciences: "[big data] is the phenomenon of massive and constant generation of data, whose treatment and organization require both technological resources and specialized computer programs and interpretation and analysis tools to achieve scientific accuracy".

The aforementioned Forbes magazine, which compiles lists that arouse great interest in the technical, business and economic sectors, published an article in September 2014 in which it mentions twelve different definitions of the term "big data" (Recovered from: https://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/#445380d913ae).

De Mauro, Greco & Grimaldi (2016: 131) established a synthesized definition based on 15 formal definitions analyzed, which they classified into 4 groups of basic characteristics: 1) information, 2) technology, 3) method, 4) impact. From these characteristics they elaborated the following definition: "[...] big data is the information asset characterized by such a High Volume, Velocity and Variety to require specific technology and analytical methods for its transformation into value". Although it is a synthesis of many definitions, it can be seen that it does not greatly differ from the previous ones, and does not contribute with something different.

Google define the term as:

> [...] extremely large data sets that can be computationally analyzed to reveal patterns, trends, and associations, especially in relation to human behavior and interactions... the current use of the term 'big data' tends to refer to the use of predictive analytics, user behavior analysis, or certain other advanced data analytics methods that extract value from data, and rarely to any particular size of data sets.

In contrast, it is worth mentioning that some authors have tried to visualize these data definitions starting from the opposite approach; i.e., defining "non-massive data" or "small data", to move from them to the definitions of big data. As an example, Pollock (2013) stated: "[...] non-massive data is data whose quantity can be conveniently stored and processed on a single computer; specifically, on a high-performance server". These type of definitions have as a common characteristic the organization's ability to locally manage its data within its own computers, without resorting to external providers or services.

In conclusion, as with many other complex technological concepts of this era, there is no universal or consensual definition of the term "big data", as established in detail by Zhan & Widén (2017, 569). However, aside from the numerous existing definitions of the concept, a characterization can be made with the common elements found among them: 1) big data consists in the processing and analysis of such large, varied, complex and disparate data sets, 2) produced at such a fast speed and from so many diverse sources, 3) that the "traditional" information processing equipment, programs and procedures, such as servers, databases, search engines, etc., are not sufficient and 4) therefore require much more powerful, sophisticated and specialized methods, equipment and programs to collect, analyze and correlate them, 5) all in order to be able to quickly extract patterns, trends and associations from these data, mainly from human behavior and interactions, and 6) from there to be able to take informed decisions to assist organizations, 7) which gives the data an enormous added value.

From this characterization it is possible to analyze in greater depth its seven components to achieve a better understanding of the concept of big data. Moreover, as Diebold stated, it is not only a concept; it has become an entire discipline, which is worth explaining.

The fundamental core of the characterization is the collection, organization, storage, and analysis of data, whose essential characteristics were stated in the three "V's": Volume, Velocity, and Variety. It is worth analyzing in more detail what these three elements refer to.

Volume: in recent decades the world has experienced an unprecedented increase in global information production, which also includes data. The first massive count in this regard is the well-known study made by Lyman & Varian (2000), who established that during that year the world generated one Exabyte (EB) of data. Although the exact amount of data per year is very difficult to estimate and there are various figures on the subject, there is a series of studies to which multiple references have been made, carried out by the IDC Corporation from 2007 to 2020. They account for the vertiginous growth in the volume of digital information. They estimated that the world produced 130 Exabytes (EB) of information in 2005, 1,227 in 2010, 2,837 in 2012; 8,591 in 2015 and 40,026 Exabytes in 2020 (Gantz & Reinsel 2007), (2012).[8] Meaning this that the amount of information produced in the world has doubled approximately every two years during the last two decades. As a reference, consider that one Exabyte of information is equivalent to 500 trillion[9] pages of text of 2,000 characters each, or two trillion books of 250 pages each, or one thousand trillion e-mail messages of one thousand characters each, or one trillion web pages of one Megabyte each, or 333 billion high-resolution photographs, or 250 billion mp3 music files of 4 minutes each, or almost two billion CD-ROMs, or 3,500 trillion tweets of 280

8 1 Exabyte = 1,000 Petabytes = 1'000,000 Terabytes = 1,000'000,000 Gigabytes = 1'000,000'000,000 Megabytes = 10^{18} bytes or characters

9 1 USA trillion = 1,000 billion = 1'000,000 million.

characters each. In other words, the world is flooded with information and data growing exponentially.

Velocity refers to the enormous rate at which data is created on a daily basis: millions of web pages, messages, social networks, news, emails, videos, images, just to mention a few. Obviously, producing such volumes of data necessarily requires that they are produced at breakneck speed. In the last two years, an average of 27 Exabytes per day have been produced. That is, 27 followed by 18 zeros of bytes every day. The site "We are social" mentions that in 2020, within the 4,500 million Internet users that existed in the world there are 3,800 million users of social networks (Recovered from: https://wearesocial.com/digital-2020). Among that group and as an example, Twitter alone generates more than 7 Terabytes of data (TB) daily, and Facebook generates 10 TB of data every day.[10] 200 billion emails were sent every day last year and 100 million images are uploaded daily to Instagram. Much of the source of this velocity comes from the users themselves, but in addition to this, much more data is already created about them by the networks in which they move, and even more is created by machines and objects connected to the Internet. Physical devices and objects have joined people and organizations in actively participating in the production of the enormous flow of data incorporated into the digital universe on a daily basis.

Finally, *Variety* is related to all possible ways to represent data, since obviously the world produces every type of imaginable data in countless formats: books, journals, newspapers, maps, music, archives, photographs, catalogs, social networks, chats, messaging, stores, electronic games, phone or video calls, TV, radio, movies, events, tutorials, websites and portals, videos, blogs; health, academic, and journalistic information; banking, taxes and commercial transactions; mail, and so on. This information is produced by people and organizations as well as by machines. The main problem associated with such variety is that there are multiple ways of

10 1 Terabyte = 1,000 Gigabytes = 1'000,000 Megabytes = 1'000,000'000,000 bytes = 10^{12} bytes or characters

understanding and classifying data: according to their form they can be quantitative or qualitative; according to their source they can be captured, derived, exhaustive or transitory; according to their type they can be primary, secondary or metadata; according to their representation or structure three types of data are considered: structured, unstructured and semi-structured. The former come from organizations and systems designed for that purpose: oceanographic and meteorological data, seismological, astronomical, banking, tax, etc., and because of this well-specified structure, they are easier to process, as they are homogeneous, standardized and ordered; these elements facilitate their management. Examples are a date in dd/mm/yy format, a ten-digit telephone number or an ISBN (Joyanes 2013; Olavsrud 2012). Unstructured ones come from web content, social networks, forums, e-mails, simple text files, spreadsheets, audios or videos, blogs, voice messages, instant messages, and so on. They do not have defined types or are not organized under any pattern; they are not stored relationally, or within a hierarchical data base; they do not have a certain standardized format, and it is not easy to identify their type or class. In order to process them, it is essential to organize and classify them, and the only way to do so is when they contain metadata from the source or they can be aggregated with some ease. Between these two extremes there are data with intermediate structure or semi-structured. All of these will be discussed in more detail later.

Many authors agree that big data cannot be entirely described with only these three characteristics, so they have been adding new ones to the originals which they considered relevant. Marr (2015), Lomotey & Deters (2014), and Affelt (2015) established that there are five "V's" or characteristics; they added to the original ones *Veracity*, which refers to the reliability or deviation of the data; other authors who also included it are Lawlor (2016), and Plale (2013). This characteristic refers to the fact that when large amounts and variety of data are produced, the quality and accuracy are less controllable, and there may be great bias and inconsistencies in them; for example, Twitter messages with hashtags, "trends", "fake

news", etc. According to Marr, data collection and analysis methods must therefore include mechanisms for data cleaning. Affelt (2015, 21) calls this feature "Verification" but it refers to the same issue: "[…] the process by which librarians and information professionals analyze data sources and retrieval systems to determine data quality". Marr, Lomotey & Deters, and Affelt also added *Value* as the fifth characteristic of big data: it refers to the real capacity to extract valuable and useful information from the data as a whole, since there is no point in collecting and processing it if no tangible benefit can be obtained for the organization. Marr affirms that this one is the most important characteristics among the five, since many organizations have fallen for the attraction of big data without having sufficient knowledge and capabilities, making substantial investments in it with little return. Cost/benefit ratios are critical in this regard. Many other authors agree with these two added characteristics and although some added even more, such as *"Variability"*, *"Volatility"*, *"Vagueness"* or *"Complexity"*, there is a consensus that the five characteristics already described are the minimum required to properly deal with big data.

Beyond definitions and characteristics, it is worth stressing at this point that the current trend in big data conceptualization emphasizes that its importance does not really lie in the techniques and tools for handling lots of data, but rather in how to extract value from them, how to make them useful to organizations. It is essential to think about how to capitalize on such data so that their potential can be truly exploited for future critical organizational decisions. Only by being able to effectively organize and exploit such asset will it be possible to obtain greater intelligence in the organization by enabling better and faster decision making. The concern for a better and greater use of data has been creating a series of trends, specialties, etc., around it.

In recent years, a trend within the field of big data called "thick data" has emerged, which establishes that in order to be fully exploited, big data need to be complemented with a qualitative approach in addition to the quantitative one, allowing the understanding of contexts, feelings, stories, opinions, emotions and the

models of the subjects' environment studied. Usually, big data is processed by mathematicians, statisticians and computer scientists and is based on mathematical and computational processes; "thick data" is processed by anthropologists, ethnologists, sociologists and social scientists, and emphasize the use of other research tools specific to these disciplines for data analysis: surveys, questionnaires, interviews, focus groups, personal diaries, and so on. The concept of "thick data" was coined by Tricia Wang based on the postulates of social anthropologists Clifford Geertz (1973) and Gilbert Ryle (1949), who established the concept of "thick description": a set of qualitative and ethnographic research techniques for the social sciences whose purpose is to elaborate detailed descriptions, contexts and interpretations of situations studied by a researcher. Closely related to this, recent years have seen the development of the discipline known as "netnography" –from "net" and "ethnography"– which adapts ethnographic research techniques for studying processes, relationships and cultural practices, virtual communities, phenomena and specific dynamics found on the World Wide Web, through the analysis of the data available on it. Paul Gibbons (2015) summarizes all these concepts splendidly in a single sentence: "The human side of analysis is the biggest challenge for the implementation of big data".

On the other hand, this concern for a better and greater use of data has created a whole professional specialty around it within data science, specifically called "data analytics". Currently, postgraduate studies, courses, diplomas, etc., have been developed around this topic by countless institutions. Basically, the purpose of this specialty is the same as data science in general, i.e., to detect trends and patterns in the data in order to propose solutions. More precisely, this specialty refers to the task of identifying which variables of the organization can be linked to certain data and thus establish correlations for the posing of questions and the eventual obtaining of solutions through specific techniques. These latter constitute the core of data analysis as a specialty. They will be discussed in more detail below.

The importance of Big Data

An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.

JOHN TUKEY

The importance of systematized data analysis to optimize decision making in organizations, especially with big data, has already been mentioned, and there is no doubt about the interest that this topic has aroused in the last two decades. Kalantari *et al.* (2017) conducted a bibliometric study in which they mentioned that they found 6,572 papers on the topic of big data in the Web of Science alone. Xu & Wu (2019) found in a similar study 10,989 documents on the subject; both studies are a clear indicator of the interest and development in this regard. Several journals devoted entire issues to the topic, such as "Research Trends" as early as 2012 (Recovered from: https://www.researchtrends.com/wp-content/uploads/2012/09/Research_Trends_Issue30.pdf).

Since 2015 Springer began to publish a journal exclusively on this topic: "Journal of Big Data"; since 2011 the first "Glossary of Big Data" already appeared (Warden 2011), and in 2013 the first "Big Data Handbook" did (Plunket *et al*. 2013). Halevi & Moed (2012) made a review of the existing literature in that year about big data as a scientific and research topic, and it can be seen from their results that the subject had already gained notoriety since then. It is now appropriate to delve into the main uses of this type

of data in general sectors in order to appreciate all its potential, and from there to glimpse its uses and benefits in libraries.

The process, use and analysis of big data is nowadays widely used in countless sectors of society, so it is not intended here to present an exhaustive list of all its practical applications, but rather to make a selection that is sufficiently representative and illustrates its uses and benefits.

1.- Financial, banking, and insurance sector
Some examples of the use of big data in this sector are: financial risk analysis, economic models, actuarial analysis for insurance, detection of fraudulent movements with credit cards, development of new personalized services for banking and insurance, among others. Today's digital banking platforms and all the related operations that can be done from home are also the result of data analysis.

2.- Industrial sector
Some uses of big data in this sector are: optimization of manufacturing processes, risk analysis, diversification or integration of business in corporations, demand and consumption forecasts, timely data from GPS or RFID - Radio Frequency IDentifiers, instant scheduling and tracking of deliveries and shipments, robotic-controlled manufacturing, 3-D printers, data feedback from "Internet of Things" products and "Wearable" personal devices, Artificial Intelligence, optimization of energy efficiency, among others.

3.- Communications and transportation
Some examples of the use of big data in this sector are: prediction of short and long term demand for Internet and telephony services –servers, data centers, channels, networks, bandwidth, the cloud, etc.–, as well as transportation services: airplanes, trains, buses, roads, ships, and so on; calculation of shipments, routes, airports, seaports, and others, to optimize times, shipping, fuel, stevedoring, and so on; instant reprogramming of resources for optimization based on usage and demand: servers, networks, channels,

hubs, and others. As an example, the cab company Uber uses the analysis of big real-time data to dynamically change the fares of its trips according to demand at any given moment.

They are also used to create new services for different target audiences and make recommendations of those services to potential new users, measure resource performance, and so on. In fact, new communications technology such as 5G[11] includes a lot of massive data analysis for dynamic resource allocation.

4.- Meteorology

Meteorology, oceanography, volcanology, services, incessantly collect large amounts of data to feed their models of long and short-term weather behavior in every part of the world, as well as for specific events such as hurricanes, tornadoes, eruptions, avalanches, etc. These models require huge amounts of real-time data from countless sensors to be updated frequently and to calculate trajectories, make timely warnings, etc.

5.- Government

The government sector has taken great advantage of data analysis. Data are used for many different tasks: design and implementation of services for citizens, monitoring and analysis of tax activities, design and monitoring of health services, education, transportation, etc. Major initiatives, such as "Open Government" rely significantly on massive processing of data. More specifically, data analysis in this sector allows changing parameters instantly for traffic control, traffic lights, access, roads, etc., based on street sensors, GPS, video cameras, and others. It allows to measure the flow at the moment of services such as bicycle rental or other public vehicles, to exchange them and reposition them appropriately between their stations. It allows to reschedule for each stage of the day the number of units to enter into service in subways, buses,

11 5G is the Fifth stage of telecommunications technologies for cell phones and Internet. It mainly involves higher bandwidths (up to 10 Gigabits per second) and greater availability to many users at the same time.

and other public transport. It facilitates to send timely alerts to citizens in the event of earthquakes, storms, floods, etc. It allows huge and instantaneous data from multiple points about air pollution to take short-term measures, among others.

6.- Health sector

In the health sector, a recent and transcendental example of the use of big data has been the design and development of the new pandemic vaccines; without the concurrence of these data exchanged and analyzed in so little time by various transdisciplinary groups all over the planet, these vaccines could not have been created and produced in such a short period of time.

In public health services, analysis of big data makes it possible to design health policies and structures at an international, regional and country levels, to visualize the health status of large sectors, to design and follow up vaccination and sanitary campaigns, etc., segmenting by region; to visualize and follow up disease outbreaks and their treatment, to predict public health eventualities and to allocate resources for their attention; to build specific epidemiological surveillance models; to purchase and maintain large inventories of medicines and other resources, etc. This includes the collection and analysis of millions of diagnostic and imaging data, prescriptions, data related to known allergies, demographics, clinical descriptions, and laboratory test results. A recent applied and transcendental example of these uses has been the design and development of new vaccines; without big data analysis they could not have been created in such a short time.

In private healthcare services, data analysis –as in other sectors– makes it possible to design the marketing strategy and follow trends. There are also more specific applications, such as measuring which treatments are most effective for certain conditions, designing new drugs for various purposes, and identifying their side effects. In both the public and private sectors, it allows the timely and instantaneous monitoring of hospitalized patients based on data sensors in them.

7.- Commerce

This is one of the sectors that has taken most advantage of data analysis. It is used to design and track products, coordinate shipping and distribution logistics at local, regional and global levels, and managing stocks and inventories. It enables the design and launch of marketing campaigns segmented by user groups and the detailed analysis of purchase events, trends and preferences. Allows for instant tracking of sales campaigns, user loyalty, advertising, etc.

Especially e-commerce has benefited from this tool. The success stories of Amazon, e-Bay, AliBaba, iStore, Mercado Libre, Wish, etc., are some clear examples of organizations using big data analytics for all stages of their business: product design, marketing, payment, shipping, tracking, user feedback, inventories, and so on.

8.- Entertainment, tourism

Several services in the leisure and entertainment sector such as Spotify, YouTube, Netflix, etc., use the analysis of big data collected from their millions of users around the world to provide informed recommendations to individual clients. Amazon Prime has been experimenting with big data to deliver a customer experience by presenting video, music and Kindle books in one place, further "personalizing" its offering with previous customer data. Electronic game manufacturers collect big data to measure product performance, user preferences and trends. Data analytics is also used to collectively identify the reason for subscribing or unsubscribing to a product, as well as the interest in a particular content, etc.

Major entertainment events: Olympic games, soccer or tennis championships, super-bowl, among others, use real-time big data analytics to measure audience, preferences, etc., and make adjustments to broadcast times, marketing, and more.

The tourism and leisure sector analyzes big data to observe and compare travel offers and prices, to analyze user trends and preferences, making recommendations to users and service providers, and so on.

9.- Food
As in other sectors, data analysis helps to design marketing strategy and follow trends. But there are more specific applications, such as finding hidden patterns and similarities which assist restaurants to detect potential customers. Image processing and artificial learning can identify the most sought-after location within restaurants; their owners can then highlight such area for advertising and marketing purposes. Many restaurants are already data-driven in their smart inventory or stock management system to order timely replenishment of stock. Fast food chains use them to predict the number of customers at specific times of the day in order that employees can anticipate preparation based on demand. When the line is long, it allows them to automatically modify menu boards to display only those dishes that can be prepared in a short time.

10.- Education
The education sector has also benefited greatly from big data analysis. Students and teachers leave a large data footprint from their activities on a daily basis. Among the main uses of such data are: the design of educational policies and structures at international, regional and country level; as well as visualizing the state of this sector in large cross-sections. They allow educational institutions to design and implement new curricula, courses, subjects, assignments, etc. They also facilitate the monitoring and feedback of the educational sector and systems. As a matter of fact, the Publications Office of the European Union highlights this type of educational monitoring as the greatest achievement of big data analytics (Berendt *et al.* 2017). They assist managers and teachers in the production of Learning Management Systems, as well as to be aware of and give feedback on the success of students in their courses. They assist teachers in the creation of assignments, exercises and exams using already existing information online. They also allow social networks to be incorporated into the teaching process, analyzing students' blogs, uploads, messages, “likes”, etc., integrating relevant tools to measure how much they are interested in

a specific topic or course. Parry (2018) studied how universities use big data to attract and retain students, assisting them to choose their courses, and providing relevant and timely advice. For students, such data allow them to visualize the curricula and road maps of a certain degree or plan and –with the pertinent recommendations– design the best path, timing, priorities, etc., for obtaining their degree. A comprehensive overview of all these applications in education can be found in Baker (2015).

Much of the day-to-day data production of students comes from their activities of seeking academic information for their needs; obviously this part is closely related to libraries. Because of its nature and particular interest, the analysis of uses and applications of big data in libraries will be treated in a special section.

The aforementioned are not all the sectors where big data is used nowadays; they are only the main ones. In addition, even though libraries were included in this list in the education sector, it should be noted that in fact there are libraries and information centers in each and every one of the named sectors. Nor are these all the possible uses of data analysis within them; it is only a representative sample with some specific examples to illustrate all the current possibilities already existing in this type of analysis. For all these reasons, it is valid to assume that big data can also be useful and interesting for libraries and their professional staff and therefore, it is worthwhile to study them.

Big Data in libraries

In a sense, the world of libraries is a microcosm of the wider world, buoyed by technology but daunted by the unknown, changing in ways that most of us understand dimly, if at all

MICHAEL GORMAN
Our Enduring Values Revisited:
Librarianship in the 20th Century.

Big data is becoming an ever-increasing presence in the field of information-related organizations and therefore is already having an impact on all types of libraries. Nicholson & Bennet (2016, 86) mention: "[...] the rapid increase in the volume, velocity, and variety of library data generated by different library tools offers innovative ways of understanding interactions with users in the library environment". In fact, although the ultimate purpose of libraries is patrons, there are numerous additional applications of big data in practically all their fields and tasks. Several authors have already reflected on this aspect, pointing to the possibilities and projects initiated in this regard. Data analytics –especially with big data– can generally be used in libraries in the same way as in other sectors, in many of the data fields: data mining, machine learning or Artificial Intelligence, statistics, visualization, etc. But these are just broad areas of data fields according to their techniques and tools. When cross-referenced with the different library fields this becomes a huge matrix with multiple specific subfields of application. For example, library data mining

or "bibliomining"[12] can be used for metadata, for extended catalogs, for users' follow-up, for bibliometrics and other types of impact analysis of all kinds of documentary materials, for "usability" testing of collections and services, for text analysis, to name a few uses. The same is true for the other areas of data analysis when applied in the library.

The IFLA (International Federation of Library Associations and Institutions), in response to the findings of its previous 2013 Trends Report proposed the creation of the Big Data Special Interest Group during its 2014 World Library and Information Congress or WLIC in Lyon, with the aim of making libraries a proactive part of the data movement and not just bystanders (Recovered from: https://www.ifla.org/big-data). The group was formally established during WLIC 2015 in Cape Town and has since conducted a series of studies, events and papers on the subject.

The main precedent for the use of big data in libraries –still a current project– is the Worldcat global catalog, operated by the OCLC organization, which, according to data from the site itself, contained in 2019 450 million catalog records in almost 500 languages from almost 18,000 libraries around the world; the catalog also contains inventories of 2.8 billion items in those libraries. In addition, it has used entity-relationship structures to create linked data from its holdings, identifying the entities in its records and then assigning relationships among them. Obviously this catalog was not originally conceived as a big data project, and it is not the product of a single institution at a single point in time; it is a collaborative and cumulative endeavor, which undoubtedly represents the archetype of big data in libraries.

Another collective library big data endeavor is the HathiTrust digital library. Beth Plale (2016), Co-Director of the Research

12 *Bibliomining* consists of data mining and bibliometric techniques used together to extract meaningful patterns, trends, relationships, etc., from library systems and data. These techniques include: identification of topics or objects of interest, creation of a data warehouse, its refinement, processing, analysis, as well as obtaining and evaluating the results.

Center for this project described it in that year: "[...] the collection is big data in size. To step through all the nearly 14 million digitized books in 24 hours would require 14,000 computers running simultaneously". In 2020, the HathiTrust site already lists more than 17 million digitized items. Another significant example of big data in the library field is the collection named "Internet Archives", which is a non-profit system that began in 1996 to store web pages in order to prevent their total disappearance, and later extended to other digital or digitized materials: to date, it reports that it manages 330 billion web pages, 20 million books, 4 million audios, 4 million videos, 3 million images, and 200,000 computer programs: in total, more than 45 Petabytes or 45 x 1015 bytes of data (Recovered from: https://archive.org/about/).

Several authors agree on two main aspects of the use of big data in libraries: one direct and the other indirect. Jharotia (2016, 3) mentions that the direct effect exists in the use of specialized tools to analyze the large data sets coming from the libraries themselves. The indirect effect is through their patrons who increasingly require the use of products and services derived from big data in their information searches. Olendorf & Wang (2017, 191-192) stated something similar in this regard: the first strand consists of using big data in libraries as an auxiliary in their daily operations; they can use these data to improve their collections, make better use of space, evaluate their services, feed back into their instructional programs, and optimize the information provided to patrons. These authors also agree that the second area in which libraries can work with big data is the information search services for their users. In addition, some libraries are already beginning to provide numerous data services –many of them massive– for researchers and scholars, such as: design and planning of data management, collection, curation, storage and preservation. Because of their special interest, these types of services will be discussed in more depth below.

Going into further detail, there are many areas of opportunity in libraries with respect to big data. Blummer & Kenton (2018, 18-19) conducted a study in which they reviewed the specialized

literature that existed at the time about big data specifically in libraries. They found 76 papers, from which they extracted the four major themes discussed there: 1) management of big data in libraries (29); 2) provision of data analytics services by librarians (26); 3) informative papers about big data (13); 4) training opportunities for librarians in big data (8). In addition, they found eight sub-themes within the topic of big data management: 1) privacy and data management; 2) librarians' skills in privacy protection; 3) additional challenges in big data management; 4) assessment and needs detection for big data management; 5) collaboration in big data management; 6) the various factors that foster big data management projects; 7) research and studies about big data; 8) librarians' activities in big data (*Ibid*., 19). Obviously more papers have accumulated since then.

An entire issue of the library and technology journal "Library Hi Tech" was devoted exclusively to the topic of big data in 2018; Liu & Shen (2018) produced a detailed review of all its articles on the subject. The famed "Library Journal" has devoted numerous articles to the topic over the past few years. See as representative examples "Promise and Problems of Big Data" by Steven (2013) and "What Governmental Big Data May Mean for Libraries" by Schwartz (2013).

In addition to all the published papers, it is also noticeable the interest awakened by the topic in the library profession, visible through numerous events held on this matter in recent years. A good number of papers and sessions at the Special Libraries Association annual conferences[13] from 2014 to date have been focused on some aspect of data. Lawlor (2016) made a very comprehensive review of the papers at that year's NFAIS Annual Conference,[14] where data was discussed as a relevant part of research and knowledge.

---

13 See as an example: 24th Annual Conference and Exhibition of the Special Library Association. 2018. http://slaagc.org/slaagc2018/

14 NFAIS = National Federation of Advanced Information Services. Non-profit association of librarians, publishers, scholars, computer scientists, etc., in USA.

The fifteenth IEEE/ACIS International Conference on Computer and Information Science (ICIS) in 2016 had some panels discussing possible research topics in big data by libraries (Wang *et al*. 2016).

A large part of the uses of big data in libraries is being done with techniques from the field of "Artificial Intelligence" (AI). At first glance, many people –librarians included– when hearing the concept of AI in the library, think in something like a machine inside the call center answering everything is prompted, or a "personal assistant" robot installed in the lobby responding to everything it is asked, kind of a "Siri", "Alexa" or "Cortana" in the library.[15] While there are indeed some libraries that have dabbled in this variety of applications, it is merely a technological curiosity with not much effectiveness given their limited repertoire, and will still remain at this stage for quite some time. Nonetheless, they serve very well as "spotlight" applications to draw patrons' attention to other library services. See in this regard (Harada 2019*)*.

In practice, there are many variants and applications of Artificial Intelligence, from chess player machines, robots for industrial manufacturing, traffic control devices, image recognition systems, driverless cars, etc. Two main fields of study can be distinguished in this discipline: the theoretical field and specific applications. The first has to do with the whole concept and theory of the behavior and "intelligent" capabilities of machines, their philosophy and deontology, and so on. The second deals with the construction of specific applications to solve certain specific problems; this second variant is the one of interest in libraries, and it is not as distant as could be thought. There are many applications in everyday life. In fact, any personal computer or smart phone already has a good number of applications used by the public of which users are not aware that they are AI; for example, systems giving instructions for driving from one point to another in the city, offering

15 These so-called "personal assistant" applications use people's natural language processing to answer questions, make recommendations and perform certain actions by connecting to an ever-growing set of web services and other applications.

options of less time, shorter distance, cheaper route, etc. Another visible example of these applications are those suggestions from shopping sites –such as Amazon– indicating to the user personal details such as "people who bought that also bought this", or suggesting similar products in concordance with user's previous visits. A further everyday example is the spelling and structure suggestion systems built within document processors. An additional example of applied uses of AI are text translation programs, being the best known of them Google Translator, but in which there are other impressive products going far beyond simple literal translation, such as DeepL, Reverso Translation, Wordlingo, BabelFish, or Translation2.

The antecedents of AI can be traced back to the "automaton" figurines or dolls that have existed for more than twenty centuries.[16] In modern times, Alan Turing is considered the father of AI. He was the inventor of the "Bombe" machine for decoding encrypted messages from the German army during World War II, and designed his famous "Turing Test", a criterion by which the intelligence of a machine can be assessed when its answers in the test cannot be differentiated from those given by a human. The coining of the term "Artificial Intelligence" is attributed to John McCarthy in 1956, who introduced it at the first meeting on the subject; he stated there:

> [...] every aspect of learning or any other feature of intelligence can, in principle, be so precisely described that a machine can be made to simulate it… The most basic concept of ai defines any type of intelligence that does not arise through natural processes, or where intelligence can be understood and measured in such a way that it can be recreated (Dartmouth… 1956).

16 Heron of Alexandria wrote a book in the First century A.D. in which he explained the creation of mechanisms for entertainment imitating human or animal movement, doors opening automatically, etc. He detailed in there his "Automaton Theater", with mechanical puppets. Ahmad, Muhammad and Hasan bin Musa ibn Shakir compiled in 805 a book describing more than a hundred mechanisms and automatons. There are many more texts about it.

The most basic concept of AI defines any type of intelligence that does not arise through natural processes, or where intelligence can be understood and measured in such a way that it can be recreated. There are many definitions of the concept, depending on the intended focus and the discipline addressing it. For the field of libraries, the specialized definition and the study of some of its most used subfields in them are certainly more interesting than the canonic definition of Artificial Intelligence. Contrary to what one might think, the interest of libraries in the AI field is not new; three decades ago, in 1991, Charles Bailey already made an interesting and complete review of the applications of such technology in libraries (Bailey 1991).

Nowadays, there are numerous applications of AI in libraries, which should be studied and divided according to the major thematic chapters on library collections and services which deserve to be discussed more specifically.

Big Data in taxonomies and metadata in libraries?

> *Few people as yet, outside the world of expert librarians and museum curators, know how manageable well-ordered facts can be made, however multitudinous, and how swiftly and completely, even the rarest visions and the most recondite matters can be recalled, once they have been put in place in a well-ordered scheme of reference and reproduction.*
>
> H.G. Wells,
> "World Brain: The Idea of a Permanent World Encyclopaedia", 1938

At this point, it is worthwhile to go into more detail and examples about where big data can be found in the library environment. As previously pointed, there are numerous activities, tasks and

services in libraries where big data can be applied and exploited for their benefit: studies and analysis of collections, services, and patrons; deep learning, expert systems, translation, OCR, text analysis projects with a large range of tools for a wide variety of purposes; robotic assistants, to name a few. Many specific strands can be found along these general lines.

In the first instance, the use of big data in libraries can be found in the design and creation of metadata schemas and new information taxonomies. It is impossible to think of exploiting data –of any volume– in the absence of adequate metadata. Without them, datasets, especially massive ones, become an amorphous mass of sterile entities with little or no utility. In the aforementioned series of IDC Corporation studies about big data, in the 2014 study they established that metadata was added in a systematized way to only 3% of the immense amount of data being produced in the world (*The Digital Universe…* 2014). Metadata is important in any kind of information structure, but it becomes crucial in the domain of big data, as it informs everything about the data: what it is, who generated it, how, when, where and why it was generated. And they not only inform about the data itself, but also about its associated elements: transactions, forms, programs or applications, computing resources, devices, histories, and countless other potentially useful and interesting elements for an organization. In the realm of big data, metadata can and does become so comprehensive that it turns into "meta-information". Zeng & Qin (2008, 15) established four types of metadata standards already used in library practice:

- "Structures" – as the Dublin Core Metadata Element Set;
- "Content" – as Anglo-American Cataloging Rules;
- "Values", as Library of Congress Subject Headings;
- "Exchange", as MARC 21 Format for Bibliographic Data.

As mentioned before, according to the nature of the data and the sender, data can be structured, semi-structured or unstructured. Depending on this, there may be much or little metadata

embedded in each type of data, but often such metadata is not obvious or evident to everyone; a thorough understanding of its structure and essence is necessary to be able to extract something coherent from it. For example, Schmarzo (2018) established that there are a staggering twenty pieces of metadata associated with each Twitter message, beyond the content itself. Few people realize this enormous amount of metadata associated with something as simple as a 280-character tweet. It is worth noting at this point that for those who study this type of communication, what is said in the message as such has no statistical value, but these twenty metadata represent a gold mine for the analysis of this type of social network.

Librarians have long been aware of the great value of metadata in the information world and therefore they are familiar with its design, creation and use. Likewise, they have been creating and using all kinds of information taxonomies since quite a long time.[17] As is well known, the first Anglo-American Cataloguing Code with its derived tables and schemes exists since 1908 with the purpose of unifying the rules of registration and description in different libraries, and that was not the first antecedent of these efforts.[18] It has been an incessant process that in recent decades has taken on unprecedented dimensions, and has become increasingly

---

17 "A taxonomy is a controlled vocabulary that is organized in a hierarchy. Each term designates a category, type or class. There is only one type of link, which means 'is a variety of' and corresponds to a subclass relationship. Strictly speaking, each node in a taxonomy has exactly one 'parent', but the term 'taxonomy' often refers to hierarchies with multiple 'parents'. It is also sometimes used to refer to networks with more than one link type" (Uschold & Grüninger 1996, 94).

18 Frederick Rostgaard wrote theories of documentary classification since 1697. The Vatican's documentary organization systems, the ones from Jacques-Charles Brunet's, Antonio Panizzi's (1841), William Harris' (1870), Dewey's decimal (1876), and Cutter's Library of Congress (1891) are from the 19th century. The CDU of Otlet & La Fontaine dates from 1905 and the first Anglo-American Cataloguing Code with its derived tables and schemes exists since 1908.

sophisticated until it has now reached the level of complex ontologies, passing through a whole series of intermediate levels. This concept of "ontology" is neither absolute nor monolithic, and varies significantly with the approach of the discipline that defines it. In addition to these various possible approaches, Souza *et al.* (2011) have established various levels or depths of "ontological precision"; from the simplest to the most complex they defined: 1) Lexicon or vocabulary with definitions in natural language; 2) Simple taxonomy, formed by data dictionaries and hierarchies; 3) Thesaurus or taxonomy with related terms; 4) Relational model, which includes type restrictions and arbitrary relations between entities; 5) Complete axiomatic theory.

All these levels of precision and descriptions have been explored and developed by library professionals in recent times. Obviously some other authors divide the levels or depth for ontologies differently. In general terms, they classify as "light ontologies" those comprising only the vocabulary, classification or thesaurus stages, and as "complex ontologies" those already including axioms, constraints, and so on. Milton (1998, 86-88) divides them into theory-centered ontologies and those that are pragmatically oriented. For this author, the former are those created from a certain scientific, humanistic or social theory and therefore emphasize it, while the latter are those emanating from the consensual practice of a discipline and are generally designed having in mind that they can eventually be used by computer systems. The latter are therefore the most common in the practice of information sciences and are aimed at specific practical areas such as library or archival sciences.

Interesting current examples of these developments are the conceptual models underlying RDA (Resources, Description and Access), the cataloging standard for the formulation of bibliographic records used in libraries, archives, museums, etc. As is well known, RDA is a set of guidelines, data elements, and instructions to create properly formed library and cultural heritage resource metadata in accordance with international models for user-oriented linked data applications. These underlying conceptual models of

RDA are: the Functional Requirements for Bibliographic Records or FRBR, the Functional Requirements for Authority Data or FRAD, the Functional Requirements for Subject Authority Data or FRSAD, and the PRESS ontology,[19] endorsed by IFLA and consolidated with their Library Reference Model. For a good overview of all these conceptual models, see IFLA's Bibliographic Conceptual Models page, https://www.ifla.org/node/2016

The central point of interest here is that shifting theory into practice for each of these conceptual models involves the handling of large amounts of data. The fastest and most complete way to build vocabularies with definitions in natural language, simple taxonomies –data dictionaries and hierarchies–, thesauri or taxonomies with related terms, relational models with attributes, constraints, relationships, functional requirements models, etc., for each discipline, has been through the collection, process and analysis of extensive amounts of data elements. Since all the mentioned taxonomic elements must be built for each of the fields of human knowledge, the task has just begun, and most of it is yet to be done: an extremely extensive potential field of study and development.

It should be stressed at this point that all these conceptual models are not just theoretical ideas to be used in academic disquisitions: they have innumerable daily uses and practical applications to optimize the registration and retrieval of information both within libraries as well as in the environment currently imposed by the World Wide Web. In fact, it is reasonable to assume that these practical uses and applications are a fundamental part of what will keep libraries in the near future in the concert of universal information, and for this reason it is worth analyzing them in more detail.

---

19 "PRESS" is a formal ontology designed to represent bibliographic information about continuing resources, and more specifically about serial publications (journals, newspapers, etc.). It aims to propose answers to long-standing problems with the application of the FRBR family of models to these serials and continuing resources.

Big Data in catalogs

> *He is wise who knows the sources of knowledge; Who knows who has written and where it is to be found.*
>
> Archibald A. Hodges

Many people might think that big data is in catalogs, given their current large volumes. Huge as they may be today, especially in libraries such as the British Library, the Library of Congress, or the France National Library; or in major compilations like OCLC or HathiTrust, actually big data is not in their catalogs, but rather in all the information associated inside them.

Library holdings catalogs intrinsically possess an immense amount of interlinked data which constitutes a big data network not evident at first glance. Embedded within catalogs are innumerable authors –persons and organizations–, events, places, publishers, periods, subjects, dates, citations, etc., but more importantly, they conform an immense web of interrelationships that are not easily discernible or extractable, and which do not exist in each individual record: they are only found in the whole. In addition, library catalogs are usually separated by type of material: catalogs of books, journals and their tables of contents, dissertations, images, audio, vertical, etc., which makes it even more difficult to detect and establish the interrelationships among data from different catalogs, since they are usually completely separate entities, especially when the library manages catalogs for "traditional" materials as well as for digital materials. These non-obvious interrelationships are the ones actually making up a large set of big data that –being not the catalog– arise from it, and which are the subject of important eventual studies in libraries, since they offer unprecedented and powerful options for information search and retrieval. All these developments have their origin in the 2015 "Open Archives Initiative Protocol for Metadata Harvesting" (OAI-PMH, https://www.openarchives.org/OAI/openarchivesprotocol.html) originally developed in 2002 with successive updates, which set the guidelines for the

collection, analysis and interrelation of metadata internally in a library among sets of them, as well as with those coming from publishers, suppliers, and so on. Much more evolved projects have since been derived based on this initiative.

In addition to this, initiatives exist around the concept of Linked Data". These are a set of structures and principles for data capturing and recording for globally sharing interconnected machine-readable data on the web. The basic theory of this concept is that data is more valuable the more it can relate to other data in the global environment of the World Wide Web. The more things, events, people, places, etc., connected together in a structured way, the more powerful the data network will be, regardless of whether they come from diverse sources and their formats are not homogeneous. The purpose of this principle is to enhance knowledge discovery as well as the effectiveness of data analysis Such data structures have been defined under the HTTP, RDF and URI standards being one of the foundations of the semantic web proposed by Tim Berners-Lee and the W3C Consortium since 2006.[20] There is a major common theme between library metadata and the postulates of the Semantic Web: how to make the implicit relationships found in traditional library metadata –which are obvious to humans– explicit enough to be understood by machines. Given the relevance of the topic, the W3C Consortium created an interest group called the Library Linked Data Incubator Group, which performed a detailed study on the subject, the results of which were published in a final report. The document states as its goal:

> […] to help increase global interoperability of library data on the Web, by bringing together people involved in Semantic Web

20 The term "linked data" is attributed to Sir Tim Berners-Lee, considered the creator of the World Wide Web, in his note: "Linked Data Web architecture: Design Issues" (2006). Last updated: 06/18/2009. He mentions there a style of publication on the Web with interrelated structured data http://www.w3.org/DesignIssues/LinkedData.html

> activities –focusing on Linked Data– in the library community and beyond, building on existing initiatives, and identifying collaboration tracks for the future (W3C 2011).

The document reviews some ongoing projects in this regard and presents interesting recommendations on the subject. Heery (2004, 270) mentioned there ware main similarities between traditional library metadata and linked data:

> […] what is perhaps the most striking aspect of the Semantic Web for the library community is the commonality between traditional information management and library interests –constructing vocabularies, describing properties of resources, identifying resources, exchanging and aggregating metadata– and the concerns that are driving the development of Semantic Web technologies (Ávila 2020-1, 80)

mentions in this regard:

> [...] the integration of Linked Data in the bibliographic record has two essential purposes. On the one hand, to link library data with other data sources available on the web. On the other hand, to promote the generation of a method for the optimal retrieval of information in libraries, according to the current users' demands.

Given the importance of the topic, IFLA (2017) has already also summarized the use of linked data in libraries.

As representative examples of metadata with linked data, the British Library and the Library of Congress have already begun to study these data interlinks among their hundreds of respective collections –involving many millions of items– trying to model the interrelationships among people, events, places, etc., contained in their holdings. See as examples the *British Library Data Model – Books*, recovered from: https://www.bl.uk/bibliographic/pdfs/bldatamodelbook.pdf, and the *Library of Congress Linked Data Service*, Recovered from: https://id.loc.gov/, Basically, these institutions have extracted immense amounts of data from their

catalogs by assembling interrelationship matrices among them, creating numerous interlinks mappings to form unprecedented sets of metadata. Springer Nature, a division of this publisher, is building a project called SN SciGraph around the concept of "linked open data" (Recovered from: https://www.springernature.com/gp/researchers/scigraph). This project consists of a natural sciences "discoverer" compiling data from this publisher's editions in this field along with those coming from other academic partner institutions. The database collects information on research papers, books and chapters, conferences, citations, institutions, researchers, etc., making semantic links amongst them. They claim will eventually have some two billion linked items. OCLC has also been doing some developments about linked data from its catalogs; it has a project entitled WorldCat Linked Data, of which it has already published some by-products: VIAF – Virtual International Authority File, a linked authority catalog structure, and FAST – Faceted Application of Subject Terminology, derived from the Library of Congress Subject Headings (Recovered from: https://www.oclc.org/en/worldcat/oclc-and-linked-data.html).

Although all these projects are still in their early stages due to the vast number of data to be extracted and correlated, the partial results obtained are highly interesting and encouraging: indeed, the general concept of this type of initiatives can establish unprecedented dimensions which are extremely useful and powerful in the field of libraries, and certainly fall into the environment of big data management, as they indeed contain the established basic characteristics for them: volume, velocity, and variety. The *volume* of data eventually contained in the interrelationships of all these elements –authors, events, places, subjects, periods, publishers, dates, citations, etc.– can easily climb to several million, especially if more than one library's catalogs are involved simultaneously: books, journals, dissertations, multimedia, etc. The *velocity* at which data and therefore their interrelationships change is very rapid, since new elements are added daily to all those interconnected documents belonging to a certain library and consequently the relationships mapping changes. Finally, the *variety*

is also extremely wide, since the structural forms of these data even though they are standardized in libraries– are not the same throughout different catalogs, and may differ even more with information coming from external sources: tables of contents, citation systems, authority catalogs, external ontologies, etc.

The assertion that this type of projects can establish unprecedented dimensions of great practical utility and power in the field of libraries has been increasingly consolidated in recent years. Current studies conducted in countless libraries around the world indicate that one of the main factors of users satisfaction or frustration in a large number of libraries is precisely their internal information search engine for the catalogs, due to three main reasons: first, many libraries accept the first search engine offered to them as part of their ILS (Integrated Library System), without checking whether it had the minimum required elements: Boolean operators, exact terms, truncated or "wildcard" words, delimitation by date, language or other parameters, and so on. As a general rule these library search engines perform the most basic possible retrieval through all records containing any of the searching words in any position, presenting everything that resembles what the user entered without discernment, thus retrieving huge sets of information of little relevance and therefore of little use or interest. Secondly, users' search habits have changed substantially due to the web, and in general the usual library search engines can only search on the "typical" fields of catalog records: author, title, subject, publisher, series, etc.; a set already considered by users as too limited and complicated in order to find what they are looking for. Thirdly, as a general rule, library search engines operate vertically in the catalogs, one by one, so users must repeat their search in each one of them to cover all exhaustively.

Aware of this problem as well as the new uses of data, many libraries are already aggregating large additional datasets to their catalogs, thereby radically optimizing their search engines. For example, they are capturing and adding to them the table of contents and glossaries for each of their books, linking these terms with the original catalog record. This enhances the search immensely,

since the search engine no longer has only a few words from the author, title or subject, but many words contained in the index or glossary of each book. Some specialized libraries –i.e., in literature– do not add tables of contents, since these hardly exist and contribute little to this type of publications, but they extract and add all the characters, places, periods, events, etc., contained in novels, tales, theatrical plays, and other similar works. Something similar can be done in almost all disciplines, adapting the idea to the context and characteristics of each one of them; for example in chemistry, extracting from the book index and adding to the original record formulas, processes, substances, etc., useful for the search engine; and this could be adapted for many other disciplines. If the library also provides that searches can be made horizontally across several of its catalogs, and verifies that its search engine has the pertinent search and narrowing elements: Boolean operators, exact terms, truncated or wildcard words, delimitation by dates, languages or other parameters, changes in card display formats, etc., the library will then have a formidable search tool that will compete very advantageously with searches on the web or in other sources. Regarding this last point about the desired options in search engines useful for patrons, Markey (2007), Borgman (2007), Calhoun (2006), and De Rosa (2006) among other authors, have elaborated extensive and varied lists to this respect. New Zealand National Library (2017) prepared an interesting checklist for librarians to verify all the attributes and capabilities of a certain ILS they intended to acquire. The list is very comprehensive and well developed according to those elements considered desirable today.

Some libraries are already adding to their search engine various interrelated elements such as those mentioned above: authors, events, places, citations, etc., built horizontally across several catalogs. This is the real immense value of interrelationships in library catalogs: interlinked data models instantly allow the search engine to inform the user that a certain author searched for –say in the book catalog– appears too in other catalogs, or perhaps that such author is cited in other texts, be they books, articles, dissertations,

etc., or if that author generally publishes with other certain authors. The same can be done with topics or other parameters. The search engine can report that people who consulted a certain book also consulted other related books, pointing to them; the possibilities are endless. The central point of all this is that certain additional data added to the catalogs can greatly enhance them. Obviously such data extraction, their interrelationships mapping and aggregations cannot be done by basic manual or computational methods: they require special treatment that is certainly close to big data because of their volume, their velocity of change and their variety of structures.

Such applied projects are only a small sample of the possibilities; Schilling (2012) provided a good overview on how to transform metadata into linked data within libraries. Ávila (2020-2) developed an excellent compendium about linked data models and platforms in libraries. The projects aforementioned in this regard: the "British Library Data Model – Books", the "Library of Congress Linked Data Service", and the expanded catalogs are just a few examples of what can be achieved in practice with the use of big data and metadata for library catalogs optimization. It should be emphasized that these types of developments are undoubtedly fully applied and tangible examples of the theoretical concept of semantic libraries and linked data.

Big Data in metric studies of documentary information

> *We are no longer in the information age. We are in the age of information management.*
>
> Chris Hardwick

Big data are also found in libraries in the metric studies of documentary information, in all its specialties: bibliometrics, informetrics, librametry, as well as in other associated ones: scientometrics, webmetrics, altmetrics, and emerging archival metrics. All of them

have in common the application of mathematical and statistical models and methods to library, bibliographic, and archival tasks; social networks, research in sciences and humanities, communication and dissemination, among many others. These are another classic examples of applied data mining.

Factors ruling the impact of articles in academic journals obviously have had and continue to have strong repercussions on the marketing, demand and prices of these journals, as well as on the prestige, incentives, and recognition of their authors and their institutions. For this reason, these have been elements of great interest through decades for publishers, research centers, libraries, etc., and within these organizations for authors, editors, librarians and those responsible for research policies in all disciplines. These factors remained stable for decades, but in recent years they have been questioned, and new elements have come into play initiating a radical change in such conceptualization: open access journals and open research repositories, social networks among researchers, breakdown of citation structures into more finely differentiated categories, just to mention a few. In fact, webmetrics and altmetrics as specialties of document metrics have emerged in recent times due to the web. In addition to all this, new data analysis techniques have introduced fresh approaches and perspectives to the study of documentary information. Gorbea mentioned since that year in this regard:

> [...] bibliometric indicators have often been used to support the level of scientific development achieved by a particular discipline, institution or country, a practice that has led to proposals for scientific and information policies based on the recognition of high levels of productivity, impact, visibility and growth of scientific literature generated in these instances. This has meant that those disciplines, information sources, institutions and countries that are in the so-called 'mainstream' always appear better represented by this type of indicators... the above behavior has raised controversy about the validity and usefulness of this type of indicators in the evaluation of science, which is why the study, recognition

> and *definition of bibliometric indicators in the specialized literature is extensive* (2013, 154).

Today's great capabilities and tools for data collection, management and use have been expanding the possibilities of this type of models and techniques to create new fields of application and new types of studies in this regard. Moed (2012) also mentions:

> [...] the availability of full-text research articles in electronic format gives us the opportunity to conduct textual analyses of all of an article's content, not just the metadata, extracted by indexing databases. The citation contexts can be analyzed linguistically, and sentiment analyses can be conducted to reveal how the citing author appreciates a cited work.

Several authors, e.g. Small & Klavans (2011), and He & Chen (2018) established that by using text data mining techniques, visual analysis, and so on, it was possible to obtain interpretable results from the different contexts of citations in scientific journals, provided that the data was sufficiently structured. This allowed differentiating citations in those journals into several weights and categories to develop diverse typologies of those citations. He & Chen also mention that citation contexts were also used in order to characterize publications for various applied uses, such as publication summarization (Qazvinian *et al.* 2010), survey article generation (Mohammad *et al.* (2009), and information retrieval (Huang *et al.* (2015). Lourdes Feria (2020) described an interesting applied example of bibliomining for user diagnostics in Mexico.

Other authors, such as Duy & Vaughan (2006), (2005) and (2003) have established that –having considerable volumes of data– the analysis of access and use of scientific journals is a more accurate alternative for measuring their impact instead of the traditional citation analysis; they also conducted studies to verify to what extent both the impact and the use indexes of certain scientific journals coincided with those provided by the publishers. These studies and their results are also very useful in libraries to

feed back into the eventual selection of journal subscriptions, and are very helpful when making decisions about renewing or cancelling them. Moreover, impact analyses have already begun to be carried out not only for journals, but also for books; see Halevi, Nicolas & Bar-Ilan (2016).

Halevi (2014) compiled a very interesting summary of the different types of uses of big data in bibliometrics, categorizing them into five: citations, references, keywords, usage, and full-text analysis. She adds:

> [...] data availability and technological capabilities led to a strong proliferation of bibliometric databases and better data analysis tools for the development of: more sophisticated and personalized scientific evaluation indicators; measurements of the behavior of researchers and journal editors; indicators of the social impact of research, both in its academic value and in its contribution to the enlightenment of the general public; creation and analysis of macro-datasets by combining multiple data sets.

Other authors have studied internal patterns, geographical distribution, influential journals and institutions, international collaboration between authors, institutions and countries or territories which use big data publications. Coincidence and correlations of the keywords of authors publications' have been also studied. There are bibliometric reviews on various disciplines performing with big data what they did before with "small data", which have detected trends and current topics, their dispersions, research groups, etc. There is indeed a difference when using much larger amounts of data and new analysis tools, for different results can be obtained rather than when recurring to "traditional" tools, the former being finer and allow more detailed approximations.

It should be emphasized at this point that text analysis or text mining are not limited only to metrics studies. With the use of data mining, linguistic techniques, statistics, machine learning, information retrieval, natural language understanding, case-based reasoning, and others, this type of studies can help people and

organizations to obtain new knowledge by extracting meaningful information from large amounts of unstructured documentary texts available on the Internet and corporate intranets, using elements as varied as lexicographic and semantic analysis, groupings, categorizations and taxonomies; links, relationships and associations between entities; sentiment analysis or opinion mining, word frequency, and so on. Consequently, their applications are very diverse: texts identification, extraction of elements from them, texts categorization and/or taxonomy, extraction of their concepts, entities, relations, events; texts translations; text trends, to name only a few.

As can be seen, the use of big data has opened new possibilities and fields of interest for this type of metrics and its professionals within libraries and other related organizations.

Big Data in machine learning in libraries

> *Ambiguity is not today a lack of data, but a deluge of data.*
>
> Paul Gibbons 2015

The next field of data analytics in libraries consists of one of the subfields of AI (Artificial Intelligence), the so-called machine learning,[21] in which a certain specific system of this type of AI is designed and programmed to be susceptible to be taught, trained, or prepared to perform various optional actions without direct human intervention; these specific systems receive data that they can interpret, and extract from it meaningful patterns; depending on such data and its interpretations, a certain system will respond in one way or another (*Artificial Intelligence...* 2018, 6). Any computer program can make simple action decisions according to its "conditional statements": *if... then... else*. The main difference with

21 The term "machine learning" is attributed to Arthur Samuel in 1959, a pioneer in computer games and Artificial Intelligence. He defined it as "the ability of computers to learn without the need for explicit programming".

machine learning systems is that the latter also use complex decisions of pondering results as they process more and more data, assigning "weights" and iterating over and over again the possible decisions by feeding back new data and rearranging the importance and order of the results after each iteration. Machine learning is similar to data mining in the sense that both are processes for exploring large sets of data to discover patterns and correlations; the main difference between them is that machine learning goes as far as pattern prediction and does not merely remain at pattern discovery. Experience with this type of systems has shown that the more specific and precise the task performed by a system, the more accurately it can perform it; systems pretending to have very broad fields of action tend to lose precision. Although the term "machine learning" suggests that the machine is teaching itself, it is necessary to emphasize that in practice the most common machine learning techniques are supervised by people, and require enormous amounts of aggregated human work and knowledge, as well as the careful design and feedback of training data.

Machine learning is used today not only in libraries, but in the entire LIS (Library and Information Services) industry for many different purposes: indexing, cataloging, classification, online information retrieval, abstracting, reference services, tables of contents, user and trend analysis, and so on.

Many of them have already started to build practical applications of machine learning in several ways: first of all, there is document analysis and synthesis. This consists of programs able to "read" a certain document and extract information from it. As already mentioned, systems of this type are built for very specific fields of action, meaning this: documents; there is not yet a universal "interpreter system" for all types of them. The programs for this purpose are built to specifically interpret certain types of documents: there are programs for text, for images, for video, etc., and within these there are even more specializations: there exist now programs built to read books, others for journals, others for dissertations, for e-mails, and so on. There are programs for interpreting photographs; others for maps, for famous paintings, for

sheet music, and others. Their function is to extract certain types of specific information from these documents; this can consist of elements as varied as a summary of the text, or about what types of people or objects appear in the image, or certain metadata such as people, events, places, publisher, etc. Obviously, this extracted information can be used later for certain projects or uses in the library. It is worth stressing at this point that these systems can become powerful auxiliaries for the library, but they still cannot fully replace human activity in projects. To date there is no system that can, systematically and reliably, read books and build their complete catalogue records from them, but it can extract enough coherent information to provide valuable elements for people, such as cataloguers, or for systems, such as library discoverers. This use is one of the points that show the potential application of these systems which deserves further reflection in libraries. In fact, one of the most radical questions in libraries today is whether catalogs should continue to be built in the "traditional" style, or whether there should be a change towards new structures for document registration and retrieval. An interesting reflection about this can be found in Bourg (2017).

Another well-known practical example of "machine learning" programs in libraries are the so-called "OCR" – Optical Character Recognition, which interpret text that has been scanned in image form to convert it into computer interpretable text formats, such as txt, doc, rtf, odt, pdf, etc. These types of programs belong to the field of AI, since their function is to read and interpret letters from a graphic form, just as humans do, and they fall into the subfield of "machine learning" since these programs can "learn" issues with human feedback, such as: interpretation errors; old, discontinued, and serif typefaces; blemishes, and their eventual corrections.

Libraries have been using these devices for much longer than is thought. In the early 1930s –long before computers– character recognition associated with microfilm was invented for rapid searching on them. By the late 1940s, such systems were widely used in libraries and archives. For instance, Alicia Perales (1962, 21-22) reports the Microfilm Rapid Selector, which stored consecutively

cards with information on a microfilm tape. Attached to each card was simultaneously stored a pattern of white marks encoded on the film similar to perforations on cards, which was then optically searchable by a machine using photoelectric cells to find the combination of encoded marks containing the desired information on the cards, at a rate of up to 36,000 cards per hour. A similar device also widely used in libraries during the 1950s and later was the "Kodak Minicard"; the difference was that it stored each record individually. In the mid-1970s the first OCR programs for interpreting text with computers were introduced commercially. At first they were very crude and inefficient; however, libraries began to acquire and use them for text conversion because of their usefulness. Over the years, these programs have achieved very advanced levels of efficiency, as their learning and correction features have evolved enormously to make them very accurate, and therefore virtually all digitization projects in libraries and archives include the use of an OCR system.[22] Moreover, since other variants of these systems can also interpret and convert written text to speech, as well as written text to the Braille alphabet, they are widely used in libraries with typhlological departments to serve people who require them. Few librarians have reflected on the fact that the OCR and typhlological systems they may have in their libraries are longstanding practical applications of AI, and especially machine learning.

22 Today, the most commonly used OCR systems for library and archival projects are: OneNote, Google Drive, SimpleOCR, FreeOCR, PhotoScan, OmniPage, Abby FineReader, and Capture2Text. Obviously there are more.

Big Data and expert systems in libraries

> *Although expert systems may create new functions for librarians and free them for other high-level tasks, such systems will somehow invade their professional domains. Therefore, librarians should familiarize themselves with current research and applications of expert systems that may affect libraries.*
>
> S. E. B.,
> "THE CUTTING Edge," American
> Libraries, December 1983, 730.

Another subdivision of AI widely used in libraries is the so-called expert systems. They have been of interest to librarians starting in the 1980s, and since then numerous texts can be found about the subject, dealing with knowledge-based indexing, natural language processing, cataloging, query information retrieval, and so on. Expert systems are computer programs using Artificial Intelligence principles and methods to solve problems within a specialized field that would usually require the skills of expert personnel. They incorporate the accumulated know-how of experts in a certain subject and are designed to function as closely as possible to them. Basically, they are constructed upon a knowledge base of facts and relationships represented in the form of data, and have the ability to make inferences based on them. The creators of these systems use various techniques for the acquisition of this knowledge base, such as the analysis of written protocols and procedures, the verbal description of tasks performed by persons, questionnaires, surveys and interviews; the discovery and documentation of tacit knowledge within the organization, as well as the observation of processes and their simulation.[23] This is a very

23 The first expert system was developed in 1965 at Stanford University by Edward Feigenbaum and Joshua Lederberg. It was called "Dendral" and was built for the analysis of chemical compounds.

valuable resource in libraries, since much of the librarians' knowledge about the management and exploitation of information falls under this heading: the tacit or internal knowledge of librarians is their accumulated knowledge, one generated by their experience, inherent to library staff and which has been interiorized through different processes. Olivares (2020) made a very complete review on this topic.

Machine learning and expert systems can be used in many other practical applications in libraries:

- Some libraries massively extract and store data about their patrons' searches to learn more about the logic and ways they approach to information, and thus improve internal catalogs, information finders and discoverers, etc.;
- Many libraries specifically study the "natural language" that patrons –like all people– use to communicate, in an attempt to teach computers to understand such language. Under these principles, the machine can identify the language's key concepts within a question and its possible solution through AI processing of natural language. The goal of these applications is to design and create programs capable to analyze the plain language used by a person and apply it to textual information extraction, information retrieval in databases and catalogs, machine translation, speech recognition and synthesis, and so on. All these applications require large amounts of data;
- Similarly, many libraries store information from previous patrons' searches in order to "personalize" each user's page, "memorizing" what they have previously searched for, in order to establish patterns, just as e-commerce sites do. By saving this type of information for each user, the system can later make suggestions such as "people who consulted this text also consulted these others..." or "this author is related to that one" or "such topic is related to that other one".

- "Personalized" library pages are generally constructed in such a way that they also allow users to save the appearance of their pages, the display formats, their previous searches, etc., so that the look and behavior of each user's page can be different, to their preference and convenience;
- Some libraries extract data from the social networks of their patrons linked to the library services to retrieve suggestions for the acquisition of items, to detect "trending topics", to count "likes" and other similar events about their services or information, to verify effectiveness and follow up on their services, to measure "usability" of new services and options, to detect failures or problems, and to design new tutorials, among many other uses.

An interesting compendium of such applications for libraries can be found in *Artificial Intelligence and Machine Learning in Libraries* (2018).

Norman Jacknis (2017) summarizes the interplay between libraries and Artificial Intelligence splendidly:

> [...] the issue is not 'either ai or libraries' but both reinforcing each other in the interest of providing the best service to patrons. Instead of being purely ai, artificial intelligence, this new service would [include] what is beginning to be a new buzzword – IA, Intelligence Augmentation for human beings.

Libraries as vast data repositories

It is easy to lie with statistics. It is hard to tell the truth without it.
ANDREJS DUNKELS

During the last decade, academic and information communities realized that the data collected throughout scientific, journalistic, social, etc., researches had an added value after the conclusion of their projects as it could be reused later by other people, since there is no doubt that a certain set of data from a research is susceptible to be analyzed from new approaches and perspectives by different groups, and eventually new results can be obtained from such data. Based on this consideration, it is no longer merely the raw material to produce information, but an object of information in itself with its own inherent value, and therefore requiring specific treatment.

In addition, the worldwide trend in the dissemination of research results underwent a change: researchers were no longer obliged to publish only in "renowned" journals. Increasingly, governmental academic funding agencies began to require that both the results and data from government-funded research be made

public.[24] But disclosing data sets to the public requires method and standardization. Until that time, virtually every researcher or group designed their own formats, forms, repositories, etc., for their data for each research project; any method was good if it worked for the project. Suddenly, research communities were faced with the need to start managing their data in a systematized and standardized way for later storage and access. Researchers found themselves without sufficient time, skills and resources to handle their data in such a way throughout their projects, with the additional problem of finding appropriate repositories for their data. Research institutions –especially those in universities– were urged to start creating repositories of their project data: this was a new boost to science and data management. Many institutions turned to their libraries for advice and started hosting these datasets. Hence the arrival of data repositories in libraries. From that point on, in the words of Rafael Ball (2019): "[...] the work of the library no longer focuses only on books, journals, and catalogs, but also on all types of data –structured and unstructured– as well as their forms: texts, metadata, images, audio and video collections, research data, and software".

Several library organizations started to outline these new challenges, such as the Association of College and Research Libraries, a subdivision of the ALA (Tenopir *et al.* 2012, 2015), and LIBER, the "Ligue des Bibliothèques Europeénnes de Recherche" (Tenopir *et al.* 2016). IFLA also conducted detailed studies of the topics related to data in libraries; in the last issue of 2016 and first issue of 2017 of its journal, this organization compiled about twenty texts and reflections on the subject, dividing it into four main headings: the researchers' needs, the skills required from librarians, the possible services to offer, and data literacy (Recovered from: https://www.

24 There is already the series of well-known initiatives called "Open Data", which, similar to open access journals, promotes the creation and dissemination of open data repositories. In addition, they must be governed by the so-called "FAIR" principles for data: Findable, Accessible, Interoperable and Reusable.

ifla.org/publications/node/1691). Based on these preliminary studies, IFLA created an initiative called the Data Curator Project (Recovered from: https://www.ifla.org/library-theory-and-research/projects). Its main objective was to determine the roles and responsibilities of library professionals already engaged in these tasks in various countries. The study also focused on the terminology used to describe emerging practices and new professional roles.

Witt & Horstmann (2016, 251) made a very representative and concise list of the main activities required to librarians in this regard: 1) helping researchers to understand and resolve needs throughout the research data lifecycle; 2) advising on the construction of data and metadata management plans; 3) designing data publishing and curation solutions; 4) creating web guides and tutorials to train researchers and users; 5) hosting and maintaining repositories in their current holdings.

All these new needs, concepts and solutions gave rise to a new specialty in the world of information, called Research Data Management or RDM. Whyte & Tedds (2011) defined it as: "[...] the organization of data, from its entry into the research cycle to the dissemination and archiving of valuable results." Basically, RDM deals with all aspects related to the management, storage and distribution of research data: data lifecycle, collections, capture, cleaning, consistency, standardization, and formats; metadata, repositories and data querying services; data anonymization and security; data preservation, the skills and roles required for data operators, data literacy for researchers, and even data citation. Pinfield, Cox, & Smith (2014) noted that there are seven major areas of development or "drivers" for the study of RDM: storage, security, preservation, policy and legal compliance, quality, dissemination, and engagement. Obviously data management, and specifically Research Data Management, is a multidisciplinary activity, but it is clear that librarians must be among the professionals who manage it. Certainly it requires new knowledge, training and instruction, but librarians clearly do have the proper professional background for this task.

Interest in the topic of data and libraries has been growing strongly in recent years. The Association of Research Libraries (ARL), and the National Science Foundation (NSF), created a few years ago a special union to jointly develop projects in what is now known as e-Science; among them, studies and developments in RDM. Some authors have already addressed the topic of libraries in RDM, such as Cox *et al.* (2017), Alvaro *et al.* (2011), Matusiak (2019), and Lewis (2010). In Mexico, the (IIBI) Instituto de Investigaciones Bibliotecológicas y de la Información" Institute of Library and Information Science Research of the National Autonomous University of Mexico (UNAM), organized in November 2018 a forum exclusively dedicated to data management and its relationship with libraries: the "Second Congress on Information Studies: Data Management" = "Segundo Congreso de Estudios de la Información: Manejo de datos" (*El manejo de datos.…*, 2020).

Many universities and research institutes are already creating data repositories for this purpose, and many libraries and library systems are already working in this direction. A very representative example of this is the data repository of the Network of the National Library of Medicine (NNLM). This repository was created by this library so that researchers from associated institutions willing to do so, can store the results of their projects in there, obviously in the health sciences area (Recovered from: https://nnlm.gov/data/thesaurus/data-repository). Another interesting example are the guidelines created by the Smithsonian Institution libraries for the creation and deposit of data in repositories, covering a wide variety of specifications due to the vast diversity of interests and disciplines encompassed by this organization (Recovered from: https://library.si.edu/research/data-repositories). In the field of social sciences, representative example is the Inter-University Consortium for Political and Social Sciences Research or ICPSR of the Massachusetts Institute of Technology libraries, considered the largest repository worldwide in this field of knowledge. In addition to the data repositories of academic institutions –being already very numerous– some other repositories for data hosting in general are now proliferating, such as Zenodo, Dryad or Dataverse.

In Latin America (ECLAC), the UNO Economic Commission for Latin America and the Caribbean is associated with a project called Leaders Activating Research Networks (LEARN), as part of the European Union's "Horizon" research and innovation program. The purpose of this project is to promote and develop Research Data Management projects. As a result of this initiative, the Brazilian *Scielo* system of academic journals from this region has recently installed a pilot version of a data repository precisely for researchers and data from this geographical area; the repository is built on Dataverse (Recovered from: https://data.scielo.org/). Some countries in the region have begun to legislate in this regard and/or to build data repositories in universities and related academic institutions: Argentina, Brazil, Chile, Colombia, Mexico and Peru (Andaur 2016). There are also already some registries or catalogs worldwide providing information about scientific data repositories, such as *"re3data"*, collecting information about more than two thousand repositories of this type (Recovered from: https://www.re3data.org/).

As can be seen, today's Research Data Management is a new field of action offering broad new opportunities for information professionals; it obviously requires new skills and specific knowledge; among those who are most likely to be trained in this new approach to information are undoubtedly librarians, due to their proper professional background and experience.

But data management is not limited to research data; as already mentioned, in recent years a number of initiatives have been developed under the umbrella of Open Data, aimed to promote creation, dissemination and use of open data repositories. These data initiatives are the sequel of other previous movements in favor of openness: free software, Open Government, open academic journals, and so on. This is because –beyond academic research data– open data increasingly occupies a preponderant place in the modern world: it allows a more complete understanding of global problems and universal issues, such as diseases, education, insecurity, employment, or famine. It is a fundamental factor in the principles of Open Government, with transparency and accountability; they

empower citizens and thus strengthen democracy. It can streamline the processes and social structures that governments and societies are building. It can support in an outstanding way movements for racial equality, gender equality, and so on. In short, it can help to transform the way we understand and interact with the modern world (Recovered from: https://opendatacharter.net/principles/).

There are already some representative projects in operation; for example, the World Bank Open Data, whose repository contains more than 3,000 global datasets on development, economics, etc., in open form. There is also the "World Health Organization Open Data Repository", which includes statistical information on this subject from its almost 200 members. Then there is the "European Union Open Data Portal", with 12,000 datasets from governments, agencies, institutions, etc., from that geographical area. There is also the Wikipedia's DBPedia project, which allows semantic searching and exploration of the relationships and properties of 4.6 million elements of that encyclopedia, such as people, places, events, etc. Another project is RODA – Registry of Open Data on AWS Resources from Amazon, which allows searching in a single site for open data captured on that platform. Similar to the above is "Google Public Data Explorer", which enables searching multiple open data banks worldwide. To these should be added numerous agency-specific open data sets, such as the "National Oceanographic and Atmospheric Administration" or NOAA, and the "National Center for Atmospheric Research" or NCAR, which collect and distribute climate, weather, etc., open data from all over North America; and similarly seismological, volcanological, census, etc., services at the regional level or from numerous countries.

But there is more: initiatives around the concept of Linked Data in libraries have already been mentioned. When such data is also installed as open, i.e., it can be freely used and distributed, it is called Linked Open Data or LOD. Theoretically, this is shaped as a virtual cloud of data in which anyone can access any authorized data as well as add new ones, which provides an open, structured and interoperable environment that favors that data can be created, interconnected and consumed on a global scale. It should be

noted that these are two related but distinct concepts: open data is made available to all without the need to be interlinked with other data; data can be linked without having to be freely available for use and distribution. In short, data can be open but not linked, and data can be linked but not open. When data is both linked and open, it becomes linked open data. As with other types of data projects, there are already proposals around linked open data specifically in libraries – see the compendium of papers on the subject resulting from the IFLA Satellite Meeting in France in 2014 http://ifla2014-satdata.bnf.fr/.

All of these are examples of large open data projects already in operation; new ones are being added to the list daily. The main point is that the design, management and exploitation of this type of data covers a field infinitely larger than just research data management. Even if these projects are not directly managed or embedded in a library, they inevitably require staff with experience and expertise in data management. Many of these developments already involve library staff, but there could certainly be more. Library staff undoubtedly have great opportunities for professional development in these open data projects, beyond academic research.

All of the above uses of big data in libraries are not hypothetical; they already exist and are used in some libraries around the world. Every day new theories and discoveries become technological applications in this field. The context of digital libraries is relentlessly changing. Pierre Piganiol, as early as 1971, splendidly summarized this:

> [...] information should not build up a dead structure: the body of knowledge is in continuous evolution and it is vital, in order to forecast and influence the future, that information should contain at least the seeds of tomorrow's progress and discoveries. What distinguishes modern information science from traditional documentation is precisely the introduction of this heuristic element (Piganiol 1971, 13).

The downside of Big Data in libraries

For if we are observed in all matters, we are constantly under threat of correction, judgment, criticism, even plagiarism of our own uniqueness. We become children, fettered under watchful eyes, constantly fearful that –either now or in the uncertain future– patterns we leave behind will be brought back to implicate us, by whatever authority has now become focused upon our once-private and innocent acts. We lose our individuality, because everything we do is observable and recordable.

BRUCE SCHNEIER
"The Eternal value of privacy", 2006.

Like any other technological development –along with its many uses and benefits– big data also has many significant risks, problems and disadvantages, which librarians need to be aware of and study, in order to avoid or at least reduce them.

Big Data is difficult to manage, in part because of its inherent immense volume and in part because there is a general lack of knowledge on how to handle it properly, existing very few data experts worldwide. This scarcity of knowledge and qualified personnel frequently leads to a poor approach to objectives and techniques, data duplication, data inconsistency or bias, improper selection of analysis tools, erroneous interpretations, etc., with the subsequent negative consequences. It is very easy to get lost in a sea of data. Moreover, big data projects require a certain budget, and pose great technical challenges. Obviously all this is

a drawback; but undoubtedly the most negative aspect of the use of big data is its potential abuse of privacy and confidentiality of personal data.

All of the aforementioned applications of big data –which are neither good nor bad in themselves– can be and are already being used for the great benefit of individuals; unfortunately, they entail a huge problem: since they often involve the collection of personal data, they also introduce great threats to their privacy. Compiling and storing large amounts of data –when personal data are included among them– introduces the risk of them being used for purposes other than statistics and mass processing, and that they can be used for unethical or even illegal purposes. Even if the data compiler does not misuse the data, there is always a risk of information leaks from the data servers, deliberate or by mistake, or attempts to extract information from them by third parties with mischievous intentions. This is not a simple data protection issue: it involves from ethical principles and decisions to management of large commercial interests, data protection legislation and standards, technical and administrative responsibilities, data governance, accountability and IT security. There are different components in all of it: data protection is essentially a technical issue, mostly involving securing data against unauthorized access: who takes care for it and how. Data privacy goes even further: it is an ethical and legal issue, implying even deeper aspects: who can own personal data, for how long, who defines those who can access it, who can access it with permission and under which circumstances, who can modify it, to whom and how it can be transferred, and so on. Therefore, data protection is a necessary but not sufficient a precondition to achieve a greater goal: data privacy.

The European Union's General Data Protection Regulation or GDPR establishes personal data as:

> [...] any information relating to an identified or identifiable natural person; [that is], a person who can be identified, directly or indirectly, and in particular, by reference to an identifier such as a name, an identification number, location data, an online identifier

> or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person (Recovered from: https://gdpr.eu/tag/gdpr/).

Mexican law defines personal data as: any information concerning an identified or identifiable natural person. A person is considered identifiable when his or her identity can be determined directly or indirectly through any information (Mexico. Ley General… 2017). GDPR –in force since 2018– has as predecessor the e-Privacy Directive of 2002, also from the European Union, and is currently considered the most advanced regulation worldwide for the benefit of users' privacy (Recovered from: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32002L0058).

Undoubtedly, data is nowadays a huge business worldwide, and among this, personal data is even more so. Large global companies such as Google, Facebook, Twitter, etc., have multi-million dollar profits coming in large part from data sales; this fact in itself is not against ethics or the law. There are countless legitimate and ethical applications of the use and/or sale of data which create business, commerce, advertising, government applications, jobs, etc., all according to legally and morally acceptable principles. The problem is that the line between legal and illegal, ethical and unethical, is actually blurred and its boundaries are not clear. It is pointless to discuss here all the eventual bad uses for data, but a clear example of it was the scandal of the company Cambridge Analytica, which used personal data extracted from Facebook to create bias in the US elections in 2016; Facebook faced multi-million dollar losses as a consequence. There are also the lawsuits, sanctions and fines that Google, Apple and Amazon have received and continue to receive in Europe for dubious and abusive handling of the data they collect. There are thousands of examples like these; we all have received countless phone calls, e-mails and tags offering us goods and services which we never requested, and even African inheritances. The central point is that there are many entities interested in appropriating personal data and there are many others willing to hand them over, not always within the legal and

ethical framework. The temptation in this regard has been and continues to be enormous.

Derived from this, currently over a hundred countries have some level of legislation and regulations about privacy and data protection (CNIL 2019). Some countries have raised regulations to very high levels for the benefit of their citizens, as is the case of those belonging to the European Union, but there are others that, in practice, although they have issued laws for personal data protection are known for exercising strong surveillance and government censorship of their citizens, such as Russia or China. There are countries that do not exercise censorship over their citizens, but the extent to which they exercise surveillance over their citizens is highly debated; this is not a history of the past. In recognition of the above, the United Nations General Assembly adopted in 2013 and 2014 several resolutions on the Right to Privacy in the Digital Age, urging all countries to "respect and protect the right to privacy, including in the context of digital communications...." and "to take steps to stop violations of these rights and create the necessary conditions to prevent them, such as ensuring that relevant national legislation complies with their obligations under international human rights law" (UNO 2013).

Austin (2016) established that globally there were ten major themes of attention and debate about privacy and personal data protection: 1) The physical location and jurisdiction regulations of data servers; 2) The Internet of Things and ubiquitous computing; 3) Official privacy regulations; 4) Regulations relaxing compliance obligations on privacy and data protection issues; 5) Government surveillance; 6) The development of new cybersecurity standards; 7) Big data; 8) The new global framework about data transfers; 9) Recent laws and regulations about data security; and 10) The new European Union regulations on personal data protection.

Regarding our field of interest, libraries and archives, activities of data compilation and use –especially big data– also frequently involve the collection of personal data. This entails with it responsibilities that librarians must be aware of, as well as skills that must be acquired for its proper handling. This is inevitable and

indeed is not a result from big data: it has been around for a long time and has simply become more emphasized with technology. Libraries have long defended their users' right to privacy: the ALA – American Library Association, adopted as far back as 1939 a set of principles known as the "Library Bill of Rights." Among several rights, it was defined in there that whatever any person decides to read does not concern anyone else, and there is no valid reason for governments, organizations or individuals to interfere or find out about it. In its current version, it establishes in its clause VII:

> All persons, regardless of origin, age, background or views, possess a right to privacy and confidentiality in their library use. Libraries should advocate for, educate about, and protect people's privacy, safeguarding all library use data, including personally identifiable information (ALA 1939).

In its "Interpretation of the Library Bill of Rights" about privacy, the ALA clearly states:

> [...] privacy is essential to the exercise of free speech, free thought, and free association… lack of privacy and confidentiality diminishes users' choices, thus suppressing access to ideas… the possibility of surveillance, either directly or through access to records of speech, research and search, undermines a democratic society (ALA 2002).

IFLA has also pronounced itself on this issue for many years. Its principles are based on Articles 12 and 19 of the Universal Declaration of Human Rights, among which, freedom of access to information and freedom of expression have been fundamental principles for the profession in which –most especially– privacy is considered an indispensable part of safeguarding these rights. Article 12 of the Universal Declaration includes privacy as a human right, and states*:* [...] no one shall be subjected to arbitrary interference with his privacy, family, home or correspondence. Privacy is therefore fundamental for accessing and using information without fear of

consequences. The IFLA Code of Ethics (2012) takes up these principles and explicitly states:

> [...] librarians and other information workers respect personal privacy, and the protection of personal data, necessarily shared between individuals and institutions… The relationship between the library and the user is one of confidentiality and librarians and other information workers will take appropriate measures to ensure that user data is not shared beyond the original transaction.

This organization further added in its "IFLA Internet Manifesto": "[...] library and information services... have a responsibility to... strive to ensure the privacy of their users, and that the resources and services that they use remain confidential..." (IFLA 2015). Similarly, the "IFLA Statement on Access to Personally Identifiable Information in Historical Records" says:

> [...] librarians should recognize an obligation to monitor their governments' legislation in regard to confidentiality of data records. In particular, librarians should support the need for privacy laws to protect library users from such abuses such as government agencies monitoring their reading and research habits (IFLA 2008). IFLA recently devoted an entire issue of its journal to the topic of privacy (IFLA 2008).

As can be seen, IFLA is another of the institutions that has also established for a long time fundamental principles in regard to privacy, protection of personal data and confidentiality in the user's relationship with the library or information service. From them, many other countries and organizations worldwide have collected and made these principles their own; for example, in Mexico, in addition to the law, the Code of Professional Ethics of the National Board of Librarians states: "[library professionals] shall keep total confidentiality of the facts regarding the information requested or received, the user's personal data, as well as materials consulted or borrowed, unless authorized by the interested parties". Many

other deontological codes of librarian practice include these principles; this points to an ethical obligation in addition to the law.

The aforementioned principles refer to privacy and confidentiality of personal data. It is necessary to clarify their difference within libraries: they emerged many decades ago with privacy, which means, in a library, the right of all users to read and consult whatever they wish without the subjects of interest being examined or scrutinized by third parties. Confidentiality stems from the fact that a library comes into possession of personal data making the user identifiable and therefore it must take the necessary measures to prevent unauthorized access. In other words, confidentiality is a process which protects –among many other things– the privacy. The latter is a right of every user; confidentiality is an ethical and legal obligation of the library to protect that right.

Since these principles were established nearly a century ago, countless libraries around the world have incorporated privacy policies and procedures into their information services, as well as measures to maintain user confidentiality. During the last century –before the digitally globalized world– this task was relatively simple, since the library processes which included personal data collecting were very few: the access to catalogs, indexes and other printed material in traditional media did not leave an associated trace of patrons and their interests. The only points of registration between user and the consulted material were, on the one hand, the loan slips, which were always destroyed once the book was returned, and on the other hand, the book cards, on which it was common practice to replace the user's name with a id number that made it impossible for third parties to associate names with readings. Libraries which kept microfilmed loan records periodically destroyed them, precisely for the sake of confidentiality. All libraries that extracted data for collection usage statistics always did so anonymously, also due to privacy principles. In the first decades of library automation, search and documentation services were offered by them through their own computer systems, so they had total control over access and therefore over the privacy of their users. Since the material to be consulted was inside their computers,

libraries could fully control it and prevent unauthorized access or transfer of sensitive data. As a general rule, libraries always refused to provide personal information of any kind to outside individuals or organizations.

Unfortunately, among the problems brought about by data globalization in the digital world there is the massive invasion of privacy and confidentiality of personal data. Freedom of information, privacy and confidentiality have been seriously threatened in the last two decades due to the rise of large-scale collection of such data, electronic surveillance and interception of digital communications. This is especially sensitive in big data, and hence its negative face in libraries. Obviously, this is not a problem exclusive to libraries, but it undoubtedly affects them greatly and therefore requires their meticulous attention.

During the 2019 Annual Meeting of the World Economic Forum in Davos, they released preliminary results from a global study on the public's perception about data privacy (Recovered from: https://www.ipsos.com/sites/default/files/global-citizens-data-privacy.pdf). The main findings highlighted that: 1) the vast majority of the public expresses very low levels of trust regarding the use of personal data by companies and governments globally; 2) only one third of the public has acceptable knowledge about data privacy; 3) two thirds of the public would feel more comfortable if they knew clearly how organizations use and share their data.

This is relevant because if libraries want to be among the institutions perceived by the public as "trustworthy" in terms of privacy and confidentiality of personal data, it is essential they handle and share personal data in a totally efficient and secure way, and also they make this process very transparent to users, so that libraries can effectively build in them the perception of trustworthiness in data security within the institution. Otherwise, libraries will unfailingly become part of that large conglomerate of organizations perceived as "untrustworthy" by the public, either because they do not handle personal data adequately or because –despite doing it well– they are unable to communicate it properly and transparently to the public. Gorman (2000, 36) points out that

respect for personal privacy is one of the eight fundamental values" for trust between the library and its users. In addition to the above, it should not be forgotten that in many countries –as is the case of Mexico– laws require all organizations storing personal data to comply with a series of provisions in this regard. For this reason, libraries are obligated subjects.

This is not trivial: due to the immense boom in digital information services and products with which the library and its patrons interact today, numerous aspects of users' data need to be taken care of. Unlike the paper slips of the last century, today there are numerous points of eventual collection of personal data in the library, even without the latter intending to do so; ALA indicates no less than a list of 32 possible points in this regard: they include electronic loan records, catalog search logs, associated items such as search histories, caches, cookies and certificates; e-mails, selective dissemination of information services, to name a few (ALA 2007). In addition to these "typical" data collection points, there will be those derived from special big data projects, such as the already mentioned about users' search styles and preferences, personalization of web pages, trends extraction in social networks, etc.

As if that were not enough, a large part of the privacy problem has been introduced by the interaction of digital networked publications and information services from a vendor outside the library. Increasingly, the works consulted, the discoverers, the search and documentation services, tables of contents, etc., come from commercial third parties. When a user decides to purchase or subscribe to certain information goods and services directly on a personal basis, it is common for the vendor to impose conditions which often threaten –among many other things– user's privacy. Whether that person wants to accept these conditions when accessing directly a digital service from a vendor is his/her choice. The problem becomes more serious and concerns libraries when vendors want such conditions to be extended inside the former, for whom they are obviously totally unacceptable. Among all the drawbacks of these products and business models, the aspect of privacy and respect for personal data has become one of the most

critical issues, as libraries and their patrons face serious challenges in this respect:

- Outside the institution, suppliers of commercial content and information services used by libraries may –and actually do– collect data about users' searches, activities, and transactions, or make it a pre-condition for the delivery of their services or content that libraries collect and transfer data to them.
- Cloud services hosting library systems can collect, store, and transfer user data outside of the library or information institution. Often the library is unaware of how and where such data is processed and stored in the cloud. It should be remembered that a huge proportion of these services are hosted on servers under other legal jurisdictions.
- The vast majority of all applications –apps– offered for mobile devices collect data on the identity, location, preferences and habits of their users. Many of these commercial apps are used by library or information services on a regular basis, and obviously these companies share the data they collect with third parties. A large number of these applications are apparently "free", but it should be kept in mind that when a user does not pay for a product or service on the net, the user inevitably becomes the product. In such cases, users always "pay" for the application with their data.

While privacy of library patrons is not an easy task, it is not an impossible one: it is a solvable problem. Although it has worsened significantly with the development of technology, it is not a problem that can be solved on a technological basis: like many others in the library, it is largely a matter of method and procedure. The problem of data protection –which is the technological part– must be addressed, but it is only a minor component. As has already been established, the major part contemplates even broader aspects.

First and foremost –in order to develop a method– each library must construct its privacy and data protection policies specific to

its context and features. Policies provide the library with a major structural foundation for planning and developing action programs to protect the privacy and personal data of its patrons. Policies should address the ethical and legal issues which constitute the organization's frame of reference, and should set out the broad outlines of who will define the issues within the organization about privacy and personal data, who will make and update related institutional plans and programs, who will supervise it, who will define the ones that can access the data, who will define how they are transferred, and who will safeguard them. Policies are drafted at a theoretical and macro-institutional level and therefore tend to be much more stable over time. Obviously policies must be consistent with the current legislation applicable to each country.

Starting from the policies as a basis, the library can then develop procedures, guidelines, best practices, standards, etc. Procedures and guidelines are the practical versions implementing the concepts outlined in the policies, detailing pre-established and sequential actions covering a variety of processes and sections of the library. Unlike policies, procedures and guidelines are specific, and therefore can and should change with some frequency as required. Therefore, elements of procedures and guidelines should not be incorporated into policies or vice versa. Policies establish *why*, *what* and *who*; procedures and guidelines establish *how*, *when*, *where* and –where appropriate– specify the *who*. The accumulated experience will generate good practices and standards.

To build the set of policies and procedures, it is suggested to start from the principles and practices recommended by "data governance". With them, the organization can develop in detail all aspects of who may collect and hold personal data in the organization, who may have authorized access to them and under what circumstances, who may modify them, to whom and how they may be transferred, and on the data protection side who has their custody. In order to ensure that all data collection points have been reviewed and procedures are in place, audits of all library services –both internal and those provided through vendors– should be

designed and carried out; obviously this includes big data projects. Based on all these documentary elements, there will be no process or set of contents in the library having not a person in charge, as well as a series of procedures and guidelines for the proper handling of each group of personal data.

With regard to audits in the library on the points of eventual collection of personal data in order to design procedures and persons responsible for their protection, the main points are as follows:

- Confidentiality of loan registers and books reservation;
- Confidentiality in the search of the library's internal catalogs. With regularity and method histories, cookies, caches, etc., should be deleted;
- Review and certification by the library about the privacy conditions of the applications —apps—generally used in the library;
- Review and certification about the privacy conditions of vendors of documentary goods and services to which the library subscribes. Libraries should always reject those providers and services which do not comply with the minimum conditions of patrons privacy, and warn other libraries about it;
- Review and certification by the library about the privacy conditions of its services installed in the cloud;
- Special security and privacy measures for young users and/or children;
- Security of library computers to prevent the introduction of malware designed to spy on or extract information from patrons;
- Security of the library's internal network, especially the wireless ones;
- Security of the library's social networks;
- Firewalls and encryption of sensitive data from patron records collected by the library;

- Warning, advice and training issued by the library to patrons about security risks while using search engines, services and tools external to the library;
- The library should anonymize as much as possible all data used for its projects, statistics or feedback.

This last item listed above has become the golden rule. Data that is not kept cannot leak or be extracted. In any service or application designed or built by the library, it should always be questioned beforehand exactly which data should be collected for it. In a vast majority of cases it happens that projects perform well without including personal data or with a minimum of them. Never, in any project or service, should such data be collected if it is not indispensable. In many cases of library data collection for analysis, it is entirely feasible to compile it without capturing sensitive user data; a typical example of this are the statistical usage of a university library's collections. In these cases, item data is usually recorded, such as call number, authors and title, as well as user's field of study, semester, age, etc., which are obtained from his/her library id number. But if the final data capture omits this last number and other sensitive data –not necessary at all for statistics– the library can perform extensive analyses for the desired purpose without the need to capture personal data that would eventually be at risk.

In services that necessarily require patron's identification, such as loan, books reservation, reference, or full-texts from a vendor, etc., the library must always ensure that it is collecting only the minimum amount of personal data, i.e., the strictly necessary data. In many cases, well-known methods can be used to hide it inside the applications, like "anonymizing" the data, as libraries have done for a long time. The above-mentioned old example of substituting id numbers instead of the user's name on book cards is still valid in the digital world. The use of "aliases" throughout the library's computer systems and files –whether numeric or textual– in substitution of user's name still works very well when recording loans, reserving books, computer or cabinet time, consulting catalogs, accessing vendor services, and many other points which

require identifying a certain user. Obviously, inside the library there is a main registry that has all the patron's data, but it should be centralized in one place, in a good computer "vault", behind some firewalls and other security mechanisms and preferably in encrypted files.

In relation to goods and services purchased from vendors, the aforementioned General Data Protection Regulation or GDPR of the European Union, besides being currently considered the most advanced regulatory framework on data privacy, is also one of the most valuable allies for librarians in this task, as it includes all the essential requirements towards suppliers. It is gradually becoming a worldwide reference framework, and therefore it is recommended that libraries outside the European region always check the terms of service of their vendors to verify to what extent they comply with this regulatory scope. The essential points covered by the GDPR in this regard, which libraries are therefore recommended to check meticulously are:

1) Privacy by design: companies should build and deliver their business processes including data privacy from the beginning, not as an afterthought. They should always have a data protection and privacy officer, easily identifiable, and independent of the operation;
2) Explicit consent: users should always be able to explicitly accept or reject privacy terms and conditions, including acceptance of cookies, before accessing a service or product;
3) Data Restriction: companies definitively may not collect "highly sensitive" data such as race, religion, political affiliation or sexual orientation;
4) Rights of access and portability: users have the right to request which personal information has been collected about them, and to request their personal information from other companies that it has received from second or third parties;
5) Right to be forgotten: users may request their data be deleted from a certain list;

6) Notification of intrusions: companies must always inform users when there has been a data breach within 72 hours of discovery;
7) Jurisdiction review: personal data must reside on known servers in a certain identified jurisdiction, and must not be arbitrarily moved out of it without express consent of the users.[25]

As has been seen, there are relatively simple methodologies to cut down or even eliminate the use of personal data throughout many of the library's applications and services. It is not desirable to duplicate personal data records across every department, section, service or project in the library. By minimizing the number of points where personal data is handled in the library, the number of places to care for reduces proportionally, making the task much easier for those responsible to do so. The set of policies, procedures, guidelines, standards, good practices, audits, and methodologies presented here makes evident the aforementioned concept that most of confidentiality consists of method and procedure, and only a minimal part of it are technological elements. The most important thing is still *how* things are done, not *with what* technology they are done. Techniques and tools do indeed exist to increase privacy in organizations, the use of which can and should be contemplated from the outset in libraries, but it is of utmost importance to emphasize that in no way, under any circumstances, can they replace proper methods to do things. Technological tools are complements, not substitutes.

Finally, the library should raise awareness and provide regular training to both its internal staff and its patrons on personal data privacy and confidentiality, computer security and related topics, so that they learn how to handle it properly. With respect to staff, all new employees, interns, volunteers, assistants, etc., should be

25 The European Court of Justice ruled in July 2020 that it is not valid for vendors to transfer personal data from Europe to the USA https://curia.europa.eu/jcms/upload/docs/application/pdf/2020-07/cp200091en.pdf

made aware from the very beginning that they must protect the privacy rights of library patrons. With regard to users, even though more than half of humankind is already connected in some way to the digital world, there is still a vast functional illiteracy in this area. Let's recall the second finding aforementioned presented at the 2019 World Economic Forum: only one-third of the public has acceptable knowledge about data privacy. The best way for the library to strengthen its privacy initiatives is to make its patrons and staff as literate as possible on this topic on a regular basis.

The tools for Big Data

It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.

SHERLOCK HOLMES,
'A Scandal in Bohemia', 1891.

For Big Data management there is nowadays a myriad of "tools"; this name is given to systems, software programs and applications, methodologies, algorithms, "sandboxes",[26] services, etc., both commercial and open access. To date, there is no great universal application covering all or at least most of the needs in this regard; instead, there are multiple specific tools with high specialties for each type of purpose, issued by various producers. No data project is equal to any other; each one has its own context and features. Therefore, when building a data project, it is seldom possible to acquire a single product for its solution; usually, it is necessary to assemble a set of software tools for its implementation and solution. Thus, one of the key starting insights in the field of big data consists in gathering a good general idea about all the possibilities and typologies among those tools, in order to be able to select the right one for each case and need.

26 A "data sandbox" in the context of big data is a development and test platform used for organizations to simulate small-scale data sets, their behavior, analysis and results, to verify how the fully operational set will behave.

There are many types: as an indicative and general list, there are software applications for:

- Data extraction from a wide variety of sources: social networks, specific programs and or files, the web, etc.;
- Data extraction from text, images, or voice;
- Subsequent data analysis and pattern discovery;
- Textual data analysis;
- Visualization, graphing and/or presentation of results in coherent forms;
- Data handling and management; obviously in large volumes;
- Artificial Intelligence; especially machine learning. Expert systems.

Because of their high specialization, certain tools have become commonly used in some key sectors: commercial, financial, health, education; and indeed, there are also tools for specific use in libraries and other information organizations.

Obviously, since there are so many of them, no one masters or even knows all the existing tools, but it is essential for every library professional to learn in general terms the offer as well as their possibilities and capabilities, in order to be able to make an eventual selection or discard any of them, just as they do with ILS – Integrated Library Management Systems, the specialized search engines or discoverers, automated catalogs or OPAC,[27] library automation systems, and other similar ones. No library manager masters or knows in depth all of the existing ones, but generally knows enough about all of them to eventually be able to select and acquire one of them properly; mostly, such manager learns what to ask, what to look for, and what tests to run to come to a right selection of an eventual system.

---

27 OPAC - Online Public Access Catalog, also known as "Automated Library Catalog", is a generic name for an online database of the records of documentary materials held by a certain library or group of libraries.

In addition, there are currently many data products, systems, applications, etc., already being developed by large libraries becoming available to smaller ones, which can take advantage of all these products and services for their benefit with little or no investment. The examples of the Linked Data Service of the Library of Congress, the similar one from the British Library, the National Library of Medicine Network (NNLM), the guidelines of the Smithsonian Institution Libraries for the creation and deposit of data in repositories, etc., have already been mentioned on this regard. Like these, there are already many cases of large library systems developing big data projects which can be used for the benefit of smaller systems or libraries without large investments. This has been gradually becoming a great source for them to acquire good tools in this regard. Likewise, library managers should start to learn about and study all these possibilities in the field of big data to identify the options and thus be able to take initiatives without large expenditures.

Besides, many of the libraries' usual suppliers are also making developments in the field of big data, with whom libraries can make strategic alliances for their use and promotion, or simply acquire on preferential terms some of the products resulting from these developments for their benefit. The SN SciGraph project of linked open data from the Natural Sciences division of Springer Nature, or the textual studies of OCLC, have already been mentioned as examples.

Outside the area of libraries –but not unrelated to them– there are also multiple options of useful tools. Today there are a large number of vendors dedicated to some facet of data processing and analysis. There are two main aspects in this regard: on the one hand, vendors offering a specific product to be acquired by an organization and integrated into its projects, and on the other hand, vendors supplying everything in the form of "bundles" or "packages", offering hardware, software, processing, storage, etc., in the scheme known as "the Cloud". All the main platforms of this modality are already offering a series of services related to big data; some of them free and others for a fee. This second aspect of

cloud services is beginning to be increasingly used, as it allows users to acquire large resources with relatively lower investments.

BIG DATA MANAGEMENT IN *THE CLOUD*

> *If you think you've seen this movie before, you're right. Cloud computing is based on the time-sharing model we leveraged years ago before we could afford our own computers.*
>
> DAVID LINTHICUM.
> "Cloud computing and SOA convergence in your enterprise", 2009

There are many definitions and conceptualizations of "the Cloud": it is not the case here to go into a detailed study about it. To understand the concept and to put it simply, "Cloud computing", or just "the cloud", consists of a set of computing resources of hardware, software and applications, storage, processing, communication, information, etc., which can be rapidly and ubiquitously delivered as a service via a network by a certain provider and widely scaled according to the needs of a certain user. The essential difference between this scheme and its predecessors is that for several decades the business model for the supply of computer hardware, software, communications, etc., was managed as the provision of products. Under this new concept, the "cloud computing" business model consists of the delivery of such resources as a service rather than as a product, through shared resources over a network, in which equipment, applications, storage, information, data, infrastructure, communications, etc., are provided in the same way as community services such as water, electricity or gas, paying only for what is consumed.

Originally, about a decade ago, cloud computing variants were divided into three basic service models, also called "layers":

- Software as a Service –or simply SaaS. In this service model, the customer uses applications or programs running on a remote server of a network provider –in the cloud– and does not manage or control the infrastructure or platforms on which these applications run, such as the type and/or model of servers, operating system, networks, etc.
- Platform as a Service –or simply PaaS. In this model, customers can develop their own applications or systems, either for their local network or in the cloud, and for this purpose they rent access to a programming platform; i.e., a place where they can select and use the operating system, libraries, compilers, packages, storage capacity, etc.
- Infrastructure as a Service –or simply IaaS. In this service model, the supplier provides its customers with a basic computing and telecommunications infrastructure, normally under a scheme which provides variable resources on demand operated by the customer at will. This infrastructure is usually an assembly of hardware, software, networks, storage, support, security, training, and so on.

Subsequently, with the evolution of the cloud, many other variants of service models have been added: Among others, we find now: Content as a Service –CaaS, Network as a Service –NaaS, Security as a Service –SaaS, Preservation as a Service –PaaST, and of course, new services specifically dedicated to data: Data as a Service –DaaS, as well as Data and Platform as a Service –DaPaaS.

In the pre-cloud stage, organizations hosted and processed their own data in a proprietary, stand-alone computer system. The downside of such model was that –as the data became increasingly voluminous and complex– it became increasingly expensive to operate and maintain. With the Data-as-a-Service or DaaS model in the cloud, data are compiled, organized and delivered by a provider as a rented service, and is easily accessible over the network. Customer management and tracking thus became a rented service on platforms external to the organization. But Data as a Service is only the most obvious part of the resources used in the cloud

environment in regard to data: as can be seen from the above, all types of infrastructure can be additionally rented: hardware, storage, network; software for projects; operating systems, generic and specific programs and applications for data, software libraries, processing, training, and so on.

In addition, cloud services are constantly diversifying and becoming more sophisticated, so it is necessary to be alert to study the emerging variations, in order to evaluate their eventual selection or discard. An example of this is the so-called Edge computing. In cases of Data as a Service or DaaS rental, sending data generated by devices to a central server in the cloud may cause bandwidth problems and delay times. Edge computing proposes a more efficient alternative: processing and analyzing data at or near the user's physical location, at or near the data source. By not having to travel through the entire network to a central server for processing, the delay time is significantly reduced, thus faster and more reliable services are obtained. The utility and efficiency of this modality is not yet evident, but Gartner asserts that by 2025 this type of cloud service variant will comprise 25% of the activity in this environment, so it is worth re-evaluating it from time to time (Van der Meulen 2018). Like this one, there are other variants emerging regularly.

Due to its massive nature, it is common nowadays that a good part of data management is to be done in a cloud resource. Therefore, one of the first decisions to be made in this regard is to decide what proportion of services will be used in that environment and in which of them, and obviously the counterpart: what proportion of data or applications will be used within the computing environment owned by the library or the organization on which it depends. This should be part of the plan from the beginning, as it has a direct bearing on cost, time, infrastructure, security and privacy. Although cloud services are generally purchased for economy of scale, this is not necessarily an absolute rule: many libraries already owning their IT infrastructure can use their remaining capacity for data projects at very low costs, without the need to purchase new external services. Cost-benefit studies and

comparisons should be carried out from the outset to be able to make the right decisions in this regard.

This makes the combinations between owned and rented infrastructure very broad: it means that storage can be rented but not process, software but not equipment, process but not data, and so on. As in many other technology initiatives in the library, a commonly used approach is to initially acquire free or open-use platforms to start testing ideas, developing applications, verifying objectives, and so on. Obviously these platforms are limited in scope or features by their very nature of being free, but they serve very well to validate projects. Once stable and if the library has been convinced of their usefulness and benefit, a shift can be made to the paid versions of these platforms.

Nowadays, practically all the major players in the world of Information and Communication Technologies –Google, Amazon, Microsoft, Apple, Facebook, Oracle, IBM, etc.– offer some kind of product or service related to data management and analysis, some of them open or free for promotional purposes, while others are paid. Libraries have been taking advantage of these resources for some years now, incorporating them into their work. The possibilities are extremely extensive: from minimal data extraction and insertion for specific services to large projects related to them. Obviously not everything offered by these large corporations has to do with data, but a very significant part of their products and profits come from this sector. Therefore, their main services and products closely related to data will be reviewed here.

Beyond its famous search engine, Google started offering some basic services in the cloud, and gradually incorporated new tools, applications, services, etc.: a whole range of processing possibilities and technological services; among them at some point they started with data management. In recent times, this company decided to integrate everything related to cloud services in a single site to better give visibility and access, and thus created the structure called "Google Cloud". On this site, they offer a large number of possible services of all kinds in the cloud, such as those already mentioned: access to servers, storage, software, etc. Many of these

services and products have to do or can be used closely with data management. For instance: Google Charts. This open access tool is easily implemented by embedding JavaScript programming into the HTML code of a certain website allowing to connect to a database and then extract, sort, filter, modify, and visualize its data. Google's cloud also offers access to super processors called TPU or Tensor Process Unit –used in its search engine, translator, photo identifier, Google Assistant and Gmail– in aim of enabling customers to develop AI applications on extended-capacity computers.

Another typical example of data extraction and use through the Google platform are the developments known as "mashups". Basically, it consists of a web page using and combining data, presentation and functionalities from one or more sources to create new rich services easily and quickly, while integrating applications, data sources and open sites through a graphical interface. In doing so, a library can collect a certain set of data and then integrate it with the open access application Google Maps, and thereby display to users visualizations, graphically presenting a set of data and information specific to that library, with the typical presentation, appearance and facilities of that company's maps. Google Mashup Editor is another tool of this platform enabling mashups of all kinds to be made with library data; multiple case studies of library services made with mashups can be seen in Engard (2009, 2012), and in Stephens (2011). There are other available platforms to create a mashup: Amazon, Facebook, Twitter, LibraryThing, Flickr, Ebay, YouTube, JournalTOC, and so on. In addition, there are multiple software pieces or editors from vendors to build the mashup, both in free and paid versions; among them: Yahoo Pipes, Microsoft Popfly, IBM Mashup Starter Kit, Intel's Mash Maker, and DreamFace Interactive. As mentioned, their free versions are adequate and sufficient options to start with, despite being basic; once the technique has been mastered, it is convenient to migrate to paid editors to be able to build more complex and sophisticated mashups, since they offer more options and detail.

From the wide number of data products and providers, it is important for the library manager to begin learning about and

studying all the possibilities in order to discern the options and thus be able to take initiatives in the field of data in a concerted manner. Since all platforms offer similar services, two approaches can be used to properly start this knowledge: on the one hand by studying the offer of each of the platforms in this matter in their different variants, and thus know the total offer of a certain platform about data, and on the other hand, studying the tools by type to learn all the options in a specific variety of service through all platforms and thus enter into the possibility of making comparisons between them. Both approaches work well in practice, and it is a matter of preference.

In the first approach, i.e., by company or platform, practically all the major and well known technology vendors offer various data handling and analysis tools: Google, Amazon, Facebook, Twitter, Apple, IBM, Microsoft, LinkedIn, YouTube, Flickr, Ebay. To these should be added some large providers of these tools not so well-known to the general public, such as Apache, Oracle, Hana, or Gaia-X. The study of the tools offered by each of these platforms serves as the first approach to them. In other words, a company is selected and all the possible data management and analysis tools provided by it are revised in order to have an overview of the entire data offer of such organization. Obviously, the same is done with all the other companies.

The second approach consists of analyzing the tools for each type of them, and then comparing the various offers provided by different companies. This second approach will be used here for the review of possibilities. For this purpose, we have arbitrarily divided the tools into the following types:

- Database managers, both SQL and NoSQL;
- Documentary data managers;
- Tools for data normalization and data mapping;
- Tools for big data analysis, in order to extract patterns or trends from them;
- Tools for visualization, interpretation or presentation of results;

- Artificial Intelligence tools;
- Specific application tools, such as those for mashups.

SQL AND NOSQL DATABASE MANAGEMENT SYSTEMS

> *Petabytes allow us to say 'correlation is enough'. We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let the statistical algorithms find patterns where science cannot.*
>
> CHRIS ANDERSON,
> Former Editor-In-Chief of Wired magazine, 2018.

SQL and NoSQL database managers have been mentioned. The fundamental difference between them is that the former were designed to handle structured data and the latter specialize in unstructured ones. According to the nature of the data and the sender, it can be divided into structured, semi-structured and unstructured data. It is important to discuss these concepts in a basic level to understand their importance and differences.

Structured data have a well-defined shape and format, and therefore can be easily represented homogeneously in tables with rows and columns. Each row is known as a record[28] and each column as a "field"; thus a contiguous sequence of columns or fields makes up a row or record. They derive from a data model or data schema; i.e., a pre-established way on how they can be processed,

28 This concept of "record" comes from computer science and obviously differs from the meanings of library and archival sciences. From this conceptualization, a "record" is a set of consecutive "fields" in the form of a row. Each field contains data of a certain type –numeric, alphabetical, date, etc.– A set of consecutive records conforms a "file".

stored and accessed; consequently, they are easier to manage or handle. Because of their structure, each field contains a single, easily identifiable element and can therefore be readily accessed separately or in conjunction with data from other fields.

Data models uses certain typical elements for this type of data, divided into two categories; *static* elements: fields, records, arrays, character strings; and *dynamic* elements: lists, stacks, queues, trees. Each of these terms has a particular significance within data management. We can briefly mention some examples in the library field to better understand such particular meanings: A typical case from this conceptualization of record is each entry in the library's user directory. Each record of the same contains one user's data, in the form of perfectly defined and pre-established fields with single data for each of them: id number, name, school, degree program, address, date of birth, telephone, e-mail, or alike. The consecutive addition of all these records for each user placed in this form becomes the directory. Each and every one of these fields has a type predefined by a data mode: numeric, alphanumeric, date, and so on; as well as a certain format with predefined length: ten characters for a phone, eight characters for the date of birth in mm/dd/yy format, five digits for the zip code, thirty alphanumeric characters for the e-mail, or something similar.

An array is a finite collection of elements of the same type in a defined format, i.e., they are homogeneous and have as distinguishing feature being ordered by means of an index. A typical example of an array is a predefined table with the numeric keys of the names of the different libraries within a university. In order to keep the data homogeneous, during the processes the numerical key is captured instead of the name, and only eventually, when desired, the key/name duple is associated again and the full name is displayed. This saves space and maintains homogeneity in the names.

In turn, fields or certain elements of an array may contain character strings: alphanumeric character sequences with a certain significance. In their simplest expression, they are short elements, such as a person's name, a postal address, an e-mail address, and

so on. When these strings are short and precise, they can be handled as structured data.

The *dynamic* type elements are: lists, stacks, queues, trees; and they are used in computer systems as auxiliary structures for storage and retrieval; a typical example of this are the word lists built for retrieval in an OPAC catalog; in practice, computer system stores lists where each word has been translated into a number, and points to all the card numbers containing it; this saves time and space and significantly increases the computer retrieval speed. Similar operations are performed with stacks, queues, trees, and so on; each one of them used for a specific purpose. .

As can be seen, in this type of data their access and handling is facilitated due to their structure. For decades, theories, principles, languages, algorithms, tools, etc., have been created in order to manage them. All the initial theory of databases was created around the aforementioned typology. Since then, hierarchical and relational Data Base Management Systems or DBMS were designed and built in countless versions and by multiple manufacturers around these concepts. For more information about relational databases see (Oracle, s.d.). All of these managers belong to a large category known as SQL managers, since the form or common language which enabled queries to be made on this type of data was called from the beginning SQL (Structured Query Language). Hence the name SQL Databases.

Unstructured data has an internal form but is not structured by predefined data models or schemas. Such data are not necessarily numerical or textual, and are generated by both humans and machines. It is very difficult to normalize them, since they do not have defined types and are not organized under any pattern. However, in many cases it is desirable to process this type of data, and therefore it is necessary to be able to organize, classify, store, search, delete, etc., them in some way, and therefore it is essential to find ways to do so. A typical example of this data is the contents of a web page: the page itself has an internal structure in HTML format, but its contents are scattered and it is very difficult to identify and separate the data within its parts: texts, images,

audios, videos, buttons, boxes, etc., embedded in it. Extracting information from its internal data is a very complicated task. The only thing that the page usually contains as metadata are its HTML "meta" tags: keywords, description, etc., but these are minimal data making it very difficult to extract the rest of them from the page for further processing.

Another typical example of this type of data is video content. If these are not accompanied by a fiche or similar record with metadata, it is extremely difficult to extract information from them: Who are the people who appear there? What are the dates, places, format, and duration? Who produced, directed, edited, wrote, photographed, etc.? What is the video about? What is its theme? In some cases this information is written as credits on the video, but in many others it is not. And even when it appears, it is necessary that someone observes the video to extract personally such information, since it is difficult to get a machine to do it, because the way of recording them has never been standardized. There are many other types of documents that –if they are not already accompanied by their metadata– it is difficult to extract their data: texts from office processors, Power-point presentations, audios, photographs, blogs, voice messages, digitized texts or images without metadata, and so on. Because of their nature, special databases were developed to store and manage these contents in structures called non-relational databases. These types of databases are known as NoSQL databases, or Not Only SQL databases. The major difference between them and structured data is the ease of analysis. For structured data, mature analytical tools already exist, but tools for analyzing and extracting unstructured data are still in the early stages of development and have a long way to go.

The intermediate stage between these two types of data is semi-structured data. As the name implies, it contains structure and form in some parts of its content, but it does not in others. The most typical example of this type of data within a library is a catalog entry; as is well known, it contains a series of "fields", each with an identifying label or tag followed by indicators, and then the content of the field itself; it may also include some markers: comma,

percent, colon, and so on. A complete catalog entry is a complex string of characters, since it is in turn made up of sub-strings, i.e., each of the fields of the entry, all of which are also strings of characters: call number, author, title, imprint, subject, etc. Given its dimensions and shape, it is no longer considered structured data, as is the case of a short or simple character string. It is considered semi-structured data because it does contain pre-established and identifiable divisions –those indicated by tags– but within them textual elements can be very long and complex, such as corporate authors or titles, which are not easily disentangled. Moreover, their very long and totally variable length makes it impractical to define fixed-length fields for this purpose.[29]

Another clear example of this type of data are e-mail messages, which by their own nature on the one hand, are area series of well-defined and pre-established fields: name and address and IP of the sender, name and address of the recipient(s), date, subject, message identification number, attachments, and so on. However, on the other hand, the actual content of the message is not structured; it is simply a text in the form of a character string without further definition, and this is the most important part. Since this entire part lacks structure, it is difficult to extract data from this section, which usually is the most important one. A further example of such data is XML documents. The part of the defined tags has a good structure; the part of the actual contents of those tags does not necessarily have a good structure. The great advantage is that this tag-based structure is highly flexible and adaptable to various needs to homogenize as much as possible certain data forms

29 An example of this is the "title" field; there are ones as short as "She", by Rider Haggard to ones as long as "Western Central Atlantic Fishery Commission: Report of the fifth session of the Scientific Advisory Group, Puerto Morelos, Mexico, 28-29 October 2011 = Rapport de la cinquième session du Groupe Scientifique Consultatif, Puerto Morelos, Mexique, 28-29 october 2011 = Informe de la quinta sesión del Grupo Asesor Científico, Puerto Morelos, México, 28-29 de octubre de 2011".

of many types of documents, and these structures are machine-readable, which makes the work easier.

NoSQL unstructured database management systems can also handle semi-structured data, as they differ from relational ones for not separating the data from its organization model or schema. This makes them a better choice for storing and processing information not easily fitting into a table or other static record format; this is the case for texts of highly variable lengths. In addition, these systems allow easier and more fluid data exchange between different databases. The aforementioned library catalog entries are an excellent example of this type of data, difficult to handle in the form of a table or record of fixed and length-predetermined rows and columns in a computer system. Therefore other ways of representing and storing them efficiently had to be devised from the outset. MARC's development of tags or labels for variable-length fields was one of the great methodological contributions to this type of data.[30] Data management systems specialized in library catalogs are a long-standing example of this type of tools.

Obviously, the more structured the data is, the easier it is to manage. A major part of the problem is that structured data in the digital world is estimated to represent only 5% of what is produced. Semi-structured data comprises 15%, and the rest is unstructured data; metadata is added to 3% of it (*The Digital Universe…* 2014). Clearly, the opposite is also true: the less structured the data is, the more difficult it is to extract valuable information from it; this is why metadata is so important. As is well known, metadata is "data about data"; without it, it would not be possible to derive anything usable from these data sets, especially the massive ones, and they would become an amorphous and useless bulk with little or no utility. From this follows the great importance of the aforementioned big metadata extraction and correlation

30 The origin of the "tag" structure and "text markup" comes from the metalanguage called "GML" or "General Markup Language", originally developed by IBM for editing their manuals. The LC MARC project perfected the idea and made it universally applicable in the 1960s.

projects such as those of the British Library or the US Library of Congress, reaffirming the importance of libraries regularly adding metadata to their information and collections.

As can be deduced from the above, the tools for the management and analysis of big data vary greatly depending on the type of data they are built to handle, and this is why it is important for librarians to study and understand the characteristics of the different types of data from the outset in order to be able to select the appropriate tools for each case. As can be seen, there is neither a universal data manager nor a universal data analyzer, so the selection and use of the appropriate tool for each case is fundamental. Again, it should also be kept in mind that some of them are open access and others are commercial products and services. Both have their advantages and disadvantages.

Regarding database management systems, without being an exhaustive list, the following can be found:

For structured data, the management systems traditionally used for these purposes: Oracle, MySQL, PostgreSQL, Microsoft Access, SQL Server, FileMaker, MariaDB, RDBMS, OpenOffice Base – https://www.openoffice.org/product/base.html. In turn, large cloud platforms offer services based on and compatible with these managers, such as Google's CloudSQL – https://cloud.google.com/products/databases/, AWS Relational DataBase Service – https://aws.amazon.com/es/rds/ and AWS Aurora, both from Amazon – https://aws.amazon.com/es/rds/aurora/, Azure Database – https://azure.microsoft.com/en-us/services/sql-database/ and Azure SQL Database – https://azure.microsoft.com/en-us/product-categories/databases/, both from Microsoft, Hive from Apache – https://hive.apache.org/.

For unstructured data, the following NoSQL database management systems are found; almost all of them in the form of a cloud service: Amazon's AWS DynamoDB – https://aws.amazon.com/dynamodb/?nc1=h_ls, Google's Cloud BigTable – https://cloud.google.com/products/databases/, HBase – https://docs.microsoft.com/es-es/azure/hdinsight/hbase/apache-hbase-overview and Cassandra – https://cassandra.apache.org/, both from Apache, Azure CosmosDB from Microsoft – https://docs.microsoft.com/es-es/azur

e/cosmos-db/introduction, Voldemort from LinkedIn – https://www.project-voldemort.com/voldemort/, Redis – https://redis.io/, and Riak – https://riak.com/.

As they are of special interest to libraries, some of the specific mangers for text or documentary databases are highlighted here: CouchDB or Cluster Of Unreliable Commodity from Apache – https://docs.couchdb.org/en/stable/, MongoDB – https://www.mongodb.com/es, BaseX – http://www.basex.org/, etc. As in the other NoSQL type managers, in this type of tools the data is not stored in tables, but the database is composed by "documents" – of very variable lengths– which in turn function as "objects". They are usually associated with other tools for extracting and analyzing information from text, which will be discussed later. As in many other cases, some of them are commercial products or services, and others are open source.

Tools for data normalization and data mapping

> *You can have data without information, but you cannot have information without data.*
>
> Daniel Keys Moran

In addition to the database management systems –relational or not– for big data handling a set of tools it is required prior to their exploitation, to enable certain preparatory processes to be carried out on them: data normalizing, data "mapping", arranging it in logical sets, debugging, dividing it into manageable segments, etc. This is one of the important preliminary steps of "data curation" for proper data sorting, discarding duplicate, irrelevant, excessive, inaccurate data, and making it suitable for processing and analysis. Among this type of tools, the following stand out:

Hadoop, from the Apache software developer foundation. It is an open source tool for general data processing. For many organizations, it is currently the reference software for the flexible handling of large volumes and varieties of data. It allows to group,

manage, and process large sets of massive structured, semi-structured and unstructured data. By implementing this tool in an organization, it becomes possible to organize and process data in a certain standardized way for further exploitation, as it allows to map data sets to manageable logical structures. Data mapping basically means homogenizing data represented distinctly in different systems used simultaneously; i.e., the names of the countries of the world: in one system they may be captured with their full English names: United Arab Emirates, United Kingdom, Mexico, United States of America, and so on. In another application with acronyms: UAE, UK, MEX, USA. In another system with their Internet domain acronyms: AE, UK, MX, US; in yet another with their names in other language; for instance in Spanish: Emiratos Árabes Unidos, Reino Unido, Estados Unidos de América, etc. By mapping these data sets they are normalized to a certain unique form for a given application in order to make them manageable, otherwise they would be unusable. In many data processes such as integration, migration, synchronization, data warehouses, data mining, and so on, the cleansing of unusable data and its normalization is inevitably required: Hence the importance of mapping, and therefore why this type of tool is frequently required for these tasks. There are other similar ones: to "clean" or refine the data, there are tools such as Open Refine; when data becomes too voluminous, tools such as MapReduce are used to distribute it into several more manageable logical sets. Adaptive MapReduce is IBM's version for this type of redistribution and disaggregation tasks; Talend Open Studio –https://www.talend.com/es/resources/introduction-talend-open-studio-data-integration/ is another open source tool for data integration and/or synchronization. It allows to build connectors or junction points among all source and target systems of the organization, by means of data integration models, in order to homogenize and standardize them. In fact, already standardized data sets to be used across several departments or areas in an organization, constitute a special and relevant category within data management called master data, and thus receive special treatment.

When the developer of a library project is building the basic idea, after selecting the database manager the next step is to think about how data is going to be organized for a certain purpose, and subsequently with which and how data is going to be exploited. This is the right time to select all the relevant tools to prepare, organize, normalize, and clean the data, having a number of them to choose from for this purpose. The above-mentioned are some examples. This should not be considered a superfluous step and therefore be avoided, since starting from disorganized, non-debugged or excessive data may cause delays and failures.

TOOLS FOR BIG DATA ANALYSIS, TO EXTRACT PATTERNS OR TRENDS FROM THEM

The value of big data is not in the data; it is in the analytics.

GARY KING,
Harvard University

The next type of big data tools are those specialized in analyzing and extracting information, trends, patterns, etc. from big data, according to the objectives of each project. Generally speaking, this area contains the core of big data; i.e., detecting trends and patterns in it, in order to propose solutions. By the very nature of this area, there is an amazing variety and subdivisions of computer applications, for an infinite number of different purposes. As with the previous types, there are those of multiple manufacturers, platforms, and specialties; commercial and open access. The following list is not intended to be exhaustive, but to provide a representative overview of the very extensive choice and purposes of this category of tools:

Google's BigQuery – https://cloud.google.com/bigquery, which is a cloud service in the form of a highly scalable data warehouse capable of analyzing huge amounts of data with SQL techniques for information extraction. This company also offers a tool called Google Analytics –https://analytics.google.com/analytics/academ

y/, which allows to obtain aggregated information about the queries to a certain website according to its audience, behavior, trends, etc. They are widely used for website promotion, as well as evaluation and monitoring purposes.

Infosphere Streams – https://www.ibm.com/developerworks/library/bd-streamsintro/index.html, a platform developed by IBM. It is designed to discover in minutes to hours, meaningful patterns from dynamic streams of data, in windows. From the same company is ThinkUp; an open source tool for data analysis which enables extracting information from Twitter, Facebook and Google+.

BigInsights is an analytics platform allowing organizations to turn Internet-scale complex information sets into insights. It consists of a packaged distribution of Apache Hadoop, with a greatly simplified installation process, as well as associated tools for application development, data movement and data clusters management. A *cluster* is a platform of several computers synchronized to achieve high performance.

SAP Hana or System Applications Products High-Performance Analytic Appliance – https://www.sap.com/spain/products/hana.html. Together with SAP Predictive Analytics they have the ability to integrate and process big data workloads to be analyzed in real time.

Oracle Big Data Appliance – https://www.oracle.com/engineered-systems/big-data-appliance/. It is a platform developed by Oracle with wide coverage of functions to acquire, organize and analyze big data loads from various sources at high speed.

Azure HDInsight – https://docs.microsoft.com/en-us/azure/hdinsight/ is a Microsoft cloud service based on Apache Hadoop, which allows interacting with many programs of that platform such as Apache Spark –for real-time data analysis–, Apache Hive, Apache Kafka, and Apache HBase in order to process and analyze multiple data.

Splunk – https://www.softtek.com/es/tecnologias/splunk, which typically specializes in leveraging machine-created data from many different sources, such as websites, applications, IoT, and sensors.

SPSS or Statistical Package for the Social Sciences – https://www.ibm.com/products/spss-statistics. One of the oldest and

most classic software tools for statistical analysis. This system was created in 1968 by scientists at the University of Chicago for large computers or mainframes, becoming a pioneer of this type of tools. In 1975 SPSS Inc. was created for this purpose, and later acquired by IBM in 2009. Currently, version 26 of this tool is issued for multiple platforms. Despite being the oldest, this software is still one of the most widely used for statistical analysis due to the enormous amount of tests and analyses it can perform; it also has interfaces to other tools such as SAS, Matlab, Statistica, etc., as well as to routines in R language. There is a quite acceptable free version of this tool called PSPPire – https://www.softpedia.com/get/Office-tools/Other-Office-Tools/PSPP.shtml.

Statwing – https://www.statwing.com/. Similar to the previous one, it is a software tool used to perform classical statistical analysis on data sets, to extract from them the typical elements of this discipline: parameters, regressions, correlations, etc.

LibInsight from Springshare – https://springshare.com/libinsight/ is a software tool used to collect and analyze statistics specifically in libraries. It stores all library data in a single platform, and uses cross-dataset analysis techniques to optimize decision making based on it.

DisplayR – https://www.displayr.com/migration/ is a general-purpose software tool, which includes modules for statistical analysis, machine learning, data analysis and visualization, with an interface based on the R language.

"R" is a programming language developed by the Foundation for Statistical Computing which is widely used for computer-assisted statistics and data mining, as well as for graphical visualizations. Systems and applications for data analysis can be quickly developed with this language – https://www.r-project.org/.

SAS language was designed to operate mainly on data tables: it has various options for reading, transforming, combining, summarizing, and displaying them, as well as for multiple statistical analyses of the data: its main modules to this respect are: (a) SAS/STAT, with procedures for performing typical statistical analyses – regressions, correlations, etc.–; (b) SAS/ETS for statistical analysis

of time series; (c) SAS/IML, for implementing alternative languages similar to Octave, Matlab,[31] or R; (d) SAS/OR for solving Operations Research type problems; and (e) SAS/GRAPH for generating graphs. It also has additional modules for other tasks, such as SAS Enterprise Guide for training, SAS Data Integration Studio for data mapping, and SAS Enterprise Miner for data mining – https://www.sas.com/en_us/solutions/analytics.html#.

Finally, it is worth mentioning that the classic and ancient SQL – Structured Query Language, despite its simplicity –or perhaps because of it– is still widely used to extract and analyze data in Relational Database Management Systems or DBMS. To date, anyone who wants to start learning database management tools should begin by learning SQL.

Tools for text analysis in big data, to extract their information or trends

The goal is to turn data into information, and information into insight.
CARLY FIORINA,
Former President of Hewlett Packard

As has been mentioned, text analysis or text mining is one of the areas of particular interest in the library environment, due to the multiple applications that can be given to this type of data in this field: identification of texts, extraction of elements from them, categorization and/or taxonomy of texts, extraction of concepts, entities, relationships, events and other metadata; translations, and so on. There are a number of products developed expressly for the analysis of information coming from texts, in practically all of their variants, which is why they are of particular interest for the library environment. Again, without being an exhaustive list, the following stand out among them:

31 Matlab is a specialized programming language for developing projects involving heavy numerical calculation. GNU Octave is similar but open access.

The programs of the Apache Software Foundation. This organization has developed several software pieces for the purpose of text analysis and extraction of valuable information from them. The core of the logical architecture of these tools consists in the concept of documents containing hypothetical fields; this allows them to be independent of the format of the computer file: pdf, html, txt, doc, odt, etc. Among them the main ones are: Apache Lucene, which is a software library for full texts based on Java language, which provides a platform for searching and indexing elements within the text – https://lucene.apache.org/. Also Apache OpenNLP, which is an open access tool based on machine learning techniques for natural language processing – https://opennlp.apache.org/. They are complemented by Apache UIMA, an unstructured information management software for capturing plain text and identifying internal entities, such as persons, places, organizations; or relationships, such as *located in*, or *associated with* – https://uima.apache.org/.

Google Cloud Natural Language API. It uses Google's cloud storage and processing with machine learning techniques to find the structure and meaning of unstructured text. It extracts information about people, places, events, opinions in social networks, etc. – https://cloud.google.com/natural-language/.

Textalytics. A software developed by Daedalus. It easily extracts much of the content of all types of documents, especially in social networks – https://www.programmableweb.com/api/textalytics.

IBM SPSS Text Analytics. This software allows capturing survey data, extracting key concepts, proposing results and categorizing responses – https://www.ibm.com/support/pages/downloading-ibm-spss-text-analytics-surveys-401. It is complemented by IBM Watson Natural Language Understanding, which is a cloud service using machine learning techniques to extract metadata from texts, such as entity-relationship items, syntax, keywords, categories, and opinions – https://www.ibm.com/cloud/watson-natural-language-understanding.

Microsoft Azure Text Analytics API. This software discovers ideas in unstructured text using natural language processing, and does not require expertise in machine learning systems. It identifies and extracts key phrases and entities such as people, places, organizations, opinions, among others, for the purpose of understanding common themes and trends, in a wide variety of languages – https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/.

GATE – General Architecture for Text Engineering. It is an open access tool based on machine learning to find structure and meaning in unstructured text in various formats: html, pdf, doc, text, odt, etc. It detects and extracts information about entities such as people, places, organizations, opinions or events – https://gate.ac.uk/.

DiscoverText is a set of simple and advanced cloud-based software tools which allows to quickly and accurately evaluate large amounts of unstructured text, as well as associated metadata, from surveys, chats, email, public comments, Twitter, RSS feeds and other forms of text data – https://discovertext.com/

Semantria, Semantria API and Semantria for Excel. Similar to the above, it is a set of cloud-based software tools for evaluating large amounts of unstructured text, from surveys, chats, email, public comments, Twitter, and other similar texts. It has a version that can be added to Excel for analysis of data contained in such software – https://www.lexalytics.com/semantria

Lexalytics Salience. Tool for extracting information, scanning natural language texts from content and metadata of social networks, especially Twitter – https://www.lexalytics.com/salience/server

Provalis Research Text Analytics. Assembly of Prosuite, QDA Miner, and Wordstat software tools for text mining, data capture from various sources, content analysis, extraction of entities such as keywords, phrases and topics; classification, identification of patterns and trends, visualization, maps, etc. – https://provalisresearch.com/

SAS Text Miner. Software tool based on machine learning for extraction of topics, keyword entities and relationships from unstructured data, especially texts, and from them to develop predictive models – https://www.sas.com/en_us/software/text-miner.html

Text2Data. It is a set of software tools and cloud services for extracting and classifying key entities from text; performing content and sentiment analysis. Like Semantria, it has an extension to be used in combination with Excel, and another one to be combined with Google Sheet – https://text2data.com/

The list of this type of software tools continues with a number of manufacturers and products which perform one or more of the above-mentioned text-related tasks: Rossette Text Analytics, Stratifyd, Luminoso, Bitext, NetOwl, Natural Language Toolkit, Aylien, Expert System, Smartlogic, Ascribe, Datumbox, Indico, RapidMiner Text Mining Extension, Keatext, Pingar, TextualETL, KH Coder, QDA Miner Lite, TAMS, Visual Text, Pentaho, and so on.

As can be seen from the above list, the range of options is really extensive. As in other cases, and depending on whether the data project is mainly textual or not, the library manager must begin by studying the problem to be solved and the more specific characteristics of the data to be analyzed, in order to be able to select the ideal set of tools to be used for data extraction and analysis, as well as those for data visualization and interpretation. Many of the tools mentioned in the previous lists are open access, making them excellent options for librarians to begin to venture into textual big data management, Artificial Intelligence, machine learning, etc., avoiding large initial outlays.

Tools for visualization, interpretation or presentation of results

> *[...] statistics, like cakes, are good only if you know who made them and you are sure of the ingredients.*
>
> LAWRENCE LOWELL,
> Dean of Harvard University, 1909.

The next group of software tools consists of those for data presentation and visualization. This is a very important component: there is no point in collecting a lot of data if it cannot be properly

interpreted for decision making. The data itself is of little value; the knowledge that can be extracted and interpreted from it is the real valuable contribution. Large sets of figures whether in tables with rows and columns or in endless lists can represent very little in terms of pertinent information, hence the importance of data presentation and visualization. This is particularly relevant in the big data environment, where millions of data of very different natures and significances can be collected.

Being able to abstract all that into simple forms of visualization is in some sense a technique and in other is an art, since it must combine simple design with logic and numbers, something that is not easy to match, but is undoubtedly a primary factor of success while handling big data. In fact, there is already a whole sub-discipline around it: "visual literacy", also called "visual skills". Orland-Barak and Mazkit (2017, 11) define it as:

> [...] the ability to interpret, negotiate and make meaning of information presented in the form of an image, extending the meaning of 'literacy', which commonly signifies interpretation of a written or printed text. Visual literacy is based on the idea that images can also be 'read' since their meaning can be understood through a process of reading.

In regard with this discipline, there is already a Journal of Visual Literacy and an International Visual Literacy Association.

This discipline has evolved significantly in recent years, precisely because of the growing need of being able to express adequate information in an efficient visual form. This is especially important in these times, plagued by information which is biased, wrong, misleading, deceiving, unfounded, poorly interpreted, ill-intentioned, incomplete, or simply fake. To maintain the perception of credibility, it is essential that the information displayed in libraries be of high quality, and therefore visual skills are especially important in these organizations. Edward Tufte, a pioneer in the field and currently one of the most recognized authors on the subject, recommends the use of data-rich yet simple illustrations,

and summarizes the purpose of data visualization as follows: "[...] the designer's task is not the complication of the simple, but rather to give visual access to the subtle and the difficult; that is, the revelation of the complex... graphic excellence is that which gives the reader the greatest number of ideas in the least amount of time with the least amount of ink in the least amount of space" (Tufte 2001, 16-18). This author has also coined some terms widely used in the discipline; i.e., *chartjunk* to refer to the inclusion of useless, uninformative or obstructive elements in the information of reports; also the concept of the *ink-data ratio*, which refers to the abuse of ornamentation in the graphic representations of quantitative information. He stated that "[...] in some cases decoration can help to make editorials about the core of the graphic. But it is a serious mistake to abuse data measures to make an editorial comment or to fit a decorative scheme" (*Ibidem*, 59). This is not new: in his 1954 book "How to Lie with Statistics", Darrell Huff already highlighted the importance that common readers learn to correctly interpret both what certain published statistics appear to be saying, as well as what they really say, in order to distinguish and discard those being incomplete or manipulated. The author has since clearly highlighted the frequent techniques of abuse of statistics and their visualization methods –especially in the mass media– in order to manipulate, bias, distort, minimize or sensationalize certain data. His text became a classic and is still valid after six decades, and continues to be required reading in current statistics courses (Huff 1954).

Tufte recaptures these concepts now applied to contemporary resources: the World Wide Web, Power-Point, Big Data, and so on. Basically, both authors stress on the one hand the fact that everybody must always critically review with a discerning mind presentations and conclusions of others drawn from data or visualizations, as they may be affected by biases, errors or omissions of the issuer, as well as sensationalist intentions, either deliberately or by mistake. On the other hand, they highlight the importance of issuing results, tables, graphs, etc., with the highest accuracy and quality. Therefore, the proper selection of tools in

this regard is essential. As with any tool, not knowing how and when to use it properly makes it useless, and in some cases, even dangerous. Therefore, prior acquaintance, training and proficiency of this type of tools is essential in libraries.

The first element influencing this is the proper selection of data to be visualized. At first glance, this may seem obvious, but in practice it is not. For more than half a century there has been a principle in computer science stating: "Garbage In, Garbage Out".[32] This refers to the fact that in this field, if the input data to a computer system is faulty, biased or meaningless, the result will always be something faulty with little or no utility. Despite time the principle is still fully valid. The selection of data to be displayed must be informative, constructive, interesting, challenging, yet it must also appear attractive. Libraries have been compiling and presenting their usage data for over a century, which is indeed convenient, but with the current problem that the vast majority of times the data displayed has not changed; it remains the same. Typically, they report how many users they served in a certain time period –month, semester, year–; how many queries they answered, how many documentary resources they have, how many books or journals were consulted, and other similar data. That is fine; they are still necessary. The problem is that for the present times, the data offered is too basic and elementary. By simple using the vast amount of data captured by current library automation systems in a first instance, statistics could be much richer and more varied, not even getting into big data.

Phetteplace (2012, 95) presented an overview of how to approach this issue in libraries. He suggests that –starting from brainstorming– new data collection strategies should be considered. Based on the interactions between users and collections, new questions can be posted about reporting: Which cross-referencing of dissimilar data can be done in a new, useful, and engaging way?

---

32 It is known as the "GIGO principle". It was first reported by the Free On-Line Dictionary of Computing – FOLDOC, and it is attributed to Wilf Hey in 1965.

How much additional information can be extracted from catalogs and items in the library? Can data analysis of websites of interest to the library could be done? If usable data cannot be found in the current compilation strategy, then it is worth reconsidering why abstract facts are being accumulated. This author adds that once the data have been compiled, a good understanding of the basic types of visualizations and displays allows the most appropriate ones to be selected: typical linear graphs for temporal data, maps for geographic data, bar or pie charts for simple comparisons, tables for data sets, and so on. But nowadays there are many more options, each suited to particular forms of data: tree maps, area maps, mind maps, Venn diagrams, bubble charts, graphs for Linked Open Data, to name a few. Obviously, there are many software tools to produce each of those visualizations, so it is necessary to extend the expertise towards those tools to be able to make an adequate selection of the visualization for each type of data. Hence the great importance of this knowledge in libraries. On the one hand, to be able to establish which new data sets it is convenient to collect; and on the other, how to present them in an attractive and interesting way to library patrons and staff. Most authors –as well as Phetteplace and Tufte– agree that data visualization should be designed to illustrate and not to obfuscate. If a certain design adds more complexity, it counteracts the very purpose of representation. As more layers are added and additional dimensions are revealed, the structure and message of the data should become easier to interpret. Ben Shneiderman summed all this up splendidly: *"*[...] the purpose of visualization is insight, not pictures".

Among the main software tools in this area are Tableau Public, IBM's Many Eyes, and Google Data Studio, which are general-purpose tools for easily creating multiple types of visualizations from data, such as charts, graphs, maps, etc. They are as easy to generate and use as Excel charts, but with much greater capabilities.

Artificial Intelligence tools

> *Nobody phrases it this way, but I think that artificial intelligence is almost a humanities discipline. It's really an attempt to understand human intelligence and human cognition.*
>
> Sebastian Thrun

Generally speaking, many of the above categories are software tools containing a good share of Artificial Intelligence: text analysis for the creation of indexes and catalogs, translation, expert systems, etc. All of them contain many of its components, so it is impossible to speak of big data tools "with" or "without" AI; practically all of them include some element of it to a greater or lesser degree. For the same reason, many libraries are already developing or using some service or product which incorporates such component, although this often goes unnoticed since the tool or service does not have the name Artificial Intelligence (AI) explicitly in it. If a library is using an expanded catalog beyond a simple OPAC, or collecting data from users in social networks, or optimizing the query system, or using some expert system for a task, they are most likely using some AI component in the software operating them. There are many products and services in the library that already contain some of this technology; in short, not necessarily the whole of the software is AI-based, but some parts of it are. For this reason, they are not offered in the market under the specific name of "AI tools", but they are in there. It is therefore necessary for the librarians to be aware of these developments, in order to be able to recognize these parts of such technology and use them to their advantage. Discarding the stereotypical image of a robot walking through the library, in real life there are already countless applications of Artificial Intelligence in a wide variety of uses within the library, many of them used on a daily basis in numerous libraries worldwide.

However, there are indeed some tools offered expressly associated with the name Artificial Intelligence, so a brief review of them is appropriate. These software tools can be found both in

the form of bundles, i.e., broad sets of AI-based applications and services contained in a single package offered by a vendor, and also in the form of AI punctual applications which serve to solve a single specific purpose. As with the previous categories, they can be found in open access or commercial products. Obviously, they exist in all possible variants of AI: expert systems, text analysis, speech analysis, robotic assistants, and so on. Many libraries worldwide have built specific applications for some of their services based on AI, which serve as an example to think about specific developments of this type.

Commercial bundles include Savannah – https://www.orangeboyinc.com/benefits-and-features/, a specialized intelligent platform for libraries offering in one package data warehousing, patron segmentation, distributed patron communications, performance reporting, NPS feedback, and GIS mapping capabilities.[33] Many libraries combine their users' databases with GIS mapping to render their geographic data, and send targeted and selective information to patrons accordingly. Similar to the above is Patron Point, an automation platform with features to provide libraries with enhanced digital communication to their community. Basically, it is a SDI – Selective Dissemination of Information service with aggregated Artificial Intelligence. It combines patron data with library information systems and third-party services: catalogs, electronic resources, databases, event registration systems and many other elements, segmenting patrons in detail to automate communications to them in a much more targeted and personalized way.

As mentioned, some punctual applications are built to solve a certain specific problem with the help of AI techniques. Among

33 The Net Promoter Score (NPS) survey consists of a single question measuring the extent to which a customer or user is likely to recommend an institution or service to others. It is an indicator of user experience, satisfaction and eventual future loyalty. A Geographic Information System (GIS) is a technique for the collection, management and analysis of geographically arranged data. It analyzes spatial location and organizes layers of information into visualizations using maps and 3D scenes.

these we can mention as a representative example Collection HQ – https://www.collectionhq.com/, a tool developed especially for public libraries to analyze the usage data of their collections in order to improve their utilization and performance. It has AI components which contribute to optimize the acquisition of materials for collections, their use, management and promotion.

Another representative example of these specific AI tools is Gale Analytics – https://www.gale.com/databases/gale-analytics, which integrates demographic components to know in detail the community of a certain library, optimize the use and promotion of its materials, as well as extract information from the Integrated Library Management System – ILS, to add value to existing data, optimizing its scope, promotion and use. It allows to attract new users, create and promote new library services, and allocate resources more efficiently.

Some of these tools have been created by libraries. Among the expert systems developed by libraries is Plexus: an expert system developed by the British Library and the University of London for public libraries. It consists of a tool which performs certain reference tasks typical of the reference librarian: it obtains the description of a patron's information need and –if necessary– complements it by inferring additional concepts or asking the patron to answer some clarifying and delimiting questions. The system then develops a search strategy to be applied to the library's holdings and/or databases or to other related reference sources.

A significant number of the applications and tools in this area developed by libraries deal with Learning Analytics. Although this topic is of general interest to academic institutions, because of its proximity to libraries, it is often developed within or associated with the latter. This concept is defined as "[...] the measurement, collection, analysis and reporting of data about students and their contexts in order to understand and optimize learning as well as the environments in which it occurs". This definition comes from the First International Conference on Learning and Knowledge Analytics (2011).

Ten further conferences on the subject have been held since then and –since the subject has attracted a great deal of interest in recent years– a dedicated Learning Analytics Research Society was created. The New Media Consortium (2013, 5) defines the matter as:

> Learning Analytics is the field associated with determining trends and patterns in big data in education, or large data sets relating to students, to further advance the personalized higher education support system.

This specific type of data analysis is used for:

- Prediction; for example, to identify students "at risk" of dropping out or failing out of school; or conversely, to detect students with above-average skills or potential;
- Personalization and customization, to provide students with tailored methods, channels, learning tools and even assessment materials;
- Coaching, to provide teachers with relevant information to mentor and support students;
- Feedback, to evaluate the interest and satisfaction of courses, educational materials, information services, instructional techniques and modalities, etc.;
- Visualization of information, usually in the form of "learning dashboards", which provide general learning data through data visualization tools.

There is an additional variant of Learning Analytics, the Academic Analysis, which also uses data science and AI. Baepler & Murdoch (2010, 3) define it as "[...] an area that combines selected institutional data, statistical analysis and predictive modeling to create intelligence upon which students, instructors or administrators can influence and change academic behavior".

An applied example of these concepts can be seen at Washburn University Library in Kansas. Young (2017) mentions that a project was developed there to establish the extent to which there

was a correlation between students consuming more library materials and their academic success. Over several years, they obtained thorough data on how library use compares to other metrics of academic success. As a result, it was quantitatively demonstrated that indeed increased use of the library and its resources had a significant impact on that success. In consequence, the university made certain changes: moved the tutoring department and the writing lab to the library; these changes were designed both to attract more students to it as well as to make tutoring more acceptable and efficient for students. From these library modifications, retention increased 12% at that university. Derived from this experience, many university library managers –such as those at Georgia, USA– have developed other Learning Analytics projects to measure the extent to which students who use more of the library's information resources tend to achieve greater academic success. Since the results are typically favorable but at the same time quantitative, they obviously have a significant bearing on the presence, importance and budget of libraries within their universities. It had always been intuited and asserted that students who use the library to a greater extent are more academically successful, but until the use of big data and AI in the library it had not been possible to demonstrate this quantitatively.

The National Autonomous University of Mexico –UNAM– has already initiated some projects around the principles and techniques of academic analysis, precisely for the purpose of studying issues that are extremely difficult to address with traditional data techniques, in order to improve learning and the terminal efficiency of its students. For example, the development of the AppUNAM project, which aims to collect relevant data from that community through a mobile App. This project is part of a global effort of several universities called Student Retention Workflow (Salazar 2020, 96). Also at UNAM, in the Coordination of Open University, Educational Innovation and Distance Education (CUAIEED), a special Coordination of Artificial Intelligence, Machine Learning and Learning Analytics has been created within its organizational structure, in which some projects on this topic have already been

initiated. In Latin America, the Revista Iberoamericana de Educación / Educaçao meaning Ibero American Journal of Education, dedicated a whole issue to the topic of learning analytics under a particular focus of this region (Revista Iberoamericana... 2019).

Google's cloud offers access to super processors called TPU or Tensor Process Unit –used in its search engine, translator, photo identifier, Google Assistant and Gmail– in aim of enabling customers to develop AI applications on extended-capacity computers. Amazon Web Services and Microsoft Azure also provide access to super-processors called GPUs or Graphics Processing Unit to develop such applications.

In Artificial Intelligence, there are also numerous software tools that are not built as a plug and play solution, but rather as a set of programming languages, routines, and libraries performing specific purposes and whose parts are assembled as a scale model to obtain the solution to a certain specific need. This type of tools is known as AI frameworks. The following are representative examples:[34]

- Python – https://www.python.org/, language and libraries with numerous AI elements used for countless applications; it has been used to build many of the academic and data repositories seen today;
- Amazon Machine Learning – https://docs.aws.amazon.com/machine-learning/latest/dg/what-is-amazon-machine-learning.html, a broad platform for automated learning used by thousands of institutions around the world, which contains a wide variety of AI solutions ready to assemble and build interesting applications from them;

34 Python official site https://www.python.org/ Python tutorial https://docs.python.org/3/tutorial/index.html
Amazon Machine Learning https://docs.aws.amazon.com/machine-learning/latest/dg/what-is-amazon-machine-learning.html
Microsoft Cognitive Toolkit https://docs.microsoft.com/en-us/cognitive-toolkit/
Accord Framework Net http://accord-framework.net/

- Microsoft Cognitive Toolkit – https://docs.microsoft.com/en-us/cognitive-toolkit/, a platform similar to the previous one, with the advantage of being an open access tool with numerous routines and elements for AI;
- Accord – http://accord-framework.net/, a machine learning platform with additional libraries for audio and image processing. It is an AI framework for building visualization applications, statistics processing, predictive analytics, etc.

The above is just a small representative sample of all that is available on the subject of AI applied to libraries. As can be seen, AI applications are actually very varied and go far beyond the library assistant robot. Like other aspects of ICT, they have been in libraries for quite some time now, supporting a wide range of tasks, and the best part is that the possibilities ahead are even broader. Many of them are simply ignored by librarians or used by them without the awareness that they actually are AI applications. In this regard, Wheatley & Hervieux (2019) conducted a study to measure the presence of AI developments in universities in USA and Canada, especially within their libraries, and concluded that librarians in those institutions were not as aware of the topic or as attracted to it compared to other fields of technological application, contrary to what one might think in such geographic locations.

This may be due to the natural distrust that the topic of Artificial Intelligence has always aroused in many professionals –in all fields– as a threat to their work, which causes them not to delve deeper into it. This is not new: many of the aspects of ICT that for decades have been totally common in libraries were viewed with skepticism and distrust for a long time, such as the production of catalog cards, online catalogs, or electronic indexes, to cite a few examples. This is natural, but the truth is that in spite of this and as time went by, they became an integral part of the every day's library tools, and in the end their managers accepted their usefulness and benefits, and took advantage of them for what they are: technological tools to support librarians' work.

The same is true for those of AI: beyond myths and hypes they are useful tools to assist library tasks. As early as 1995 it was stated:

> No computer program can completely emulate a librarian's knowledge gained through a solid theoretical understanding of information processes and real-world experience. But algorithms built upon years of experience can answer many questions, thus freeing the library staff to do more unique time –and labor intensive– advanced tasks (Expert Systems… 2014).

As can be concluded, this is still fully valid after more than twenty five years.

THE SPECIALTY OF DATA ANALYTICS

> *Today and in the future, companies will have more data than they can imagine and will have the means to capture and manage it. What is more necessary than ever is the capability to analyze the right data in a timely enough fashion to make decisions and take actions.*
>
> JUDITH HURWITS,
> "The Big Data Paradox"

As has been seen throughout this chapter, there are countless software tools used in the field of data, especially big data; there are tools for all types of purposes, organizations, projects, and budgets. However, in recent times there has been a special emphasis on the field and on tools for data analytics. An overview of tools for this purpose has already been given in Sections 8.4 and 8.5. But it is necessary to emphasize the fact that this is a whole specialty in itself.

Generally speaking, it embodies the core of data science; i.e., detecting trends and patterns in data in order to propose solutions; but more specifically, it refers in particular to the task of identifying

which variables of the organization can be related to certain data and thus establishing correlations for the posing of questions and the eventual obtaining of answers through specific techniques. These techniques certainly form the core of data analytics as a specialty. Several authors have emphasized the importance of the design method and procedure for data analyzing. Basically, those studies establish that different people analyzing the same event, under the same conditions, performing different procedures, may find the same value with respect to a statistical parameter, but the evaluation on the significance of the data obtained and the actions derived from it may be different depending on the analysis procedure used. This is called the "data paradox" and highlights the importance of its correct analysis (Berdondini 2019).

Data analytics should be studied with caution and understood in its full extent. In recent times, countless websites, texts, articles, software tools, tutorials, undergraduate and graduate courses, consulting, products, certifications, etc., have been offered by a myriad of organizations and companies on the subject of data analytics. Obviously the core of all data science and management is to become able to extract valuable and useful information from data sets, since there is no point in collecting and processing them if no tangible benefit can be obtained for the organization, and certainly this becomes a reality only when they are analyzed. "Value" has already been mentioned as the fifth essential characteristic of big data from the extended theory of the three "V's", and the fact that multiple authors consider it to be the most valuable of them all has been remarked. This is logical, but not absolute.

There is no doubt that data analytics is the core and most important part of the data science and management, especially in big data, since this is where the relevant solutions and decisions for an organization are extracted; therefore it deserves special attention and study, and as a logical consequence there are more tools, texts, and products in this regard compared to other areas of data management. But it is essential and unavoidable to contextualize this field along with the others: not doing so jeopardizes the ultimate outcome and purpose of data management. From browsing

at websites, product offers, training and consulting, courses and diplomas, certifications, and so on, it would seem at first glance that this is all that needs to be addressed with respect to data science and management: data analytics is everything. If the organization solves the data analysis issue, the whole problem is solved.

This is the result of the aforementioned serious global lack of experts in data science and management. There is obviously an urgent need to educate and train personnel in this area, but it is essential to educate and train them in all areas of data science and management. It is not possible to develop only one of its fields, no matter how essential it may appear, as this introduces a harmful and dangerous unbalance for the final purpose. Every field of knowledge has core areas deserving special study and training, but it also has complementary areas without which the core field loses its value and integrity. All human knowledge domains have a dialectical part with theories, bases, principles, fundamentals, concepts, methodologies, regulations, as well as a practical part with procedures, experiences, studies, recommendations, manuals, standards, techniques, and obviously, tools in their respect.

The field of data is no exception: it has a theoretical part –data science– and a practical part –data management–, both with all their subfields and related disciplines. As in all other areas of knowledge, no one can know everything. Therefore, some people become experts in one or more of the subfields of theory and others in one or more of the subfields of practice; a few of them get to acquire some experience in both. People specialize according to their studies and experience, given the impossibility of knowing everything. As in all applied disciplines, there are logically more people dedicated to practice than to theory, and there will also be a greater number of people dedicated to the most relevant subfields having the greatest market demand. Consequently, it sounds logical that in data management –and especially in big data– there should be more people dedicated to data analytics, as it is one of the most important areas. But it is by no means desirable for the discipline or for organizations that everyone be trained in that subfield. Obviously every organization must have people who know how to analyze data and

extract solutions from it, but it must also have people who know how to design, collect, normalize, audit, store, access, and protect it. If the organization has only one expert dedicated to data, this person must know about all of this in a balanced and contextualized way. There is no use in having an expert in data analytics if all other areas have been neglected. If an organization has a section or area dedicated to data, obviously each of the people therein must be trained in one or some of the areas of data management and together they must cover all or at least most aspects of it.

The whole knowledge of the personnel of a certain organization comes on the one hand from professional education, and on the other from training, continuing education and practical instruction. It cannot be stated absolutely that one of the two aspects is more important than the other, having both of them a preponderant value and usefulness in each context. Some organizations find it useful to have professional staff for certain roles; others find it useful to have trained and experienced staff in their practical tasks; still others find it useful to have both types of staff. In all cases, the widest variety of specialties should be covered according to the needs. The central point of the above is that no organization is well served by having only data analytics experts if it has no one who can contend with the other aspects of its management. Likewise with individuals; it is not desirable for a person to be trained only in data analytics; if there are no other experts in the other areas within the organization, this knowledge and skills will be sterile for successful data management, as there is no context and balance of these within the institution.

Data management fields are like the legs of a table: all of them are required to stand firm, give support and be functional. If one of them is removed, the whole thing may appear to hold up, but under any load or use it will collapse. If two are omitted it will be very difficult to even pretend that it works, and if three are omitted it cannot even be claimed that the remainder is a table. Data analytics is merely one leg of the table: without the other aspects it simply does not work. Therein lies the importance of the context and the balance of the whole. Lorica stated in this regard: "

> Judging from articles in popular press, the term 'data scientist' has increasingly come to refer to someone who specializes in data analysis: statistics, machine learning, etc... Far from being confined to data analysis, a typical data science workflow means jumping back and forth into a series of interdependent tasks. Data scientists tend to use a wide variety of tools..... Workflows that involve many different tools require a lot of context switching… (2014, 4-5).

As a conclusion to all of the above, it can be stated that organizations and individuals should consider in deed data analytics as one of the many training and coaching options, but never as the only option to be developed at corporate or personal level. The whole must always be weighed and with special caution *with whom* the training will be obtained. It must be remembered that today there exists a huge business model around data management, especially in training, consulting, and certifications. This has become a million-dollar business for many companies offering it; obviously this is valid to a large extent, but on the other hand, an enormous expectation has been developed around the most attractive field of data analytics, where much of it does not address the general context of what organizations or people really need, and where not all the providers are serious and genuine. Many of them sell because it is trendy and good business even if what they offer is not of quality or really useful to the customer. That is why it is so important to weigh the *whole* and *with whom*.

The central point of all of the above that should be kept in mind is that in order to solve a data problem, data analytics is not enough; all parts of data management must interact together in a comprehensive and balanced manner. It should be noted at this point that data process involves many steps; although not all of them are found in each of the related problems, in fact, several stages are required always for its proper management: design, modeling and data typing; data capture, collection or mining; coding and cleansing; normalization and structuring; inclusion of metadata; data transformation, aggregation or disaggregation; data validation; storage, governance and preservation; data

visualization and display; and, finally, analysis and interpretation. It can be concluded from this, that an organization requires more specialized personnel than just a data analyst; and furthermore, that a good data analyst not only knows and can analyze data, but must also be able to participate integrally in all the other stages.

Once the above reflection has been made, we can enter into the matter of what formally data analytics as a specialty implies: it is a subset of data management which refers to the task of identifying those variables of the organization that can be related to certain data and thus establish correlations for the posing of questions and the eventual extraction of answers through specific techniques and tools with the final purpose of proposing decisions and lines of action. Authors divide data analytics into four main areas or subdivisions:

a) *Prescriptive Analytics*: a type of analysis of raw data; specifically, it looks for different possible scenarios of the organization, derived from its available resources, past and current behavior, recommendations and past actions, and hence to suggest a strategy in the long term or a course of action in the short term (Segal 2014);
b) *Diagnostic Analytics*: consists of a type of analysis that digs deep into past data and facts trying to explain the causes and effects of certain occurred events; it is a further step to discover the reasoning behind certain results. It uses techniques such as data discovery, data dredging and mining, correlations, etc., to determine the best sources to solve a certain problem (SISENSE s.d.);
c) *Predictive Analytics*: similar to the previous one, but based on studying recent data and facts to try to predict certain events, trends, or behaviors. It uses advanced analysis techniques which take advantage of data to discover facts and situations in real time and predict future events. It is a new mix of classical statistical analysis supported by Artificial Intelligence tools (IBM s.d.);

d) *Descriptive Analytics*: analyzes the organization's existing databases attempting to describe and explain a certain state of the art or pattern of a particular set of data within it. It aims to provide descriptions or summaries of facts and figures in understandable formats, both to inform and to prepare the data for further analysis. It is based on data mining and aggregation as well as visualization techniques and tools (Dataversity 2017).

Some producers of tools and services for these tasks pretend to differentiate between simple "data analysis" and "data analytics", pointing out that the latter is a much more sophisticated version of the former, and that the differences are radical. Obviously, their products seek to present themselves as more complete and advanced versions. In reality, it is clear from reading the texts on the subject that it is only a semantic difference artificially introduced for marketing purposes, and that they are not two different concepts: they are simply two ways of referring to the same thing, emphasizing certain elements.

There are in fact specific techniques and tools for the analysis of big data, which emphasize the characteristics of such type of data defined by the "V's": volume, velocity, variety, veracity and value. These characteristics mean that statistical methods which usually work well for low-volume data do not scale well to big data. Similarly, many computational techniques which work well for moderate data volumes will struggle with significant problems in big data analysis. All of this introduces new challenges: big data storage and analysis; knowledge discovery; computational complexities; data scalability, data visualization, and information security. It is therefore very convenient to keep in mind that there are indeed different considerations of data analytics for smaller data sets with respect to big data assemblies. The latter entails differentiated or appropriate techniques and tools, such as complex statistical analysis, machine learning, extended data mining, intelligent analytics, well-governed data, cloud computing, quantum computing, and dataflow programming (Acharjya & Ahmed 2016, 512,517).

In recent times a new variant of data analysis has been developing, called "deep analytics", which consists of processes of applying data mining combined with other data processing techniques to analyze, extract and organize extremely massive amounts of data in an acceptable, useful and beneficial way for an organization in order to discover new uses and/or applications of certain data. This type of analysis is usually performed by extracting selected data from huge data sets –Petabytes or even Exabytes– distributed across complex architectures and dispersed in multiple data warehouses; therefore it requires very specialized and powerful resources to perform this process: distributed computing on multiple servers or computing nodes, cloud computing, distributed search, deep metadata analysis, and so on: deep analytics is often combined with AI applications, machine learning, data mining, complex correlations, etc., in order to achieve its purpose, and extract useful and specialized information from those extremely massive data sets.

Entering the field of libraries, it is worth to make a review about the issues and uses of data analytics specifically within these organizations. Several authors, such as Showers *et al.* (2015) already pointed out some of its practical applications: developing collections; designing new services; exploring patrons' knowledge, their preferences and needs; quantitative evidence of the value and impact of the library within its environment; improving patrons' educational and informational experience; the aforementioned "learning analytics"; and in general, to improve decision making in libraries. Data analytics, both massive and smaller scale, can certainly be used for this purpose. But each of these entails a perfect understanding of the real meaning of data analytics, so as to avoid making mistakes or inaccuracies.

As mentioned at the beginning of this section, the real core of data analysis refers precisely to the task of identifying which variables in the organization can be related to certain data and thus establishing correlations for the posing of questions and the eventual obtaining of answers through specific techniques. This means that before analyzing data, the library must be sure to select those

data that in effect represent certain facts or realities, in order to establish and obtain the correct correlations leading to real and relevant questions and answers. One of the most frequently made mistakes is to automatically use the "traditional" data that the library already collects for its statistics; the popular adage says that "he who only knows how to use a hammer sees a nail in everything". Another common mistake is to collect the data that are most easily attainable, simply for that reason: because they are easy to obtain. Data analytics expressly indicates that the selection of data must initially have a close correlation with the facts to be represented; otherwise the results will be biased or irrelevant. This may sound obvious, but it is not easy; it requires, on the one hand, a deep knowledge of the causes and effects of the library function or task to be represented, and on the other hand, a broad mastery of how to approach and design specific data correlations. This involves two specialties: the librarian with expertise in such library function or task as well as the data analyst. Ideally, they should be the same person; i.e., a librarian with expertise in such function, having at the same time good knowledge of data analysis; however, a duo of two experts in close synchrony will also accomplish the job well. Without this dual function properly applied, it is easy to fall into obtaining and using data that is too simple for the task, or the opposite: to try to obtain immense amounts of data, most of which is irrelevant or unrepresentative of the function, which in turn makes the task more difficult. From the very outset, it must be planned exactly what is to be measured, why and how. Hence the importance of having a duo of experts, or one expert with two specialties.

Areas of the library where such analyses can be applied are numerous, as has already been mentioned; the eventual points of data extraction and collection are even more so: catalog and collection searches, reference, book loan, document views and downloads, access to certain services, social networks between the library and its patrons, to name a few. Among so many possibilities, it is easy to get lost in the sea of data to be obtained and therefore it is imperative from the beginning to make a correct

design of them and their associated facts to start from there towards a relevant analysis. However, the effort is worth it; Farmer & Safer (2016, 6) put it this way:

> [...] the benefits of quantitative data analysis for the improvement of libraries leads into a better return on investment... Data analysis generally results in the maximum use of resources, since performance and productivity improve, and quality is controlled, thus not only reducing costs, but also increasing user and staff satisfaction.

This is one of a series of books that ALA – The American Library Association, has published about data analytics in libraries, which describes out many of the knowledge and skills that librarians should acquire in this regard: basics of statistical concepts; knowledge of recommended data sources for various library functions and processes, as well as guidance on how to use them; techniques for data cleaning, as well as for finding data correlations with appropriate data analysis methods; and finally, how to visualize the results. Aware of the importance of data analysis for library improvement, ALA has been producing a whole series of texts on the subject, covering together a number of specific topics; i.e., Farney (2018) discusses the extraction of patrons' data from their web browsing activities, library catalogs, information discoverers, and repositories. In this text, the author explains how to establish adequate and consistent data sets from this activity, analyzing and identifying areas that fit the library's priorities; how to clean the data and combine various sources of it; and how to perform the corresponding analyses to subsequently apply the results in improving collections and patron satisfaction. From the same ALA series, and in conjunction with the Association of College and Research Libraries – ACRL and the Public Library Association –PLA, Hernon *et al.* (2015) explain how to extract meaningful and consistent data for libraries management, and hope to improve its accountability through better use of collections, benchmarking, and other best practices.

As it can be seen from the above texts –which are only a sample– there is already a strong interest of library associations in studying, disseminating and training on the subject of data analysis for professionals in this discipline. In fact, in many of the texts on the subject it can be seen that some authors are already beginning to refer to "data analysis" applied in libraries more specifically as "library analysis", to highlight the specific meaning of this data activity in this environment and to make evident that it is already a specialty within libraries.

Big Data Governance

Quality is never an accident; it is always the result of, intelligent effort.
JOHN RUSKIN

Data is becoming increasingly important in a large number of businesses and organizations, of all types and sizes. By its nature of volume, speed of generation and variety, it must be properly managed, otherwise data can cause more problems than benefits. Hence the importance of data governance, especially for big data. There are numerous definitions of this concept, emphasizing one or another aspect of it; for example, the "Data Management Association" or DAMA defines data governance as: "[...] the planning, monitoring and control of data management and the use of data and related sources". (Techopedia 2020) defines this concept as: "[...] an overall management of the quality, usability, availability, security and consistency of an organization's data". The Data Governance Institute's definition about data governance has already been cited above. As a result of many of these definitions and in summary, it can be established that data governance aims to ensure the quality and security of the data used in an organization by establishing and monitoring a set of policies, processes, positions, standards, metrics and responsibilities which ensure the effective, efficient, secure and consistent use of data.

It is important to clearly set out the difference between data governance and data management. Olavsrud (2020) differentiates between the two concepts: the former is part of the latter; data governance –although a very important part– is only a subset of the more general concept of data management. Data governance deals with the roles, processes, standards, metrics, and personal responsibilities for establishing clear ownership and access to data assets within the organization; their protected use, consistency, and accountability. Data management is a much broader term which describes all the processes used to plan, specify, enable, create, acquire, maintain, use, store, access, control, cleanse and govern data.

Data governance creates a framework for data within an organization; data management entails the practical execution of such framework. Essentially, data governance seeks to ensure that –where data is concerned– the right people have the right responsibilities through relevant processes, standards and metrics.

By its very nature, data governance leads to benefits that are generally not achievable through data management alone:

- Provides a generalized and coherent vision and understanding, as well as a common terminology for data within the organization;
- Enables optimization of processes, procedures, roles and positions within the organization;
- Improves data quality, and ensures data accuracy, completeness and consistency;
- Optimizes the location of all data related to key areas of the organization;
- Defines clearly all persons handling data within the organization and their accountability;
- Enables full compliance with the minimum legal requirements, especially those related to privacy and personal data protection;
- Prevents conflicts and/or overlaps between different data sets and sections of the organization;
- Saves operating costs.

Essentially, data governance is designed by developing an organizational policy created specifically for this purpose. While this sounds desirable from the outset, as organizations capture more and more data, data governance will become increasingly imperative: simple data management will no longer suffice. For its design and implementation, there are already several lists or statements of basic principles. Techopedia (2020) lists six of them:

1. *Recognize data as an asset:* From the outset, the importance and value of data as assets for an organization must be recognized. This leads to actions to define, control and access them in a careful and process-oriented manner. Consequently, management can rely on the accuracy and usefulness of the data.
2. *Establish data ownership and accountability:* Clearly define within the organization who collects, accesses, processes and stores what data, and this should only be done under defined and authorized processes. Data governance should be jointly shared by all departments and not just IT section.
3. *Adopt standardized policies, rules and regulations for security:* Any data management process must follow recognized and standardized regulations and processes regarding the use of sensitive data, access, privacy policies and security norms. Such processes must be strictly enforced in order to avoid data problems.
4. *Manage data quality thoroughly and systematically:* Data quality must be consistently defined, created and maintained from the outset. The organization's data should be periodically cross-checked against pre-established quality standards.
5. *Manage change:* The data governance process should track over time changes in the data, the consequences derived, and the change management activities within the organization in a proactive manner.

6. *Audit the data:* The data governance process should support a transparent audit policy in any organization. Therefore, data audits should be a standard and ongoing process over the years. The basic principles of data governance should be kept simple and understandable to all levels of the organization.

Some companies offer data governance software platforms and tools, as well as many other related products and services. A small sample of them are:

- Erwin Data Governance – a SaaS tool, "Software as a Service" in the cloud, which attempts to manage the processes defined as part of data governance, such as process modeling, data modeling, or data library;
- Adobe Experience Platform Data Governance – offers tools to classify, manage, and enforce how data is used across the organization to maximize its benefits;
- Talend Data Fabric – a set of applications to collect, govern, transform and share data between the cloud and the organization's on-premises environment;
- Onna – application operating in the cloud, responsible for centralizing dispersed information in order to preserve both historical and current information in real time;
- ManageEngine ADAudit Plus – software tool designed to build data audits within organizations.

The above list is for illustrative purposes only. All such products should be considered with the utmost caution. It is worth emphasizing at this point that data governance means fundamentally creating a coordinated framework for data within an organization; this is achieved primarily through the construction of a data governance policy from which procedures, processes, etc. are derived: it is definitely not about the use of computer tools. Governance is a theoretical conceptualization that is written, not a computer program that is executed. Obviously, it is convenient to use

computational tools to implement some of its parts, such as data auditing, the creation of master data charts, IT security, and to verify compliance with them; but the use of applications designed to execute or supervise some of its parts should by no means be confused with governance as a whole. As has already been established, governance covers many different aspects of cohesion, quality, security, data accountability, and so on. To date, there is no software application that can universally integrate all these aspects in one place. The tools of this type available on the market can facilitate the task by addressing one or more of its parts, but they are definitely not a substitute for the intellectual process of conceiving and designing a coherent data governance strategy, nor for monitoring it.

To better understand this, we can use the Data Governance Functional Reference Framework, which is a compilation of views developed by the Data Management Association (DAMA s.d., 5), the Data Governance Institute – DGI, and IBM. It points out ten areas of attention, action and development regarding this topic; here there is an overview:

1. *Data architecture:* recognize data as assets with value, as well as the resources related to them as an integral part of the organization; establishing the overall data structure for it;
2. *Data modeling and design*: analyze, design, build, test and maintain data;
3. *Data storage and operations:* manage the structures and devices for data storage and access;
4. *Data security:* establish measures to ensure privacy, confidentiality and adequate access;
5. *Data integration and interoperability:* design, collect or extract, transform, transfer, replicate, and map data; seeking integrated and interoperable data across processes and operations;
6. *Documents and content:* transform, index and optimize access to data found in unstructured sources for integration and interoperability with structured data;

7. *Reference and master data:* define shared and standardized data between areas of the organization to reduce redundancies and ensure better data quality; create master data sets;
8. *Business intelligence:* generalize data processing and analysis to optimize decision making throughout the organization;
9. *Metadata:* design, structure, collect, classify, maintain, integrate, control, manage and deliver appropriate metadata;
10. *Data quality:* define, monitor, and oversee data consistency and integrity to improve data quality.

This fairly exhaustive list helps to better understand the above statement. It is clear from its analysis that the fundamental points of data governance are all concepts to be defined and established, from which actions and processes will follow. Not a single item in the list emphasizes the importance of acquiring and using a good IT tool. Obviously, no project managing large volumes of data will be able to do so without adequate resources in that regard, and therefore data management does indeed require adequate IT tools. But again, data governance is not data management, and in governance –unlike management– tools are not a key point. Likewise, in all the products offered on the market for this purpose, it is observed that there is none covering all the points of attention listed in the framework for proper data governance.

As with many other aspects of management in organizations, data requires a part of its stewardship to specify the fundamental details to ensure its quality, cohesion and security ensuring the effective, efficient, secure and cost-effective use. Such a part of management is undoubtedly data governance, and as in any organization, it is advisable to include it within libraries data.

Librarians and Big Data

Librarians can be expert guides along the Big Data super-highway.
AMY AFFELT,
"The accidental data scientist", 2015

As has already been established, data management –especially of big data– is a multidisciplinary discipline, and falls under the purview of a wide range of information professionals; obviously among them are librarians, due to their usual training and experience. Like all specializations, it requires new sets of knowledge, skills, attitudes and experience.

In general terms, current librarians need to know the fundamentals of data science, management and curation; the pertinent sources of information on the subject; bibliomining; how and when big data is used in the library and where it can be found; data analysis and visualization, as well as the selection of suitable software tools for its exploitation, all as part of a comprehensive education and training in their discipline. Globally, there is a severe shortage of data professionals who possess adequate knowledge of how to design, define, collect, cleanse, transform, analyze, and present data projects and structures. Failure to conceptualize any of these steps of data management to build a comprehensive idea of the value, processing and utilization of data will result in organizations engaging in uncertain projects, or remaining static without initiating them because of a lack of technical capabilities.

Qualified personnel, not tools, is the main key to success in these initiatives.

Beyond this comprehensive education and training in the discipline, and depending on the type of library and/or organization in which they work, librarians must acquire and deepen this knowledge to be able to help in their environment in many different fields of application. Here is an overview of these activities recommended by multiple authors, among them Bieraugel (2013), Showers (2015), and Lyon & Mattern (2016):

- Librarians in research institutions must understand how data now relates to scholarly research, and how it can be disseminated and reused;
- Librarians in higher education institutions need to understand how data from faculty and students can be used to improve the educational efficiency of the institution, curricula, career guidance and retention, learning experiences, and so on. They can also train students in the best use and exploitation of data;
- Business librarians in general need to know how they can leverage big data, data mining, data analytics, etc., to bring competitive advantage to their enterprise;
- Librarians in LIS – Library Information Services companies must be prepared for the changing needs and characteristics of those services and their users to be able to take advantage of big data and its analysis in order to design and create new products and services and optimize existing ones;
- Librarians in the social sciences and humanities should be aware that data and data analytics are also increasingly common in their disciplines, and that "digital humanities" has a growing need for these tools and their specialists;
- Librarians engaged in information organization and registration must study and design new taxonomies, metadata schemas, and structures to make large datasets more visible, accessible, and useful;

- Librarians engaged in the systematization of document retrieval methods must design new mechanisms and tools to achieve better information search results and discoverers;
- Librarians engaged in bibliometrics need to design new mechanisms and tools to extract better results from citation, reference, keyword, usage, and full-text analysis;
- Library managers must assimilate how the use of big data can be useful to redesign multiple administrative aspects of the library: acquisition and use of its information resources, budget application and/or re-direction decisions, operation schedules, document selection and discarding, reduction of hidden costs; design, usability, performance and perception of new services;
- Data curation librarians can advise members of their communities on the collection, storage and accessibility of large datasets, repository building, and especially on how to create structured data;
- For all of the above reasons, librarians engaged in teaching must begin to train library professionals who can understand the scope of the above activities within the discipline and acquire the basic knowledge and skills during their professional education;
- In addition to all the traditional disciplines and specialties directly related to data, it is important to emphasize that there are now many others that are becoming increasingly closer to library science, such as digital humanities, digital social sciences, social computing, etc.

As can be seen, each of the above statements opens up a wide range of possibilities for data projects in libraries and related organizations; undoubtedly the field of action in them is becoming extremely broad. Starting from general and common knowledge, each of these eventual applications requires professionals with specific knowledge, skills, attitudes, and experience, as well as mastery of particular tools. Obviously it sounds impossible for a single library to have specialists in each and every one of these

fields, but it should also be clear that nowadays it cannot be wholly lacking in some specialized staff. Therefore the library can and should be training its data specialists in the fields of application of its own interest and context. In the first place because this allows it to enter into dimensions in compliance with the current needs and circumstances of the information world, in order to remain competitive and interesting for its communities and funders. Secondly, because it allows the training of professionals who are in high demand at the present time and of whom there is a worldwide shortage, and this means new and better jobs for current professional librarians and library science students.

Lyon, Mattern, Acker, & Langmead (2015), and Lyon & Mattern (2016), conducted two studies in the field of data science in which they identified a set of six main "roles" or functions in this science based on real-world needs for actual jobs. Within the six roles are the "traditional" ones: data analyst, data engineer, and data journalist; but they also found three additional "roles" closely related to the discipline: data librarian, data archivist, and data manager/curator. They also stated that these information professionals should develop five aspects in regards to this topic (*Ibidem*, 3):

1) Education: academic qualifications and background
2) Direct Experience
3) Knowledge: understanding of/familiarity with topics/subjects/issues
4) Performing and practical skills
5) Competencies - Proficiency of tools and technologies.

As mentioned throughout this text, its intention has been to illustrate on knowledge, familiarity and understanding of the topics and partly on skills and competencies.

In addition to the vast range of courses, specialties, certifications, etc., offered commercially by countless companies on the subject of data, there is also a small but acceptable number of courses, texts and training on the subject offered on a non-profit basis by various organizations, which are ideal for librarians

to get started in this field. As already mentioned, great caution should be exercised in selecting them, as many companies and organizations offer supposedly free courses as a "hook" to attract prospective students, but upon closer examination it becomes evident that there are underlying hidden costs, some of them steep, the vast majority of them derived from the issuance of a course certification. This prudential recommendation includes interesting courses at prestigious universities, such as Harvard or MIT; if certification is not sought, the possibilities of free courses increase substantially.

As with all learning, it is recommended to progress from the general to the particular, and from the basic to the specialized. For example, it is not advisable to start learning data analysis without having general notions of data science and data management, nor is it advisable to start learning Python if the most simple principles of SQL database programming and querying have not been mastered. It is therefore recommended to develop a personal data training and learning plan, with a gradual evolution and deepening.

As examples, Bernard Marr (2020) points to a selection from Forbes Magazine of nine free courses on various topics in data science. A number of short courses and introductory items about data can also be found on the YouTube site –some of them quite good– by searching under the entries "data science", "data management" and "data analysis". The same cautionary recommendation is emphasized here with respect to "hook" courses.

The central point of all of the above is that nowadays it is highly recommended that professional librarians begin to seek some training in data science, data management, data analysis, data uses and applications, and so on. But this need not necessarily involve postgraduate degrees, diploma courses, or long bouts of formal schooling. Obviously, all these are good training options, and if they are affordable, they should undoubtedly be taken advantage of, but not all professionals are able to devote the time and money involved. The alternatives indicated in this section are valid options to start this training without investing so many resources,

at a pace that each person establishes and in the schedule that best suits him/her. It should be noted here once again that in this environment, certifications are good, but not indispensable, and their cost is high. In labor markets for disciplines where demand is equal to or less than supply, certifications are a crucial factor in determining the differences between applicants. But this is not the case: given the universal shortage of data experts, what matters most in the end in this market is to demonstrate in practice being able to perform a requested task and not so much to show a piece of paper that affirms it. Experience is a crucial factor here, so it is desirable for staff interested in the subject to engage to some extent in the projects undertaken by their institutions in this area in order to master the practice.

Summary and Conclusions

"Slowly, we are moving slowly into an era where big data is the starting point, not the end".

PEARL ZHU.
"The Digital Master"

The second half of the 20th century was characterized by an enormous growth in the amount of information produced in the world. Within this phenomenon there is a very significant segment being data; so far this century this sector has grown in proportion even more compared to information. Currently the world produces and consumes an immense amount of data; it has become in itself both a source of information and a new input. It is now added to "traditional" or "finished" information products, i.e., the final result of the analysis and synthesis of certain data by individuals or groups in the form of publications: books, academic journals, newspaper texts, manuals, dissertations, compendia, patents, and so on. By its very nature, data requires specific treatment, which gave rise to a series of theories, principles, modalities, methods, tools and technologies for its treatment and use. Data management was thus created, and with it the subfields or specialties: data engineering, data analysis, data mining, data processing, among others; the sum of all of them shaped data science: the study of organized data to identify those that are important for decision making in the context of a specific problem or a certain business model, in addition to the development of models and algorithms for large-scale problem solving in organizations.

Data analysis for problem solving dates back to time immemorial; basically, when humankind learned to collect and analyze data in a systematized and rigorous way, the modern science took shape, and thanks to this, data use and exploitation grew steadily, as well as the theories, principles, tools, etc., to do so. Throughout the 20th century, most of the theoretical elements for its study and formal treatment were created. In addition, the advent and development of electronic data processing, and especially the ability to store immense amounts of data in those devices during that century laid the groundwork for unprecedented growth in this regard. To that must be added the impressive increase of the global network and telecommunications in the last decade of the past century. All this gave a new and major boost to the production of information in its digital form: billions of pieces of information and data were thus created in this modality, adding to what already existed, exponentially multiplying the amount of accumulated information. In particular, some specific sectors of the Internet have grown enormously in recent years, such as social networks, the "Internet of Things" and personal devices connected to the net, exchanging data with other devices and/or systems automatically. All this has resulted in an even greater production of data, creating an unprecedented flow of it. When to all this is added the fact that a multi-billion dollar industry has been built around data, the result is an incessant and ever increasing production, collection, storage and use of immense amounts in today's world.

Due to the rise of the data phenomenon, numerous authors have pointed it out and studied it in the last decades. Several world-renowned organizations are systematically engaged in the counting and rendering of data production, its economic value and its management. Countless companies have been created for data related products and services. A whole body of theories, principles, methodology, and tools have been created from their study. In the early part of this century, all these factors, needs, technologies, etc., together with the immense volume of data, were in turn shaping a new and more complex phenomenon known as "Big Data". This concept refers to the processing and analysis of

sets of data so large, varied, complex and disparate, produced at such a rapid rate and from so many different sources, that "traditional" information processing equipment, programs and procedures: servers, databases, search engines, algorithms, etc., are not sufficient and therefore require much more powerful, sophisticated and specialized methods, equipment and programs, to compile, analyze and correlate them, all in order to be able to quickly extract patterns, trends and associations from these data, mainly from human behavior and interactions, and from this to be able to make informed decisions which assist organizations, conferring to data an enormous added value.

Today, data clusters are used by multiple businesses and organizations in all types of sectors: banking and finance, communications and transportation, industry and commerce, health, entertainment, government, and education. Within the latter sector are libraries, which are also likely to benefit from this development, especially in universities and research centers, where they have become a new and valuable input and have had to find a place to house and store them properly and systematically. Often, the libraries of these organizations and their staff have been designated for this purpose. Although this may seem at first glance to be just another small activity added to the libraries, it is not: it represents both an enormous challenge and a great opportunity. On the one hand, it implies that library staff must acquire new knowledge and skills for data proper management, and on the other hand, it represents new opportunities to reposition the library within the contemporary responsibilities and tasks of its community. This role has been growing in the last decade in a noticeable way, but it is not a simple and small casual addition: it requires organizational structure and qualified personnel to perform it properly. All the new needs, concepts and solutions derived from this task have risen a new specialty in the world of information in academic institutions, the "Research Data Management" –RDM. It is a whole new responsibility dedicated to the management, deposit and distribution of information within their organizations specifically in the form of data. Prominent authors, i.e., Witt &

Horstmann (2016, 251) have already listed the main tasks required of librarians in this regard: 1) helping researchers understand and solve needs throughout the research data lifecycle; 2) advising on the construction of data and metadata management plans; 3) designing data publishing and curation solutions; 4) creating web guides and tutorials to train researchers and users; 5) hosting and maintaining repositories in their holdings.

Large multi-institutional library organizations throughout the world –IFLA, ALA, ARL, JISC, etc.– have become aware of the growing importance of data within their environment and have created interest and study groups on the subject, with the conclusion that it is essential for libraries to be a proactive part of this phenomenon. Accordingly, the number of publications on data and its relationship with libraries, the number of meetings and papers on the subject, specialized journals, manuals and guides, as well as the number of projects developed by libraries in this regard are palpable evidence of a clearly growing interest and importance about data in this sector.

The custody of research data in libraries is one of the most visible part of this phenomenon, but by no means the only possible use of data in them. They are also generating and using their own data on a wide variety of projects and tasks: for patrons studies, for predictive analysis of collections and services, for the development of new and sophisticated search and retrieval tools, for deep learning studies, in expert systems, text translation, OCR, Artificial Intelligence, and to optimize management, to mention a few topics. Many specific strands can be found within these wide general lines of data use in libraries, and a good number of them go even further: they fall into the field of big data. There are already many examples of this: Initially, in the creation and design of new information taxonomies and metadata schemas. In order to exploit information and data –of any volume– adequate metadata is required; without it, datasets, especially massive ones, are a shapeless bulk with little or no utility. Nowadays, the fastest way to build all kinds of taxonomies about a field of knowledge in a comprehensive way is through the coherent management and analysis

of large amounts of linked element data: vocabularies with definitions in natural language, simple taxonomies –data dictionaries and hierarchies–, thesauri with related terms; ontologies or complete taxonomies: relational models with attributes, constraints, relations; functional requirements models, etc.

In this regard, the underlying conceptual models of RDA, the cataloging standard for the formulation of bibliographic records for libraries, archives, museums, etc., such as the Functional Requirements for Bibliographic Records or FRBR, the Functional Requirements for Authority Data or FRAD, the Functional Requirements for Subject Authority Data or FRSAD, and the PRESS ontology, all endorsed by IFLA and consolidated with their "Library Reference Model", were mentioned as outstanding examples. Reference was also made to very representative examples of large library systems and organizations managing documentary information extracting data from their respective collections with many millions of items to analyze and model from their holdings the interrelationships between people, events, places, etc., contained therein. This type of project is radically modifying the way of building catalogs, search engines and information discoverers as a result of studies which have exposed the dissatisfaction of users in this regard. All these data projects are undoubtedly establishing unprecedented dimensions of great utility and acceptance in the field of libraries. As a result, many libraries are already adding large sets of additional data to their catalogs, thus optimizing their search engines: tables of contents, indexes, glossaries, subject matter, etc., are being associated with the original catalog records, greatly enhancing the search and discovery of information, since the search engine no longer has only a few words of the author, title or subject, but many words contained in the index or glossary of each book. Some libraries specializing in literature extract and aggregate all the characters, places, times, events, consigned in novels, theatrical plays, and similar. Chemistry libraries add formulas, substances, compounds, industrial processes, etc., adapting the idea to the context and characteristics of that discipline. The same can be done in other fields of knowledge.

Big data are also now being found in the library in the metric studies of documentary information, in all its various specialties: bibliometrics, informetrics, librametry, as also in other associated ones: scientometrics, webmetrics, altmetrics, and the emerging archival metrics. Their common factor is the application of mathematical and statistical models and methods to library, bibliographic and archival activities, social networks, research in sciences and humanities, communication and dissemination, among many others, becoming all of them another classic example of applied data mining. And not only in metrics studies: with the use of data mining, linguistic techniques, statistics, machine learning, information retrieval, natural language understanding, case-based reasoning, and others, text analysis studies allow organizations and individuals to obtain new knowledge by extracting significant information from large amounts of unstructured documentary texts available on the Internet and corporate intranets, using elements as varied as lexicographic and semantic analysis, groupings, categorizations and taxonomies; links, relationships and associations between entities; sentiment analysis or opinion mining, word frequency, and so on. The applications of all this are very wide-ranging: identification of texts and their corresponding extraction of elements; categorization and/or taxonomy of texts, extraction of ideas, subjects, concepts, entities, relationships, and events; translation of texts, optical character recognition, to name a few.

A good number of current uses of big data in libraries and the entire Library and Information Services or LIS industry fall within the field of Artificial Intelligence or AI, with diverse applications in several subfields within it. In the first place is "machine learning" for the most assorted purposes: indexing, cataloging, classification, online information retrieval, abstracting, reference services, tables of contents, and so on. Many of these organizations –libraries, publishers and the like– have already begun to build practical applications of machine learning in many different areas, such as document analysis and summarization. This involves programs able to "read" a certain text and extract information from it. Systems of this type are built for very specific documents –texts,

images, sheet music, etc.– as there is not yet a universal interpreter system for all types of documents. Similarly, there is no system to date capable of read books and build their complete catalog entries from them in a systematic and reliable way, yet there are those which can extract enough coherent information to provide valuable elements for individuals, such as catalogers, or for systems, such as library discoverers. This has become one of the applications deserving further reflection in libraries: there is currently a debate on whether catalogs should continue to be built in the traditional style or whether it is already necessary to make a change towards new structures of document ordering and retrieval supported by these elements (Bourg 2017).

Another practical example of machine learning programs well known in libraries are the Optical Character Recognition or OCR systems, used to interpret text which has been scanned as an image to convert it into computer interpretable text formats: doc, odt, txt, pdf, rtf, and others. These types of programs belong to the field of AI since their task is to read and interpret letters from a graphic form, in the same way as humans do, and they fall into the subfield of machine learning because these programs can learn what people tell them, such as interpretation errors, blemishes; old, discontinued, and serif typefaces, as well as the relevant corrections.

Another subfield of AI long and regularly used in libraries is the so-called expert systems. They have been of interest to librarians as early as the 1980s, as can be seen from the numerous texts on the subject back then, which dealt with knowledge-based indexing, natural language processing, cataloging, information retrieval and querying, among other topics. Expert systems are computer programs using AI principles and methods to solve problems within a specialized field which usually require the skills of expert personnel; hence the name. They incorporate the accumulated know-how of people proficient in a subject and are designed to function as closely as possible to them. In general, they contain a knowledge base of facts and relationships represented in the form of data and links, and obviously have the ability to make inferences based on

them. Different techniques for the creation of this knowledge base are used by the designers of these systems, such as the analysis of written protocols and procedures, the verbal description of tasks performed by an individual, questionnaires, surveys and interviews; the observation of processes and their simulation, as well as the discovery and documentation of tacit knowledge within the organization. The latter is a scarcely explored field but nevertheless of great value in libraries, since much of the librarians' knowledge about information management and exploitation falls within this possibility. The tacit or internal knowledge of librarians is their accumulated knowledge, generated by their experience, inherent to library staff and which has usually been materialized through different processes within the library.

Many other practical applications in today's libraries use AI expert systems and/or machine learning:

- Examples were mentioned about how some libraries take advantage of and/or conduct studies on the ways in which patrons search for and retrieve information, to learn more about the logic and ways they use to access information; all with the ultimate goal of improving library's internal catalogs, information discoverers, OPACs, etc.;
- A noticeable number of these studies have to do specifically with the natural language used by patrons for searching, in order to teach computers to understand and decipher such language, by means of AI. By their nature, these applications require very large amounts of data;
- Many library software programs store information from previous patron searches to learn from them, in order to personalize the page for each user, remembering what they have previously searched for and establishing patterns, suggesting later to the patron items like "people who consulted this text also consulted these others..." or "this author or topic is related to this one";
- With these memorizing and tailoring techniques, the library's web pages may allow each patron to create and save

the form, appearance and physical layout of his/her page, the displaying formats, previous searches, and so on, allowing the personalization of each person's page to be their own and distinct, to their own preference and convenience;

- Many libraries already extract data from the social networks of their patrons interconnected to the library services, to detect "trending topics", count "likes" and other similar events about their services or information, receive suggestions for the acquisition of items, verify effectiveness and follow up on their services, measure "usability" of new services and options, detect failures or problems, design new tutorials, and many other uses.

All these are just a few examples to illustrate the numerous applications that libraries are already using through AI and its subfields of expert systems, machine learning, etc.; all of them based on the constant and solid use and exploitation of their own data.

Despite the countless advantages and benefits for libraries which can be obtained from data management, especially big data, like all technologies created by human beings, it has a downside which librarians need to be fully aware of in order to neutralize it. In the first instance, big data is difficult to handle, partly because of its own nature of immense volume, velocity, and variety, and partly because there is a widespread lack of knowledge about how to handle it properly, due to the novelty of the subject and the worldwide shortage of experts in it. Both factors frequently cause a poor approach to objectives and techniques, duplicated data, inconsistency or bias, poor selection of analysis tools, erroneous interpretations, etc., with the subsequent negative consequences. Big data projects present great technical challenges and require a certain budget; obviously all this is a drawback. But it has been stressed that the most negative aspect of the use of big data, being the one that deserves the greatest care, is introducing the risk of eventual abuse of privacy and confidentiality of personal data.

Data management, unfortunately, often involves the collection of personal data, which entails enormous privacy risks. When large

amounts of data are compiled and stored –if personal data are included among them– the risk that they could be used for purposes different from the originally intended is introduced, due to information leaks from the organization's equipment, deliberate or by mistake, resulting in the extraction of data by third parties with mischievous intentions. As established, computer data privacy is not simply a computer security issue: it also involves ethical principles and issues, data privacy legislation and standards, data governance, technical and administrative responsibilities, as well as accountability. Data protection is essentially a technical issue which involves securing data against unauthorized access: who takes care of it and how. Data privacy goes far beyond that: it is an ethical and legal issue, involving deeper aspects in the organization: who can have personal data, for how long, who defines those who can access it, who can access it with authorization and under what circumstances, who can modify it, to whom and how it can be transferred, and so on. Therefore, data protection is a necessary but not sufficient condition to achieve a greater objective, data privacy.

Privacy is not a matter of just good intentions: it obeys international principles derived from the Universal Declaration of Human Rights. In recent years, the United Nations General Assembly adopted several resolutions on the Right to privacy in the digital age, urging all countries to *respect and protect the right to privacy*. Derived from this, many countries and regions have adopted measures in this regard, the most advanced and important of them and a model to follow being the 2018 European Union's Personal Data Protection Regulation or GDPR. This model has become also one of the most valuable allies for libraries in this task, since it includes all the essential regulations for providers, which were studied in detail in the corresponding section of this text. It is recommended that libraries outside the region always check the terms of service of their suppliers to verify the extent to which they comply with this regulatory framework. In addition, more than one hundred countries have currently some level of privacy and data protection legislation and regulations. Mexico has a broad legislation in this regard by which libraries are obliged subjects.

In the library field, the ALA has clearly established its position on the matter since many decades ago, included in its Interpretation of the Library Bill of Rights; IFLA has also pronounced itself on the subject in its IFLA Statement on Access to Personally Identifiable Information in Historical Records, taking among its principles the freedom of access to information and freedom of expression in which –in a very special and explicit way– privacy is considered as an inalienable part of the safeguarding of those rights.

The aforementioned principles address of privacy and confidentiality of personal data. The difference was clarified for their use within libraries: they emerged a long time ago with *privacy*, meaning this in a library that users have the right to read and consult whatever their wishes without the texts of his interest being subject to scrutiny by third parties. *Confidentiality* arises when a library comes into possession of personal data making the user identifiable, and therefore it must take measures to prevent unauthorized access. In short, confidentiality is a process protecting –among other things– privacy, which becomes a right of every patron. Confidentiality is the ethical and legal obligation of the library to protect that right.

Dating back almost a century, these principles have served as a guide for countless libraries around the world to build privacy policies and mechanisms into their information services, as well as measures to maintain user confidentiality; a relatively simple task with "traditional" library processes that collected personal data, as these were destroyed in the short term. Search through catalogs, indexes and other printed material left no further trace associated with the patrons and their interests at that time. All libraries which compiled and used data for collection usage statistics always did it anonymously, also under privacy principles. Regrettably, data globalization in the digital world brought about a massive invasion of privacy and confidentiality of personal data. Due to the rise of large-scale collection of such data and electronic surveillance, freedom of information, privacy and confidentiality have been seriously threatened in recent decades. This is especially sensitive in big data, hence the negative aspect of its use in

libraries. Obviously, this is not a problem exclusive to libraries, but it undoubtedly affects them to a greater extent and therefore requires total care and attention.

The collection, transfer and sale of personal data has become a huge business and therefore has introduced serious privacy risks for libraries and their patrons, as it involves data collecting from them by LIS providers, cloud services, phone apps, etc., which often is surreptitious or disguised. Library acquisition of digital and networked publications and information services: reference works, discoverers, search and documentation services, tables of contents, etc., from commercial providers outside the library therefore introduce serious privacy risks, as they all too often seek to have personal data capture requirements applied in libraries. Obviously this is totally unacceptable, as it entirely contradicts longstanding library principles about their patrons' privacy.

Whilst not an easy task, it is not impossible either, and there is a solution. It is indeed a problem aggravated by technology, but its solution does not rely on it: like many other problems in the library, it is largely a matter of method and procedure. As already established, the problem of data protection –the technological part– must be addressed, but it is only a minor component. The substantive part of it must include even broader aspects, covered by data governance techniques and principles, which were analyzed in the corresponding section. Based on these, each library can and should build its own privacy and data protection policies appropriate to its context and environment, which will provide a solid basis for planning and actions to protect the privacy and personal data of its patrons. Such policies should include ethical and legal issues and be compliant with local laws to become the organization's frame of reference, and should establish the general terms about the people who are and will be responsible for defining privacy and personal data issues within the organization, drafting and updating institutional privacy plans and programs, supervising them, accessing, transferring and safeguarding the data. Policies are documents at a theoretical level covering the entire organization as a whole; therefore they tend

to be much more stable over time. Based on policies as a foundation, the library should develop procedures, guidelines, best practices, standards, and so on. They are the practical versions which implement the concepts outlined theoretically and generally in the policies, and detail pre-established, sequential actions covering the full range of library processes and departments. Policies establish the *why*, *what* and *who*; procedures and guidelines establish the *how*, *when*, *where* and, where appropriate, detail the *who*. Accumulated experience will generate best practices and will assist to select standards. It is also advisable to design and carry out audits of all the services offered by the library, both internal and through suppliers, to ensure that all points of confidentiality have been covered, especially in big data projects. Such aspects were detailed in the corresponding chapter.

Emphasis was made on the fact that whenever practicable, the library should anonymize as much as possible the data used for its projects, statistics or feedback: data that is not held cannot be leaked or extracted. Any service or application designed or built by the library should always reflect beforehand about which data are indispensable for it and which are not: in the vast majority of cases it happens that projects work well without including personal data or with a minimum of them. This type of data should not be collected for any project or service if it is not indispensable. Most of library processes collecting data for analysis –such as statistics on the use of library collections– are entirely feasible to compile without recording sensitive patrons' data. As has been seen throughout that section, by minimizing the number of points in the library where personal data is handled, the number of places to protect is significantly reduced, which greatly facilitates the task for their managers. As it was outlined, there are relatively simple methodologies for reducing or even eliminating the use of personal data across many of the library's applications and services, without duplicating such data records across every department, section, service or project in the organization. As is evident, the set of policies, procedures, guidelines, standards, audits, and methodologies set forth herein confirm the principle stated

initially that most of it all consists of method and procedure, and only a minimal part of it are technological items. In personal data privacy, the most important thing still remains in *how* things are done, not with *what* technology they are done.

Nevertheless, technological tools are fundamental for data management, especially big data. For this reason, a complete chapter was devoted to the study of such tools: systems, software and applications, algorithms, methodologies, etc., both commercial and open access. Since there is no universal, large software system covering comprehensively all data processing needs, there are numerous highly specialized, specific tools for each type of purpose, produced by an infinite number of manufacturers. Since data projects are not the same and each one has its own context and characteristics, it is necessary in every one of them to integrate a certain set of methodological and informatics tools for their development and solution. It follows that one of the most important skills when starting out in the field of big data is to acquire a good general idea of all the varieties and capabilities of such tools, in order to be able to select the right one for each project and need.

Because of their broader general functions, some tools have become commonly used in several sectors: commercial, transportation, health, education; certainly, there are also tools for use in libraries and other information organizations. Due to their large number, no person knows or masters all the existing tools, but nevertheless it is necessary today that every library professional is acquainted in general terms with their offering, possibilities and capabilities, in order to approach the selection of any of them, in the same way that every professional librarian does when selecting an ILS system for library automation or management, a specialized search engine or discoverer, an automated catalog or OPAC, and the like. Today, the library manager is required to acquire an overview and general knowledge of data processing tools.

Due to the unprecedented growth of data management worldwide in virtually all sectors, there are countless data products, systems, services and applications already being developed by large companies and organizations. In particular, several major

library institutions have already developed and made available to smaller ones a number of tools that they can leverage to their advantage with minimal or no investment. Important examples were mentioned, as the Library of Congress Linked Data Service, that of the British Library, the Network of the National Library of Medicine (NNLM) the Smithsonian Institution Libraries' guidelines for the creation and deposit of data in repositories, among others. This has gradually become an important source for the acquisition of tools in that regard. In addition to the above, many of the usual library providers are also making developments with respect to big data, with which libraries can make strategic alliances for their use and impulse, or simply acquire those products on preferential terms. The textual studies of OCLC or the SN SciGraph project of linked open data of the Springer Natural Sciences Division were mentioned in this regard.

There are also multiple options of usable tools specifically built for libraries from numerous commercial providers dedicated to some of the aspects of data processing and analysis. The provision of these in the form of services under the commercial scheme known as *the cloud* was analyzed, of which a summary of its main modalities was presented for basic knowledge: Software as a Service or SaaS, Platform as a Service or PaaS, Infrastructure as a Service or IaaS, and of course, Data as a Service or DaaS, as well as Data and Platform as a Service (DaPaaS). It was highlighted that all the major IT commercial platforms existing today –Google, Amazon, Microsoft, Apple, Facebook, IBM, Oracle, etc.– offer some kind of big data related services, some free and some paid. This type of cloud services is increasingly used as it allows users to acquire large resources with relatively lower investments. However, it should be used with extreme caution for the security and privacy issues already mentioned. Libraries have been taking advantage of many of these tools and services for some years now, incorporating them into their work, with infinite possibilities.

For the purposes of reviewing them, tools were arbitrarily divided into the following types:

- Database management systems, both SQL and NoSQL;
- Documentary and textual data managers;
- Tools for data normalization and data mapping;
- Tools for big data analytics, in order to extract patterns or trends from it;
- Tools for visualization, interpretation or presentation of results;
- Artificial Intelligence tools;
- Specific application tools, such as those for mashups, and others.

Since in recent times a special emphasis has been developed in the field and in the tools for data analytics, a special section was dedicated to it. This specialty contains the core of data management; i.e., detecting trends and patterns from data in order to propose solutions, but more specifically, it refers to the task of identifying which variables of the organization can be linked to certain data and thus establish correlations for the posing of questions and the eventual obtaining of answers through specific techniques. As mentioned, this subject must be studied with caution and understood in its full extent. Obviously the main purpose of data science and management is to be able to extract valuable and useful information from data sets, since their simple collection and processing have little value in themselves: their benefit is only realized when they are analyzed. Therefore there is no doubt that data analysis is the central and most important part of the data science and management, especially big data, since this is where the relevant solutions and decisions for an organization are extracted, and therefore it deserves special attention and study. Consequently, there are more tools, texts, and products in this regard compared to other areas and tools of data management. But it is essential to contextualize and balance this field along with the others: not doing so jeopardizes the ultimate outcome and purpose of data management. From looking at the countless companies, websites, software tool offerings, training and consulting, courses, graduate degrees and diplomas, certifications, etc., a reader is left with

the perception that just this is all that needs to be addressed with respect to data science and data management: if the organization solves the data analytics chapter, the entire problem is solved.

This can be explained by the aforementioned global shortage of experts in data science and management. Obviously, organizations need to educate and train personnel in this area, but it is essential to do that in a balanced way in all areas of data science and management. It is not advisable to develop only one of its fields, no matter how central it may be, as this introduces a harmful and dangerous imbalance for the final purpose. It is not convenient for any organization to have only data analytics experts if it lacks skilled staff who can cope with other aspects of data management. Likewise, on a personal level, it is not recommended to get trained only in data analysis if there are no additional experts in other data areas within the institution; all this personal knowledge and skills will be useless for successful data management, since there is no context and balance across the organization.

As a conclusion to the above, it can be stated that organizations and individuals should consider data analysis as one of the multiple education and training options, but by no means as the only option to be developed on corporate or personal level: the whole offer should always be considered in context. It should be remembered that there is currently a huge business model around data training, consulting, and certification. It has become a million-dollar business by multiple providers; this is not bad in itself, but it has artificially introduced a huge expectancy around the most attractive field of data analysis, where much of it does not address the overall context of what organizations or individuals really need, and where not all providers are serious or reliable. There are many who sell because it is trendy and pays good dividends, even if what they offer is not of quality or really useful. Hence the paramount importance of weighing up the *whole* offer and *with whom* the training will be obtained. The central point of all of the above remains that it must never be overlooked that to solve a data problem, data analysis alone is not enough; all parts of data management must work together in a comprehensive and balanced

way. The data process involves many steps; although not all of them are found in each of the related problems, several stages will be always required for its proper management: data design, modeling and categorization; data capture, collection or mining; data coding and cleansing; data normalization and structuring; inclusion of metadata; data transformation; aggregation or disaggregation; data validation; storage and preservation; data visualization and displaying; and finally, data analysis and interpretation. Therefore, a good data analyst not only knows and can analyze data, but must also be able to fully participate in all the other stages.

More specifically, topics and uses of data analysis within libraries were examined. An overview was given of what has been compiled by various authors in this regard: collection development; design of new services; knowledge discovery; knowledge about patrons preferences and needs; providing quantitative evidence of the value and impact of the library within its environment; improving of the educational and informational experience of users; learning analytics; and other uses to optimize decision making, both at the level of big data as well as on a smaller scale. It was stressed the importance that before analyzing data, libraries must be sure to carefully select those which actually represent certain facts or realities, in order to establish and obtain the correct correlations leading to real and relevant questions and answers. One of the most common mistakes made is to use the most readily available data, simply because it is easy to obtain, or to forcefully use the library's typical or traditional data. Good data analysis implies that data selection must have a close correlation with the facts to be represented; otherwise the results will be biased or irrelevant. Although it may seem obvious, this is not easy; it requires, on the one hand, a deep knowledge of the causes and effects of the library function or task to be represented, and on the other hand, a broad mastery of how to approach and design specific data correlations. This implies two specialties: the expert librarian in such function or task and the expert data analyst; the ideal situation is when these two specialties are held by the same person: a librarian with expertise in the area as well as good knowledge in data

analysis; however, this is not mandatory: two experts working together will do it well. But without this dual function properly applied, it is easy to fall into obtaining and using data that is too simplistic for the task, or to get lost in a sea of data most of which is irrelevant or unrepresentative of the function. Hence the importance of planning from the outset exactly what is to be measured, why and how.

The library areas where data analysis can be applied are numerous; in consequence the eventual points of data extraction and collection are even more so: catalog and collection searches, reference, book loans, document views and downloads, access to services, social networks between the library and its patrons, to mention a few. This is why it is essential to make a correct data design and their associated facts in order to be able to make a pertinent analysis. The benefit of applying all this in libraries has been treated and described by numerous authors, the vast majority of documents edited by the ALA, which attests to the subject's current importance. Skills and knowledge librarians should acquire in this regard are also described in detail: fundamentals of statistical concepts; recommended data sources for various library functions and processes, as well as guidance on how to use them; techniques for data collecting and cleaning; how to find data correlations with appropriate data analysis method, and how to visualize the results. Many other library organizations have also expressed interest in studying, disseminating and training on the topic of data analysis for library professionals. A number of authors in this regard already refer to data analysis applied in libraries more pointedly as “library analysis”, to specifically highlight with this term this important activity related to data in this environment and to make it evident that it is already a specialty within libraries.

Considering the increasing importance of data management in a large number of businesses and organizations, the relevance of its proper handling was emphasized: otherwise, data can cause more problems than benefits. Hence the importance of the issue of data governance, especially about big data. It was stated that

data governance aims to ensure the quality and security of data used in an organization by establishing and monitoring a set of policies, processes, positions, standards, metrics and responsibilities which ensure effective, efficient, secure and consistent use of such data. The important difference between *data governance* and *data management* was highlighted. The latter is a much broader concept describing all the processes used to plan, specify, enable, create, acquire, maintain, use, store, access, control, cleanse and govern data. Data governance is a part of data management which addresses the roles, processes, standards, metrics, and personal responsibilities for establishing clear ownership and access to data assets within the organization, their protected use, consistency, and accountability.

Data governance creates a data framework within an organization; data management entails the practical execution of such framework. Essentially, data governance seeks to ensure that where data is concerned the right people have the right responsibilities through relevant processes, standards and metrics. By its very nature, data governance leads to benefits that are generally not achievable through data management alone: providing a generalized and consistent data view, understanding and terminology; optimizing processes, procedures, roles and positions within the organization; improving data quality and ensuring its accuracy, completeness and consistency, as well as optimizing data placement in key areas; enabling full compliance with minimum legal requirements, especially those related to privacy and personal data protection; preventing conflicts and/or overlaps between different data sets; and saving operational costs. Basically, data governance is designed from the setting of an organizational policy created specifically for this purpose. The fundamental points of data governance are all concepts to be defined and established, from which actions and processes will be derived. It is desirable from the outset, but will become imperative as organizations capture bigger amounts of data: the simple data management will no longer be sufficient.

As has been seen throughout this text, one of the most important aspects for the successful data management has to do with human resources dedicated to this purpose. Its treatment –especially in big data– is a multidisciplinary task, and therefore involves a great diversity of information professionals, among which librarians must obviously be included, given their traditional training and experience. As it has been established, the correct data treatment is a highly specialized task, requiring new sets of knowledge, skills, experience and attitudes. Nowadays there are relatively few data experts worldwide with adequate capabilities on how to design, define, collect, cleanse, transform, store, analyze and present data projects and structures; the lack of these technical abilities results in poorly conceptualized data management projects or total inaction. It is therefore an emerging field of professional development of utmost importance in the information sector.

As stated, librarians need to know in general terms the fundamentals and principles of data science, data management and curation, how and where big data can be found and how to leverage it; bibliomining, data analysis and visualization, as well as the selection and application of software tools for its exploitation, all as a new part of their comprehensive training and experience in the discipline.

Regardless of their professional training, and in close relation to the type of library or organization in which they work, it is recommended that librarians acquire and deepen their knowledge and experience in data processing in order to be in a position to start developing projects in the very diverse fields of application within libraries discussed throughout the text: reuse of research data; its curation and repositories; new products and services in libraries; improvement of school educational efficiency; studying and designing of new taxonomies and metadata schemes for search and retrieval; bibliometrics and text analysis; incorporation of competitive advantages in companies in general; new products and services in Library and Information Services companies; using data in digital humanities, digital social sciences, social computing; in library management and resources optimization; and

obviously, data education and training. It is emphasized that the above is an indicative but not exhaustive list of the fields of data applicability in libraries.

As it could be perceived throughout the text, its chapters and sections presented a wide possibility of data projects in libraries and related organizations; undoubtedly their field of action is extremely broad. The central point is that each of these possible applications requires professionals with knowledge, skills, attitudes, experience, and mastery of specific data-related tools. It is obvious that no library can have specialists in each and every possible field of action, but it should also be clear that it cannot remain totally without some specialized staff. Therefore, libraries can and should train their data specialists in the fields of application of their own context and interest. Initially, because it enables them to keep up with the current needs and circumstances of the information world in order to remain competitive and interesting for their communities and sponsors. Also, because it allows libraries to train professionals in high demand nowadays, which are scarce worldwide, meaning this new and better jobs for current professional librarians and students.

It was outlined in detail what relevant authors identified as the main roles or major functions in this task based on real-world needs for current jobs. In addition to three traditional roles: data analyst, data engineer, and data journalist, three others closely related to discipline were aggregated: data librarian, data archivist, and data manager/curator. Authors also established the five fundamental aspects that these information professionals should develop in this regard: academic education; practical experience; knowledge, familiarity and understanding of the subjects; execution skills; and competencies in mastering tools and technologies. As mentioned throughout this text, the intention was to illustrate knowledge, familiarity and understanding of the topics as well as to indicate skills and competencies.

Amidst the huge range of courses, specialties, certifications, etc., offered commercially by countless companies about data, there is a modest but acceptable number of courses, texts and

training on the subject offered on a not-for-profit basis by various organizations, which are ideal for librarians to get started in this field. It is emphasized that great caution should be exercised in selecting them, as numerous companies and organizations –in many cases prestigious ones– offer courses supposedly free of charge to hook prospective students, but further analysis reveals that there are underlying hidden costs, some of them quite substantial, mostly derived from the issuance of course certifications. As in all learning, it is recommended to go from the general to the particular, and from the basic to the specialized. Therefore, it is suggested to develop an institutional and/or personal plan for data learning and training, with progressive evolution and deepening.

Data management –especially big data– is no longer a trendy technological topic, but an emerging reality. Even though there are still huge myths and exaggerations about its usefulness, there is no doubt that it can indeed be used systematically for the benefit of organizations, including libraries. The world's leading library organizations have already pointed out its importance and have made numerous studies and recommendations in this regard. It is a fact that data application fields within libraries are extremely diverse. Due to its growing boom and importance within the world of information, it is a topic that cannot and should not be avoided by libraries, staying in a comfort zone on the sidelines; it does indeed imply new efforts and rearrangements, expenses and hassles; but it is by no means a casual minor addition or a whim which can eventually be adopted by libraries as a technical curiosity; it represents both an enormous challenge and a great opportunity: on the one hand it implies that its library staff must acquire new knowledge, skills, experience and attitudes for its correct management, but on the other hand it represents new and immense opportunities to reposition the library within the contemporary responsibilities and tasks of its community. For the same reason, it requires organizational structure and qualified personnel to perform the job adequately, like many of the other substantive duties of the library.

Finally, it is emphasized once again that in data projects the main key to success is not in the IT tools –important as they are– but in the qualified personnel. Therefore, it is highly recommended today that professional librarians begin to acquire some education, training and experience in data science, management, analysis, curation, uses and applications. This is not merely an academic curiosity; it is in their own and their institutions best interest. The world of information keeps evolving, so it is imperative that libraries and their staffs evolve as well. As in many other aspects of technology and libraries, the systematic and professional treatment of data has already surpassed the initial stage of boom and hype, and has become a reality that should undoubtedly be studied, included, and seriously taken advantage of in the environment. Libraries should not and cannot wait until it becomes an old field of development to consider including it in them, as has happened with other technological developments.

Jeffrey Stanton made a reflection since 2012 about the topic of data and libraries which remains fully valid today, and is a whole editorial on the subject:

> [...] a librarian does not need to become a programmer, but every librarian interested in knowledge creation should have some essential familiarity with how various software tools can transform data. A librarian need not be a database engineer, but every librarian must understand the underpinnings of information retrieval tools. A librarian does not need to be a statistician, but every librarian should have a clear understanding of how descriptive summaries and basic tests of numeric data can be used and misused. Finally, a librarian does not need to be a graphic designer, but every librarian needs to recognize the features of effective data displays. In short, to fulfill their missions, librarians can exercise a range of sophisticated skills that squarely occupy the central ground between understanding information user needs on one end and data curation on the other (Stanton 2012).

Jesse Shera (1973) remarked many years ago that "information science deals exclusively with the transmission of signals, whereas

librarianship is based on human interactions and deals with ideas and knowledge as well as information". According to this view, contemporary professional librarians can nowadays offer great and important contributions to data science from their very particular perspective because of their technical skills and professional experience, but above all, because of their humanistic formation: an opportunity that should not be missed.

Bibliographical References

(All electronic references have been verified as accurate and extant as of March 31, 2021).

Acharjya, D.P.; Ahmed, Kauser (2016). A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools. In: *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 7, num. 2, 511-518. https://www.researchgate.net/publication/296550027_A_Survey_on_Big_Data_Analytics_Challenges_Open_Research_Issues_and_Tools

Affelt, Amy (2015). *The Accidental Data Scientist*. Medford, N.J.: Information Today.

ALA – American Library Association (1939). *Bill of Rights*. Adopted by ALA Council, June 19, 1939; amended October 14, 1944; June 18, 1948; February 2, 1961; June 27, 1967; January 23, 1980; January 29, 2019. http://www.ala.org/advocacy/intfreedom/librarybill

——— (2002). *Privacy: An Interpretation of the Library Bill of Rights*. Junio 2002. http://www.ala.org/advocacy/intfreedom/librarybill/interpretations/privacy

——— (2007). *ALA Privacy Tool Kit*, 8-9. http://www.ala.org/advocacy/privacy/toolkit

Alvares, Lilian; Araújo Jr., Rogério (2010). Marcos históricos da ciência da informação: Breve cronologia dos pioneiros, das obras clássicas e dos eventos fundamentais. En *TransInformação*, Campinas, Brasil, vol. 22, num. 3, set./dez. 2010,195-205 http://www.scielo.br/pdf/tinf/v22n3/a01v22n3.pdf

Alvaro, Elsa; Brooks, Heather; and Ham, Monica (2011). E-science librarianship: Field undefined. In *Issues in Science and Technology Librarianship*, num. 66 (summer). http://www.istl.org/11-summer/index.html

Andaur, Gabriela (2016). *Panorama de la Gestión de Datos de Investigación en América Latina y el Caribe*. Blog del proyecto Learn. Entrada del 16 de mayo de 2016. http://learn-rdm.eu/es/gestion-de-datos-de-investigacion-en-america-latina/

Artificial Intelligence and Machine Learning in Libraries (2018). Jason Griffey (Ed.). *Library Technology Reports*, ALA Techsource, vol. 55, num. 1. DOI: https://doi.org/10.5860/ltr.55n1

Austin, Sidley (2016). Top Ten Data Protection and Privacy Issues to Watch in 2016. In *Lexology*, January 11, 2016. http://www.lexology.com/library/detail.aspx?g=6f3d4d67-ba04-42a9-9131-28bbd4a13f9f

Ávila, Eder (2020). *Los datos enlazados y su uso en bibliotecas*. México: UNAM. Instituto de Investigaciones Bibliotecológicas y de la Información. http://ru.iibi.unam.mx/jspui/bitstream/IIBI_UNAM/56/3/datos_enlazados..pdf

Baepler, Paul; Murdoch, Cynthia (2010). Academic Analytics and Data Mining in Higher Education. In *International Journal for the Scholarship of Teaching and Learning*, vol. 4, num. 2. doi:10.20429/ijsotl.2010.040217

Bailey, Charles (1991). Intelligent Library Systems: Artificial Intelligence Technology and Library Automation Systems. In *Advances in Library Automation and Networking*, vol. 4, 1-23. http://eprints.rclis.org/4891/

Baker, Ryan (2015). *Big Data and Education*. New York: Teachers College, Columbia University. http://www.columbia.edu/~rsb2162/bigdataeducation.html

Ball, Rafael (2019). Big Data and Their Impact on Libraries. In *American Journal of Information Science and Technology,* vol. 3, num. 1, 1-9. DOI: 10.11648/j.ajist.20190301.11

Bell, Arthur (1947). *Christian Huygens & the Development of Science in Seventeenth Century*. London: E. Arnold & Co.

Bell, Steven (2013). Promise and Problems of Big Data. In: *Library Journal*. March 13, 2013. http://lj.libraryjournal.com/2013/03/opinion/steven-bell/promise-and-pro...

Berdondini, Andrea (2019). The information paradox. In: *"Towards Data Science"* (blog), Entry from: July 9, 2019. https://towardsdatascience.com/the-information-paradox-38a411517f15

Berendt, Bettina; Littlejohn, Allison; Kern, Philippe; Mitros, Piotr; Shacklock, Xanthe and Blakemore, Michael (2017). *Big data for monitoring educational systems.* Publications Office of the European Union, Luxembourg. https://doi.org/10.2766/38557

Bieraugel, Mark (2013). Keeping up with… Big Data. In: *American Library Association – Association of College and Research Libraries*. ALA Website http://www.ala.org/acrl/publications/keeping_up_with/big_data

Blummer, Barbara; Kentin, Jeffrey (2018). Big Data and Libraries: Identifying Themes in the Literature. In: *Internet Reference Services Quarterly*, vol. 23, num. 1-2, 15-40, DOI: 10.1080/10875301.2018.1524337

Borgman, Christine (2007). *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: MIT Press.

Bourg, Chris (2017). What happens to libraries and librarians when machines can read all the books? In *Feral Librarian* (blog). Entry from: March 16th, 2017. https://chrisbourg.wordpress.com/2017/03/16what-happens-to-libraries-and-librarians-when-machines-can-read-all-the-books/

Brinton, Willard (1914). *Graphic Methods for presenting facts.* New York: The Engineering Magazine Company. https://archive.org/details/graphicmethodsfo00brinrich/page/n13/mode/2up

Calhoun, Karen (2006). *The Changing Nature of the Catalog and its Integration with Other Discovery Tools;* Final Report, Prepared for the Library of Congress. March 17, 2006. http://www.loc.gov/catdir/calhoun-report-final.pdf

Cao, Longbing (2017). Data science: A comprehensive overview. In *ACM Computing Surveys*, vol. 50, num. 3, article 43, 1-42. DOI: http://dx.doi.org/10.1145/3076253

Carlson, Scott (2006). Lost in a Sea of Science Data. In: *The Chronicle of Higher Education,* Entry from: June 23, 2006. https://chronicle.com/article/Lost-in-a-Sea-of-Science-Data/9136

CNIL - Commission Nationale de l'Informatique et des Libertés (France). (2019). Map of data protection around the world. https://www.cnil.fr/en/data-protection-around-the-world

Codd, Edgar F. (1970). A relational model of data for large shared data banks. In *Communications of the ACM*, vol. 13, num. 6, June 1970, 377-387. https://doi.org/10.1145/362384.362685

Columbus, Louis (2017). IBM predicts demand for data scientists will soar 28% by 2020. In *Forbes Magazine*. Entry from: May 13, 2017. https://www.forbes.com/sites/louiscolumbus/2017/05/13/ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020/#232bc4927e3b

Cox, Andrew; Kennan, Mary Anne; Lyon, Liz; and Pinfield, Stephen (2017). Developments in Research Data Management in Academic Libraries: Towards an Understanding of Research Data Service Maturity. In *Journal of the Association for Information Science and Technology,* vol. 68, no. 9, 2182-2200.

DAMA - Data Management Association (s.d.). Data Governance Functional Reference Framework. In: *Data Governance Part III: Frameworks – Structure for Organizing Complexity*. NASCIO Governance Series, 5 https://www.nascio.org/wp-content/uploads/2019/11/NASCIO-DataGovernancePTIII.pdf

Dans, Enrique (2011). Big Data: Una pequeña introducción. Author's blog. Entry from: October 19, 2011. https://www.enriquedans.com/2011/10/big-data-una-pequena-introducion.html

Dartmouth Summer Research Project on Artificial Intelligence – (DSRPAI). John McCarthy & Marvin Minsky, July 1956. Hanover, New Hampshire. AI Magazine (2006), vol. 27, num. 4.

Dataversity (2017). *Descriptive Analytics.* https://www.dataversity.net/fundamentals-descriptive-analytics/

Davies, Roy (1989). The Creation of New Knowledge by Information Retrieval and Classification. In *Journal of Documentation*, vol. 45, num. 4, 273-301.

Devens, Richard Miller (1865*). Cyclopaedia of Commercial and Business Anecdotes*. New York: D. Appleton and company, 210. https://archive.org/details/cyclopaediacomm00devegoog

Diebold, Francis. (2013). *On the Origin(s) and Development of "Big Data": The Phenomenon, the Term, and the Discipline*. University of Pennsylvania, Draft https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2202843

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

Duy, Joanna; Vaughan, Liwen (2005). Are citation data a valid measure of journal use? An empirical examination in an academic context. In: *Proceedings of the 10th International Conference of the International Society for Scientometrics & Infometrics,* pp.390-397. Stockholm: Karolinska University Press. https://spectrum.library.concordia.ca/6414/1/ISSIfall2005_v9.pdf

Duy, Joanna; Vaughan, Liwen (2006). Can electronic journal usage data replace citation data as a measure of journal use? An empirical examination. In *The Journal of Academic Librarianship*, num. 32, 512-517. https://doi.org/10.1016/j.acalib.2006.05.005

Duy, Joanna; Vaughan, Liwen (2003). Usage data for electronic resources: a comparison between locally-collected and vendor-provided statistics. In *The Journal of Academic Librarianship*, vol. 29, num. 1, 16-22. https://eric.ed.gov/?id=EJ673398 Engard, Nicole (Ed.). (2009). *Library Mashups: Exploring New Ways to Deliver Library Data*. Medford, N.J.: Information Today.

Engard, Nicole (Ed.). (2015). More *Library Mashups: Exploring New Ways to Deliver Library Data*. Medford, N.J.: Information Today.

Expert systems: HTML, the WWW, and the librarian. (2014) In: *The Free Library* https://www.thefreelibrary.com/Expert+systems%3a+HTML%2c+the+WWW%2c+and+the+librarian.-a016880355

Farmer, Lesley; Safer, Alan (2016). *Library Improvement through Data Analytics*. Chicago: ALA Neal-Schuman.

Farney, Tabatha (2018). *Using Digital Analytics for Smart Assessment*. Chicago: ALA.

Fayyad, Usama; Piatetsky-Shapiro, Gregory; and Smyth, Padhraic. (1996). From Data Mining to Knowledge Discovery in Databases. In: AI *Magazine*, vol. *17,* num. 3, 37-54. https://ojs.aaai.org//index.php/aimagazine/article/view/1230 Cited by: Press, Gil (2013). *A Very Short History of Data Science*. Forbes https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#64a5f94455cf

Feigenbaum, Edward (1989). Toward the Library of the Future. In *Long Range Planning*, vol. 22, num. 1, 118-123. https://doi.org/10.1016/0024-6301(89)90059-9

Feria, Lourdes (2020). Minería de texto aplicada a un diagnóstico de usuarios en Ciencia y Tecnología: Aprendizajes para fortalecer la investigación bibliotecológica, 31-42. In *El manejo de datos. Aproximación desde los estudios de la información* (2020). Araceli Torres (Coord.) México: UNAM. Instituto de Investigaciones Bibliotecológicas y de la Información. http://ru.iibi.unam.mx/jspui/handle/IIBI_UNAM/L225

First International Conference on Learning Analytics and Knowledge – Proceedings of LAK'11 (2011). New York: Association for Computing Machinery. https://dl.acm.org/doi/proceedings/10.1145/2090116

Gantz, John F.; Reinsel, David. 2007. *The Expanding Digital Universe.* An IDC White paper. (International Data Corporation). Sponsored by EMC Corporation. March 2007. https://web.archive.org/web/20090612013506/http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf

Gantz, John F.; Reinsel, David. 2012. *The Digital Universe in 2020.* IDC. (International Data Corporation). Sponsored by EMC Corporation. December 2012. https://www.slideshare.net/arms8586/the-digital-universe-in-2020

Gartner Glossary. (2005). Definition of Big Data. Gartner Group Website. Entry: Big Data https://www.gartner.com/en/information-technology/glossary/big-data

Gartner Glossary. (2005). Definition of EIM. Gartner Group Website. Entry: Enterprise Information Management https://www.gartner.com/en/information-technology/glossary/enterprise-information-management-eim

Geertz, Clifford (1973). Thick Description: Towards an Interpretative theory of culture. In: *The Interpretation of Cultures*, New York: Basic Books, 3-31 https://chairoflogicphiloscult.files.wordpress.com/2013/02/clifford-geertz-the-interpretation-of-cultures.pdf

Gibbons, Paul (2015). *The Science of Successful Organizational Change.* Boger.

Gorbea, Salvador; Piña, Maricela (2013). Propuesta de un indicador para medir el comportamiento del desarrollo disciplinar de las Ciencias Bibliotecológica y de la Información en instituciones académicas. En *Investigación Bibliotecológica: Archivonomía, bibliotecología e información,* vol. 27, núm. 60, 153-180. http://ru.iibi.unam.mx/jspui/bitstream/IIBI_UNAM/A111/1/art60-7.pdf

Gorman, Michael (2000). *Our Enduring Values Revisited: Librarianship in the Twentieth Century*. Chicago: ALA. http://publish.illinois.edu/whylibraries/files/2019/10/gorman-2015-enduring.pdf

Halevi, Gali; Nicolas, Barnaby; Bar-Ilan, Judit (2016). The Complexity of Measuring the Impact of Books. In: *Publishing Research Quarterly,* vol. 32, 187-200. https://doi.org/10.1007/s12109-016-9464-5.

Halevi, Gali (2014). *Bibliometric Big Data and its Uses* http://bdigital.unal.edu.co/12475/7/biliometricsbigdata.pdf

Halevi, Gali; Moed, Henk (2012). The evolution of big data as a research and scientific topic: Overview of the literature. In: *Research Trends*, num. 30, 3-6 https://www.researchgate.net/publication/285119834_The_evolution_of_big_data_as_a_research_and_scientific_topic_Overview_of_the_literature

Hallo, Maria; Luján-Mora, Sergio; Maté, Alejandro; and Trujillo, Juan (2015). Current state of Linked Data in digital libraries. In: *Journal of Information Science,* vol. 42, num. 2. DOI: 10.1177/0165551515594729.

Han, Jiawei; Kamber; Micheline, and Pei, Jian (2011). *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann.

Harada, Takishi (2019). *Robotics and artificial intelligence technology in Japanese libraries*. IFLA WLIC, 21-22 August 2019. http://library.ifla.org/2695/1/s08-2019-harada-en.pdf

Hernon, Peter; Dugan, Robert; Matthews, Joseph (2015). Managing with Data: Using ACRLMetrics and PLAmetrics. Chicago: ALA.

Hayashi, Chikio *et al.* (Eds.) "Data Science, Classification, and Related Methods". Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japan, March 27-30, 1996. Springer. Índice en: https://d-nb.info/955715512/04

Hayashi, Chikio (1996). What is data science? In *Data science, classification, and related methods*, 40-51. Springer: Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japan, March 1996. https://www.springer.com/gp/book/9784431702085

Heery, Rachel (2004). Metadata Futures: Steps Toward Semantic Interoperability. In *Metadata in Practice*. D. I. Hillman and E. L. Westbrooks (Eds.), 257-71. Chicago: American Library Association.

Hey, Tony; Hey, Jessie (2006). E-Science and its implications for the library community. In *Library Hi Tech*, vol. 24, num. 4, 515-528. http://www.emeraldinsight.com/doi/pdfplus/10.1108/07378830610715383

Hey, Tony; Tansley, Stewart; Tolle Kristin (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

Huang, Wenyi; Wu, Zhaohui; Liang, Chen; Mitra, Prasenjit; & Giles, Lee (2015). A neural probabilistic model for context based citation recommendation. In *AAAI 2015: Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2404-2410. https://clgiles.ist.psu.edu/pubs/AAAI2015-neural-probabilistic.pdf

Huff, Darrell (1954) *How to Lie with Statistics*. New York: W. W. Norton & Co.

IBM (s.d.). *Predictive Analytics*. https://www.ibm.com/analytics/predictive-analytics

IFLA – International Federation of Library Associations and Institutions (2012). *IFLA Code of Ethics for librarians and other information workers*. Approved August 2012. https://www.ifla.org/files/assets/faife/codesofethics/englishcodeofethicsfull.pdf

IFLA – International Federation of Library Associations and Institutions (2015). *IFLA Statement on Privacy in the Library Environment*. Endorsed by IFLA Governing Board, August 4, 2015 https://www.ifla.org/node/9803

IFLA – International Federation of Library Associations and Institutions (2017). *Linked Data for Libraries*. https://www.ifla.org/best-practice-for-national-bibliographic-agencies-in-a-digital-age/node/8549

IFLA – International Federation of Library Associations and Institutions (2018). IFLA Journal, vol. 44, num. 3, October 2018 https://www.ifla.org/files/assets/hq/publications/ifla-journal/ifla-journal-44-3_2018.pdf

Jacknis, Norman (2017). The AI-Enhanced Library. Author's Blog. Entry from: June 21, 2017. https://jacknis.com/tag/library/

Jharotia, Anil (2016). Big Data Technology: Big Opportunity for Librarians. In: *Librarianship in ICT Age*. Agra: Y. K. Publishers,1-9. https://www.researchgate.net/publication/326972552_Big_Data_Technology_Big_Opportunity_for_Librarians

Joyanes, Luis (2013). Big Data: Análisis de grandes volúmenes de datos en organizaciones. México: Alfaomega.

Kalantari, Ali; Kamsin, Amirrudi; Kamaruddin, Halim; Ebrahim, *et al.* (2017). A bibliometric approach to tracking big data research trends. In *Journal of Big Data,* vol. 4, num. 30 https://doi.org/10.1186/s40537-017-0088-1

Kenwright, David (1999), Automation or interaction: what's best for big data? In: *Proceedings Visualization '99*, San Francisco, CA, 1999, 491-495, DOI: 10.1109/VISUAL.1999.809940

King, David Lee (2018). Big Data and Libraries. Author's Blog. Entry from: December 13, 2018. https://davidleeking.com/big-data-and-libraries/

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

Lang, Charles; Siemens, George; Wise, Alyssa; and Gasevic, Dragan (Eds.) (2017). Handbook of Learning Analytics. SOLAR – Society for Learning Analytics Research https://solaresearch.org/wp-content/uploads/2017/05/hla17.pdf

Lawlor, Bonnie (2016). An overview of the NFAIS 2016 Annual Conference: Data sparks discovery of tomorrow's global knowledge. In*: Information Services & Use*, vol. 36, num. 1/2, 3-21. DOI: 10.3233/ISU-160807

Lewis, Martin (2010) Libraries and the management of research data. In McKnight, S. (ed.) *Envisioning Future Academic Library Services: Initiatives, Ideas and Challenges*. London: Facet, 145–168. https://eprints.whiterose.ac.uk/11171/1/LEWIS_Chapter_v10.pdf

Lexico Dictionary (2014). Entry: Big Data. (Lexico By Oxford) https://quizlet.com/203476200/business-intelligence-cis-chapter-13-flash-cards/

Liu, Shan; Shen, Xiao-Lang (2018), Library management and innovation in the Big Data Era. In: *Library Hi Tech*, vol. 36, num. 3, 374-377 https://doi.org/10.1108/LHT-09-2018-272

Lomotey, Richard; Deters, Ralph (2014). Towards knowledge discovery in Big Data. In *Proceedings of the 8th International Symposium on Service-oriented System Engineering*. IEEE Computer Society, 181-191. DOI: 10.1109/SOSE.2014.25

López-Yepes, José (Ed.). (2004). *Diccionario Enciclopédico de Ciencias de la Documentación*. Madrid: Síntesis, 2 vol.

Lorica, Ben (2014). *Big Data Now*. Sebastopol, CA, USA: O'Reilly Media. https://www.oreilly.com/data/free/files/bigdatanow2013.pdf

Lyman, Peter; Varian, Hal. (2000). How Much Information? In: *Journal of Electronic Publishing*. December 2000, vol. 6, num. 2. http://www.press.umich.edu/jep/06-02/lyman.html

Lyon, Liz; Mattern, Eleanor; Acker, Amelia; & Langmead, Alison (2015). Applying translational principles to data science curriculum development. In: *iPres Conference Proceedings*. Chapel Hill, November 2015 http://d-scholarship.pitt.edu/27159/1/Applying_Translational_Principles_to_Dat.pdf

Lyon, Liz; Mattern, Eleanor (2016). Education for Real-World Data Science Roles (Part 2): A Translational Approach to Curriculum Development. In *International Journal of Digital Curation*, vol. 11, num. 2. DOI: 10.2218/ijdc.v11i2.417.

Manual de Ciudades Digitales (2012). Asociación Española de Usuarios de las Telecomunicaciones y la Sociedad de la Información. https://nuevasciudadesdigitales.wordpress.com/manual-ciudades-digitales/

Markey, Karen (2007). The Online Library Catalog: Paradise Lost and Paradise Regained? In: *D-Lib Magazine*, vol. 13, nums. 1-2, Jan./Feb. 2007. http://www.dlib.org/dlib/january07/markey/01markey.html

Marr, Bernard (2015). Why only one of the 5 V's of big data really matters. In: *IBM Big Data & Analytics Hub*. Entry from: March 19, 2015 https://www.ibmbigdatahub.com/blog/why-only-one-5-vs-big-data-really-matters

Marr, Bernard (2020). Forbes Magazine. Entry from: Februray 24, 2020. https://www.forbes.com/sites/bernardmarr/2020/02/24/the-9-best-free-online-data-science-courses-in-2020/?sh=6fabfd7d2bbf

Martínez Musiño, Celso (2020). Big Data - Análisis infor-métrico de documentos indexados en Scopus y Web of Science. In: *Investigación Bibliotecológica: Archivonomía, bibliotecología e información,* vol. 34, núm. 82, 87-102. http://dx.doi.org/10.22201/iibi.24488321xe.2020.82.58035

Mashey, John (1999). Big Data… and the next wave of infrastress. In *USENIX Annual Conference*, Monterey, CA., June 6-11, 1999. https://www.usenix.org/conference/1999-usenix-annual-technical-conference/big-data-and-next-wave-infrastress-problems

Matusiak, Krystyna (2019). Research Data Management and Libraries: Opportunities and Challenges, 59-74. In *El manejo de datos. Aproximación desde los estudios de la información* (2020). Araceli Torres (Coord.) México: UNAM. Instituto de Investigaciones Bibliotecológicas y de la Información. http://ru.iibi.unam.mx/jspui/handle/IIBI_UNAM/L225

De Mauro, Andrea; Greco, Marco; and Grimaldi, Michele (2016). A Formal Definition of Big Data Based on Its Essential Features. In: *Library Review*, vol. 65, num. 3, 122-135. DOI: 10.1108/LR-06-2015-0061

McJones, Paul (Ed.) (1995). *The 1995 SQL Reunion: People, Projects, and Politics*. Recorded May 29, 1995. Computer History Museum. CHM Reference number: X7466.2015.2015. http://archive.computerhistory.org/resources/access/text/2015/07/102740133-05-01-acc.pdf

Van der Meulen, Rob (2018). *What Edge computing means for infrastructure and operations leaders*. Gartner Research Group https://www.gartner.com/smarterwithgartner/what-edge-computing-means-for-infrastructure-and-operations-leaders

Mexico: Ley General de Protección de Datos Personales en Posesión de Sujetos Obligados. (DOF – 26-01-2017). https://www.dof.gob.mx/nota_detalle.php?codigo=5469949&fecha=26/01/2017

Milton, Simon (1998). Top-Level Ontology: The problem with Naturalism. In: N. Guarino (Ed.), *Formal Ontology in Information Systems,* 85-94. Amsterdam, The Netherlands: IOS Press.

Moed, Henk F. (2012). The use of datasets in bibliometric research. In *Research Trends*, Issue 30, September 2012 https://www.researchtrends.com/issue-30-september-2012/the-use-of-big-datasets-in-bibliometric-research/

Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishan, P., Qazvinian, V., *et al.* (2009). Using citations to generate surveys of scientific paradigms. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics,* 584-592.

Morris, Robert; Truskowski, B.J. (2003). The evolution of storage systems. In: IBM *Systems Journal*, vol. 42, num. 2, 205-217. DOI: 10.1147/sj.422.0205

Murtaugh, Fionn; Devlin, Keith (2018). The Development of Data Science: Implications for Education, Employment, Research, and the Data Revolution for Sustainable Development. In: Big Data and Cognitive Computing, 2018, vol. XX, num. 1. https://web.stanford.edu/~kdevlin/Papers/Murtagh-Devlin_2018.pdf

Naur, Peter (1974). *Concise Survey of Computer Methods*. Studentlitteratur: Lund, Sweden. Cited by: Press, Gil (2013). *A Very Short History of Data Science.* Forbes Magazine https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#64a5f94455cf

New media Consortium (2013). *The NMC Horizon Report: 2013 Higher Education Edition*. https://files.eric.ed.gov/fulltext/ED559358.pdf

New Zealand National Library (2017). Integrated Library Systems (ILS) checklist. In *Choosing an Integrated Library System (ILS)*. Library's Website https://natlib.govt.nz/schools/school-libraries/library-systems-and-operations/your-library-catalogue/choosing-an-integrated-library-system-ils

Nicholson, Shawn; Bennett, Terrence. (2016), Dissemination and discovery of diverse data: Do libraries promote their unique research data collections? In *International Information & Library Review*, vol. 48, num. 2, 85-93 https://www.tandfonline.com/doi/full/10.1080/10572317.2016.1176448

Ohsumi, Noboru (2000). From data analysis to data science. In *Data Analysis, Classification, and Related Methods*. Kiers, Rasson, Groenen, & Schader (Eds.). Heidelberg: Springer, 329-334.

Olavsrud, Thor (2012). Big Data Causes Concern and Big Confusion. In CIO *Blog*. Entry from: February 24, 2012 http://www.cio.com/article/700804/Big_Data_Causes_Concern_and_Big_Confusion?page=2&taxonomyId=3002

Olavsrud, Thor (2020). What is data governance? In *CIO Blog*. Entry from: February 24, 2020 https://www.cio.com/article/3521011/what-is-data-governance-a-best-practices-framework-for-managing-data-assets.html

Olendorf, Robert; Wang, Yan (2017). Big Data in Libraries. In: Suh S. and Anthony T. (eds.) *Big Data and Visual Analytics*. Springer, Cham, 191-202. DOI: https://doi.org/10.1007/978-3-319-63917-8_11

Olivares Marín, Susana (2020). *Proceso de gestión del conocimiento tácito en bibliotecas digitales universitarias*. Ph.D. LIS Dissertation. México: UNAM.

Oracle (s.d.). *What a relational database is.* https://www.oracle.com/database/what-is-a-relational-database/

Orland-Barak, Lily; Mazkit, Ditza (2017). *Methodologies of Mediation in Professional Learning.* Springer Intl.

Oskosh Daily Northwestern Newspaper (Wisconsin), 24 May 1960, 19. https://newspaperarchive.com/oshkosh-daily-northwestern-may-24-1960-p-19/

Oxford English Dictionary. Entry by "Information" – sub-entry "Information explosion" https://www.oed.com/viewdictionaryentry/Entry/95568#eid112206197

Parry, Marc (2018). Big Data on Campus, In *The New York Times*, Entry from: July 18, 2012. https://www.nytimes.com/2012/07/22/education/edlife/colleges-awakening-to-the-opportunities-of-data-mining.html

Perales, Alicia (1962). La documentación. In *Anuario de Biblioteconomía y Archivonomía*. Año II. México: UNAM - Facultad de Filosofía y Letras, 9-34.

Phetteplace, Eric (2012). Effectively Visualizing Library Data. In *Reference & User Services Quarterly* (ALA), vol. 52, num. 2, winter 2012, 93-97. https://www.jstor.org/stable/refuseserq.52.2.93

Piganiol, Pierre (1971). In Preface to OECD Report: *Information for a Changing Society*. Paris, France: OECD, 13. https://files.eric.ed.gov/fulltext/ED057307.pdf

Pinfield Stephen; Cox, Andrew; and Smith, Jen (2014) Research Data Management and Libraries: Relationships, Activities, Drivers and Influences. In PLOS ONE vol. 9, num. 12. https://doi.org/10.1371/journal.pone.0114734

Plale, Beth (2016). In *Notes from the School of Information Sciences*, University of Illinois Urbana-Champaign. Entry from February 23, 2016. https://ischool.illinois.edu/news-events/news/2016/02/project-will-help-researchers-explore-big-data-hathitrust-digitized

——— (2013). Big data opportunities and challenges for IR, text mining and NLP. In *Proceedings of the 2013 International Workshop on Mining Unstructured Big Data Using Natural Language Processing* (UnstructureNLP'13). New York, NY: ACM, 1-2 https://dl.acm.org/doi/10.1145/2513549.2514739

Plunkett, Tom; McDonald, Brian; Nelson, Bruce (2013). *Oracle Big Data Handbook*. McGraw-Hill Osborne Media.

Pollock, Rufus (2013). *What do we mean by Small Data*. Open Knowledge Foundation Blog. Entry from: April 26, 2013. https://blog.okfn.org/2013/04/26/what-do-we-mean-by-small-data/

Qazvinian, V., Radev, D. R., & Özgür, A. (2010). Citation summarization through keyphrase extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics, COLING '10*. (Stroudsburg, PA: Association for Computational Linguistics), 895-903.

Qin, Jian; Norton, Jay (Eds.) (1999). Knowledge Discovery in Bibliographic Databases. In *Library Trends*, vol., 48, num. 1, summer 1999 https://aquila.usm.edu/fac_pubs/4791/

Rees, Alan; Saracevic, Tefko (1967). Education for Information Science and its relation to librarianship. In *Special Libraries Association Annual Conference,* New York, May 29, 1967, 2. Cited by: Shera, Jesse (1968). Sobre Bibliotecología, Documentación y Ciencia de la Información. In *Boletín* UNESCO *de Bibliotecas*, vol. XXII, núm. 2, marzo-abril 1968, 62-70.

Revista Iberoamericana de Educación / Educaçao (2019). *Analítica del aprendizaje y la educación (Learning Analytics and education): clasificación, descripción y predicción del aprendizaje de los estudiantes*, vol. 80, núm.1 https://rieoei.org/RIE/issue/view/Learning%20Analytics

De Rosa, Cathy (2006). *Perceptions of Libraries and Information Resources: A Report to the* OCLC. Columbus, OH.: OCLC.

Ryle, Gilbert (1949). *Concept of the mind*. New York: Hutchinson & Co.

Salazar, Javier (2020). Plan para el desarrollo de la Ciencia de Datos y Big Data (PDCDBD) en la UNAM con fines académicos y administrativos, 94-113. En *El manejo de datos. Aproximación desde los estudios de la información* (2020). Araceli Torres (Coord.) México: UNAM. Instituto de Investigaciones Bibliotecológicas y de la Información. http://ru.iibi.unam.mx/jspui/handle/IIBI_UNAM/L225

Saracevic, Tefko (1992). Information Science: Origin, Evolution and Relations. In Vakkari, Pertti & Cronin, Blaise (Eds.). *Conceptions of Library and Information Science. Historical, Empirical and Theoretical Perspectives*. London: Taylor Graham, 5-27.

Schilling, Virginia (2012). *Transforming Library Metadata into Linked Library Data*, 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. In *Data Science Central Blog*. Entry from: July 23, 2018. https://www.datasciencecentral.com/profiles/blogs/importance-of-metadata-in-a-big-data-world

Schwartz, Meredith (2013). What Governmental Big Data May Mean For Libraries. In *Library Journal*. May 30, 2013. http://lj.libraryjournal.com/2013/05/oa/what-governmental-big-data-may-m...

Segal, Troy (2019). Investopedia. Prescriptive Analytics. https://www.investopedia.com/terms/p/prescriptive-analytics.asp

Shera, Jesse (1973). Toward a Theory of Librarianship and Information Science. In: *Knowing books and men; knowing computers, too*. Libraries Unlimited.

Showers, Ben (Ed). (2015). *Library Analytics and Metrics: Using Data to Drive Decisions and Services*. London: Facet.

SISENSE (s.d). Data Analytics Glossary. Diagnostic Analytics. https://www.sisense.com/glossary/diagnostic-analytics/

Small, Henry; Klavans, Richard (2011). Identifying Scientific Breakthroughs by Combining Co-citation Analysis and Citation Context. In: *Proceedings of 13th International Conference of the International Society for Scientometrics and Informetrics (ISSI 2011).*

Souza, Renato; Tudhope, Douglas; and Almeida, Mauricio (2011). *Towards a taxonomy of KOS: Dimensions for classifying Knowledge Organization Systems*. International Society for Knowledge Organization (ISKO). 11th ISKO International Conference. Rome (Italy): 23-26 February 2010.

Staff, Frank (1993). *The Penny Post 1680-1918*. The Lutterworth Press.

Stanton, Jeffrey (2012). Data Science: What's in it for the New Librarian? In *Infospace. The Official Blog of the Syracuse University iSchool*, Entry from: July 16, 2012 https://ischool.syr.edu/infospace/2012/07/16/data-science-whats-in-it-for-the-new-librarian/

Stephens, Owen (2011). Mashups and open data in libraries. In: *Serials*, vol. 24, num. 3, 245–250. DOI: http://doi.org/10.1629/24245

Streitfeld, David (1989). *Infomania*. The Washington Post, February 3, 1989. https://www.washingtonpost.com/archive/lifestyle/1989/02/03/infomania/54d862a2-ba33-4ffe-a810-f96a0c2ca3a3/

Swan, Alma; Brown, Sheridan (2008). *The Skills, Role and Career Structure of Data Scientists and Curators: An Assessment of Current Practice and Future Needs.* Report to the JISC. Truro, UK: Key Perspectives.

http://www.jisc.ac.uk/publications/documents/dataskills-careersfinalreport.aspx

Techopedia (2020). *What are some core principles of data governance?* https://www.techopedia.com/7/32187/enterprise/databases/what-are-some-core-principles-of-data-governance

Tenopir Carol; Birch Ben; and Allard, Suzie (2012). Academic libraries and research data services: Current practices and plans for the future. Association of College and Research Libraries. 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? In: *Journal of eScience Librarianship*, vol. 4, num. 2. https://escholarship.umassmed.edu/jeslib/vol4/iss2/4/

Tenopir, Carol, *et al.* (2017). Research Data Services in European Academic Research Libraries. In: LIBER *Quarterly*, vol. 27, num. 1, 23-44. DOI: http://doi.org/10.18352/lq.10180

Torres Vargas, Georgina Araceli (Coord.). (2020). *El manejo de datos. Aproximación desde los estudios de la información.* México: UNAM. Instituto de Investigaciones Bibliotecológicas y de la Información. http://ru.iibi.unam.mx/jspui/handle/IIBI_UNAM/L225

Tufte, Edward (2001). *The visual display of quantitative information*. 2nd ed. Connecticut: Graphics Press.

Tukey, John W. (1962). The Future of Data Analysis. In: *The Annals of Mathematical Statistics*, vol. 33, num. 1, 1-67. DOI: 10.1214/aoms/1177704711 https://projecteuclid.org/euclid.aoms/1177704711 Cited by: Press, Gil (2013). *A Very Short History of Data Science*. Forbes Magazine https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#64a5f94455cf

UNO (2013). Resolution 68/167 about the Right to privacy in the digital age. December 18, 2013 http://www.oas.org/es/sla/ddi/docs/UN_A-C_3-68-L-45_Rev1.pdf

Uschold, Michael; Grüninger, Michael (1996). Ontologies: Principles, Methods, and Applications. In: *Knowledge Engineering Review*, vol. 11, num. 2, Cambridge Journals On-Line, 93–155. DOI: https://doi.org/10.1017/S0269888900007797

A Very Short History of Data Science. Forbes Magazine https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#64a5f94455cf

Wang, Chunning; Xu, Shaochun; Chen, Lichao; and Chen, Xuhui (2016) Exposing Library Data with Big Data Technology: A Review. In: *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*. DOI: 10.1109/ICIS.2016.7550937

Warden, Peter (2011). *Glossary of Big Data*. O'Reilly Media.

Wheatley, Amanda; Hervieux, Sandy (2019). Artificial Intelligence in Academic Libraries: An Environmental Scan. In: *Information Services & Use*, vol. 39, num. 4, 347-356. DOI: 10.3233/ISU-190065

Whyte Angus; Tedds, Jonathan (2011). Making the case for research data management. DCC Briefing Papers. Edinburgh: Digital Curation Centre. https://www.dcc.ac.uk/guidance/briefing-papers/making-case-rdm

Witt, Michael; Horstmann, Wolfram (2016). International approaches to research data services in libraries. In*: IFLA Journal*, vol. 42, num. 4, 251–252. DOI: 10.1177/0340035216678726

W3C – World Wide Web Consortium (2011). *Library Linked Data Incubator Group Final Report* https://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/

Xu, Zeshui; Yu, Dejian (2019). A Bibliometrics analysis on big data research (2009–2018). In: *Journal of Data, Information, and Management,* vol. 1, 3-15 https://doi.org/10.1007/s42488-019-00001-2

Young, Jeffrey (2017). Libraries Look to Big Data to Measure Their Worth - and Better Help Students. In: *Digital Learning in Higher Education*. Entry from: November 17, 2017 https://www.edsurge.com/news/2017-11-17-libraries-look-to-big-data-to-measure-their-worth-and-better-help-students

Zeng, Marcia Ley; Qin, Jian (2008). *Metadata*. New York: Neal-Schuman.

Zhan, Ming; Widén, Gunilla (2017). Understanding big data in librarianship. In: *Journal of Librarianship and Information Science*, vol. 51, num. 2, 561-576. DOI: 10.1177/0961000617742451

Datos masivos en bibliotecas / Big Data in Libraries.
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 y Francisco González y Ortiz; revisión de pruebas Carlos Ceballos Sosa, Valeria Guzmán González y Francisco González y Ortiz; formación editorial, Mario Ocampo Chávez. Fue impreso en papel cultural de 90 g en los talleres de Migal Impresiones Digitales, 3er. Anillo de Circunvalación no. 73, Col. Barrio Santa Bárbara, Alcaldía Iztapalapa, C.P. 09000, Ciudad de México. Se terminó de imprimir en septiembre de 2022.